Artificial Intelligence Review (2024) 57:231
https://doi.org/10.1007/s10462-024-10849-5

# Large language models in healthcare: from a systematic review on medical examinations to a comparative analysis on fundamentals of robotic surgery online test

**Andrea Moglia[1] · Konstantinos Georgiou[2] · Pietro Cerveri[1,3] · Luca Mainardi[1] · Richard M. Satava[4] · Alfred Cuschieri[5,6]**

Accepted: 3 July 2024 / Published online: 6 August 2024


**Abstract**
Large language models (LLMs) have the intrinsic potential to acquire medical knowledge. Several studies assessing LLMs on medical examinations have been published. However, there is no reported evidence on tests related to robot-assisted surgery. The aims of this study were to perform the first systematic review of LLMs on medical examinations and to establish whether ChatGPT, GPT-4, and Bard can pass the Fundamentals of Robotic Surgery (FRS) didactic test. A literature search was performed on PubMed, Web of Science, Scopus, and arXiv following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) approach. A total of 45 studies were analyzed. GPT-4 passed several national qualifying examinations with questions in English, Chinese, and Japanese using zero-shot and few-shot learning. Med-PaLM 2 obtained similar scores on the United States Medical Licensing Examination with more refined prompt engineering techniques. Five different 2023 releases of ChatGPT, one of GPT-4, and one of Bard were tested on FRS. Seven attempts were performed with each release. The pass score was 79.5%. ChatGPT achieved a mean score of 64.6%, 65.6%, 75.0%, 78.9%, and 72.7% respectively from the first to the fifth tested release on FRS vs 91.5% of GPT-4 and 79.5% of Bard. GPT-4 outperformed ChatGPT and Bard in all corresponding attempts with a statistically significant difference for ChatGPT ($p < 0.001$), but not Bard ($p = 0.002$). Our findings agree with other studies included in this systematic review. We highlighted the potential and challenges of LLMs to transform the education of healthcare professionals in the different stages of learning, by assisting teachers in the preparation of teaching contents, and trainees in the acquisition of knowledge, up to becoming an assessment framework of leaners.

**Keywords** Large language models medical examination · Large language models medical education · ChatGPT medical examination · ChatGPT medical education · Large language models healthcare · Large language models surgery

Extended author information available on the last page of the article

## 1 Introduction

Currently, the growth of artificial intelligence (AI) based computations is doubling every 6 months, hence far exceeding Moore's Law.[1] Natural language processing (NLP), a subfield of AI, focuses on the interaction of human language with computer systems (Nath et al. 2022). NLP has advanced significantly with the advent of transformers, an AI architecture that has improved NLP without requiring any recurrent or convolutional layer (Vaswani et al. 2017). Transformers exploit the attention mechanism to determine which parts of the input are more relevant. Hence, transformers-based models have been used for language translation and text completion tasks. More recently, transformers have been applied successfully to domains other than NLP, e.g., computer vision (Hatamizadeh et al. 2021).

Large language models (LLMs) constitute one of the most successful applications of transformers. In essence, they are large pre-trained AI systems based on knowledge gained from huge datasets, using language as a tool for human-AI interaction that can be adapted easily across several domains and for diverse tasks (Singhal et al. 2023a). The impressive performance of LLMs on NLP tasks has been amply demonstrated over the past few years (Singhal et al. 2023a). Kingston et al. demonstrated that LLMs' performance and data efficiency increase with both model and dataset size (Kingston et al. 2021). LLMs have been shown to exhibit promising results across a wide range of tasks, including those requiring specialized scientific knowledge and reasoning, thereby enabling them to generalize rapidly and even exhibit reasoning abilities with appropriate prompt strategies (few shot, chain of thought, and self-consistency) (Brown et al. 2020; Cobbe et al. 2021; Li et al. 2021; Wang et al. 2022; Wei et al. 2022). By employing prompt engineering, LLMs can be adapted to downstream tasks without the need for fine-tuning (Liu et al. 2023b).

In 2018, Open AI (San Francisco, CA, United States) released Generative Pre-trained Transformer-1 (GPT-1), a 117 million-parameters autoregressive LLM.[2] It was trained using unsupervised pre-training followed by supervised fine-tuning on Common Crawl (a large body of publicly available text from the Internet) and Book Corpus (a set of thousands of books of various genres). During unsupervised pre-training, GPT-1 learned the statistical patterns and structures present in the text data to predict the next word in a sentence. During supervised fine-tuning, GPT-1 was trained with input–output pairs on specific tasks (natural language inference, question answering, semantic similarity, and classification) on smaller datasets.[3] For instance, if the task was text classification, GPT-1 was trained with labeled text samples to predict the correct labels. With fine-tuning GPT-1 specialized in a particular task. GPT-1 was followed by larger models: GPT-2 in 2019 with 1.5 billion and GPT-3 in 2020 with 175 billion parameters[4] (Brown et al. 2020).

Unfortunately, LLMs like GPT-3 may amplify social biases in the training data and generate incorrect outputs (hallucinations) or reflect negative sentiments (Liévin et al. 2022). For instance, LLMs can generate different occupations and levels of respect for different genders, by imposing the idea that intellectual "brilliance" belongs only to a gender (Shihadeh et al. 2022). In part, this is because LLMs are trained to predict the next (sequential) word in a large dataset of Internet text, and hence, the results may not always align with

[1] https://blog.google/intl/en-africa/products/explore-get-answers/an-important-next-step-on-our-ai-journey/

[2] https://openai.com/index/language-unsupervised

[3] https://openai.com/index/language-unsupervised

[4] https://openai.com/index/better-language-models

users' expectations (Ouyang et al. 2022). InstructGPT by OpenAI, a fine-tuned version of GPT-3 incorporating reinforcement learning from human feedback (RLHF), has enhanced the performance of LLMs significantly (Stiennon et al. 2022). InstructGPT was trained in three stages: initially, a dataset of human-written prompts was submitted as input to the OpenAI application programming interface (API) with human annotators labeling the desired output. This dataset was used to fine-tune GPT-3 with supervised learning. Secondly, a dataset was collected on a larger set of API prompts, with human annotators ranking the outputs of different models for each prompt. A reward model was trained on this dataset to predict the preferred output by the annotators. Thirdly, the supervised learning baseline model was fine-tuned through reinforcement learning, with the reward model optimizing the policy using a proximal policy optimization algorithm (Ouyang et al. 2022).

LLaMA (Large Language Model Meta AI), a smaller LLM than InstructGPT (ranging from 7 to 65 vs. 175 billion parameters) trained on a larger number of tokens, performed better than InstructGPT on several benchmarks (Touvron et al. 2023). The latest version of LLaMA, called LLaMA 3, is available in configurations from 8 to 70 billion parameters.

ChatGPT by OpenAI, the successor of InstructGPT, was launched on November 30, 2022.[5] It was trained using RLHF following the same methods as InstructGPT but with a different dataset. In this case, the dataset included one from InstructGPT and another with conversations where human trainers assumed the roles of both the users and the AI assistants. ChatGPT was trained with data from the Internet until the end of 2021.

Bard was the response of Google (Mountain View, CA, United States) to ChatGPT and was unveiled on February 8, 2023. It was capable of answering multimodal questions, e.g. mixing text and images. At the end of the output, it added also weblinks for a Google search using some keywords of the input question. The same company has also been working on Sparrow, another LLM based on RLHF, Gemma and Gemini. Galactica, a decoder-only transformer LLM by Meta (Menlo Park, CA, United States), was developed to organize scientific literature. It was trained on over 48 million papers, textbooks and lecture notes, millions of compounds and proteins, scientific websites, and encyclopedias (Taylor et al. 2022).

Claude by Anthropic (San Francisco) was designed to rely on Constitutional AI, a set of principles provided by humans, to improve the performances of LLMs (Bai et al. 2022). Its use is currently limited to users in the United States and the United Kingdom.

On March 14, 2023, OpenAI launched GPT-4, the successor of ChatGPT, accepting both image and text inputs, and generating text output. As with ChatGPT it was trained on publicly available data on the Internet and fine-tuned with RLHF (OpenAI 2023). GPT-4 also extended the size of the text which can be prompted as input, at the cost of increasing the computational complexity. In GPT models, there is a quadratic dependency between computational complexity and the length of the tokens sequence due to the self-attention mechanism in the transformer entailing pairwise comparisons between all tokens in the sequence. The maximum tokens sequence increased from 512 in GPT-1 to 4,096 in ChatGPT up to 128,000 in GPT-4.[6] For comparison, the latest version of Claude (Claude 3) can manage a sequence of 200,000 tokens.

Other LLMs handling multimodal data (text and images) include Flamingo, Bard, BLIP-2 (Bootstrapping Language Image Pretraining), CM3Leon, PaLM-E, and LLaVA[7] (Alayrac et al. 2022; Driess et al. 2023; Li et al. 2023b; Liu 2023a).

[5] https://openai.com/index/chatgpt

[6] https://openai.com/index/gpt-4-research

[7] https://www.microsoft.com/en-us/research/project/llava-large-language-and-vision-assistant/

LLMs can potentially store, combine, and explore scientific knowledge to find hidden connections between different searches and produce systematic reviews or meta-analyses on specific topics automatically (Taylor et al. 2022).

Because of the potential of LLMs to acquire useful knowledge encoded in medical databases, they are likely to have applications in healthcare, including knowledge retrieval, clinical decision support, synopsis of key findings, and triaging patients attending primary care clinics (Singhal et al. 2023a).

The ability to answer medical questions requires full comprehension of medical text, recall of appropriate medical knowledge, and reasoning with expert information. LLMs like ChatGPT, GPT-4, Google Bard, and Claude by Anthropic were not specifically trained for healthcare applications, since they were developed for general-purpose cognitive capability. The data on healthcare used to train LLMs came from openly available medical texts, research papers, health system websites, and online available health information podcasts and videos (Lee et al. 2023b). The training data did not include privately restricted data, e.g., as those contained in an electronic health records, or any medical information that exists only on the private network of a medical organization (Lee et al. 2023b).

Several LLMs have been developed for healthcare, including the LLM by Hippocratic (Kolkata, India), ChatDoctor, DoctorGLM, Clinical Camel (derived from LLaMA), Med-Alpaca (derived from LLaMA), PMC-LLaMA, HuaTuo, and ChatCAD (Han et al. 2023; Li et al. 2023d; Toma et al. 2023; Wang et al. 2023a; Wang et al. 2023b; Wu et al. 2023; Xiong et al. 2023). Recent efforts have led to multimodal LLMs for medicine like Med-PaLM M and LLaVA-Med (Li 2023a; Tu et al. 2023).

Research on LLMs for healthcare applications is expanding rapidly. A recent review highlighted that current studies are mostly focused on: i) medical education, such as assessing performances of LLMs in medical examinations and ability to provide information support to learners and teachers; ii) clinical practice, e.g., by generating clinical reports and summarizing clinical discussions; and iii) research e.g., to develop LLMs-based applications to collect and analyze medical literature (Wu et al. 2024). A list of applications of LLMs in healthcare is reported in Table 1. Radiology is currently the medical specialty where LLMs were mostly applied, followed by surgery and dentistry (Wu et al. 2024).

This work focused on LLMs on the medical education domain, in particular on the assessment of the performances of LLMs in medical examinations.

The capability of LLMs to pass or not a medical examination may open new opportunities in medical education for both teachers and learners. If LLMs can demonstrate proficiency in answering correctly questions and demonstrating reasoning capabilities in some medical area they could be used by trainees to learn about a specific topic, or learners may ask an LLM to explain them some concepts which they did not understand. If LLMs can demonstrate reliability on knowledge related to a medical field, then teachers could trust them in preparing new teaching contents like lectures, prepare examinations, and use LLMs as assessment frameworks to evaluate the responses of students to examinations, thus saving a considerable amount of time.

**Table 1** Applications of LLMs in healthcare

| Domain | Application | Examples of applications | Published studies |
|---|---|---|---|
| Clinical practice | Decision making | Analysis of patient symptoms and data to provide diagnosis and suggest treatment | Chiesa-Estomba et al. (2024)<br>Cocci et al. (2023)<br>Koga et al. (2023)<br>Kao et al. (2023)<br>Haemmerli et al. (2023)<br>Nazario-Johnson et al. (2023) |
| | Consultation | Provide a real-time consultation to a young practitioner seeking advice from an expert | Alanzi (2023)<br>Hsu et al. (2023)<br>Liu et al. (2023) |
| | | Patient education | Ayoub et al. (2024)<br>Bellinger et al. (2024)<br>Altamimi et al. (2023)<br>Gabriel et al. (2023)<br>Karacas et al. (2023)<br>Koh et al. (2023)<br>Lyu et al. (2023)<br>Mondal et al. (2023) |
| | Generation of medical documentation | Generation of medical reports, transcription of patient-doctor conversations, or summarization of patient history | Bosbach et al. (2023)<br>Zhou (2023) |
| | Discussion | Virtual assistant helping patients to schedule appointments and provide initial triage | Abi-Rafeh et al. (2023)<br>Ayoub et al. (2023)<br>Gebrael et al. (2023a)<br>Jacob (2023)<br>Lyons et al. (2023)<br>Sarbay et al. (2023) |

**Table 1** (continued)

| Domain | Application | Examples of applications | Published studies |
|---|---|---|---|
| Medical Education | Assessing performance in medical examinations | Assessment frameworks to evaluate learners | See Tables 2, 3, and 4 of the present systematic review |
| | Support to learners and teachers | Help students to understand concepts during self-study | Xie et al. (2024)<br>Dhanvijay et al. (2023)<br>Kaarre et al. (2023)<br>Lower et al. (2023)<br>Lebhar et al. (2023)<br>Lee (2023)<br>Mohapatra et al. (2023)<br>Sallam et al. (2023)<br>Totlis et al. (2023) |
| | | Help teachers to create teaching contents, and assess students | Agarwal et al. (2023)<br>Ali et al. (2024)<br>Sevgi et al. (2023)<br>Smith et al. (2023a) |
| | | Generation of simulated clinical cases | Chung et al. (2023) |
| Research | Published documents | Writing of scientific abstracts and articles | Babl et al. (2023)<br>Gao et a. (2023)<br>Macdonald et al. (2023)<br>Valentín-Bravo et al. (2023) |
| | | Summarization of key information from published articles, and conference discussions | Almazyad et al. (2023)<br>Liu et al. (2023)<br>Xie et al. (2023) |
| | | Generation of systematic reviews and meta-analyses | Alshami et al. (2023)<br>Anghelescu et al. (2023) |
| | | Automatic revision for peer-reviewed articles | Biswas et al. (2023) |

### 1.1 Work motivation

Since LLMs have exceptional natural language comprehension abilities and are trained on massive datasets they represent ideal candidates for professional benchmarks, including those related to healthcare (Holmes et al. 2023). Several studies testing LLMs on medical examinations were conducted, for instance on the United States Medical Licensing Exam (USMLE), a three-step examination to assess clinical competence, required for licensure for independent provision of healthcare in the United States (Gilson et al. 2023; Han et al. 2023; Kung et al. 2023; Nori et al. 2023; Shama (2023); Singhal et al. 2023a; Singhal et al. 2023b; Toma et al. 2023; Tu et al. 2023; Wu et al. 2023). Step 1 of USMLE is taken by medical students after completing their preclinical training. Step 2 is taken after graduation and its scores are considered for admission into residency programs. Passing Step 3 is required for being licensed to practice medicine without supervision. However, at present there is no published systematic review of LLMs on healthcare examinations.

Although surgery is a medical specialty that generates some of the largest volume of data in healthcare which can be processed by AI algorithms, there is currently no published evidence on how LLMs perform on tests related to robot-assisted surgery (RAS). This surged ahead of traditional direct manual operations given its undoubted improved efficacy, such that the global market of RAS is predicted to grow at an average rate (CAGR) of 16.8%, reaching 21 USD billion in 2030.[8] Assuming this prediction materializes, there is an urgent need to train an increasing number of surgeons in RAS. Recognizing the need for training in RAS, several curricula have been proposed but none has received universal acceptance and widespread adoption (Satava et al. 2020). Fundamentals of Robotic Surgery (FRS) is a multi-specialty, proficiency-based curriculum of basic cognitive and technical skills to train and assess surgeons to safely and efficiently perform RAS. The threshold for attaining proficiency in FRS was computed as the mean of the expert surgeons participating in a multicenter randomized control trial involving 12 surgical training centers, accredited by the American College of Surgeons (Satava et al. 2020).

### 1.2 Contributions

The first purpose of this work was to perform a systematic review of published literature on LLMs on healthcare examinations. The second aim was to see whether ChatGPT, GPT-4, and Bard are capable of passing the FRS test. The main contributions of this paper are as follows:

- The studies on LLMs in medical examinations are presented;
- The role of prompt engineering is discussed for each group of studies;
- A comparative analysis of ChatGPT, GPT-4, and Bard on the FRS test is performed;
- The future challenges of LLMs in medical education are presented.

The rest of the paper is structured as follows: Sect. 2 describes the literature search strategy, and the process to extract and analyze studies. Section 3 states the research questions of this work. The applications of LLMs, reported in Sect. 4, are subdivided into three groups: National Qualifying examinations, Medical Specialty examinations, and other tests in medicine. In Sect. 5 a comparative analysis of ChatGPT, GPT-4, and Bard on FRS test

[8] https://www.strategicmarketresearch.com/market-report/surgical-robots-market

is presented. First FRS is described, including the source to retrieve the online question set. Then, consistency of performances of these LLMs over trials is reported. For ChatGPT scores over multiple releases are presented. Section 6 deals with the discussion on the comparison of ChatGPT, GPT-4, and Bard on FRS test, underling similarities and differences with the published evidence resulting from our systematic review. Challenges of LLMs in medical education are then discussed. Conclusions are reported in Sect. 7.

## 2 Literature search

### 2.1 Search strategy

In August 2023, a literature search was conducted on PubMed, Web of Science, Scopus, and arXiv following the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) statement and the AMSTAR 2 tool for critical appraisal of systematic reviews (Appendix A) (Page et al. 2021; Shea et al. 2017). The search was limited to articles in English language with an abstract and published from January 1st, 2018 to July 31, 2023. The following search terms were used:

"Large language models medical education"
OR "ChatGPT medical education"
OR "large language models medical exam"
OR "ChatGPT medical exam"
OR "large language models medical examination"
OR "ChatGPT medical examination"
OR "large language models medical license"
OR "ChatGPT medical license"
OR "large language models surgical education"
OR "ChatGPT surgical education"
OR "large language models medicine"
OR "ChatGPT medicine"
OR "large language models healthcare"
OR "ChatGPT healthcare"
OR "large language models surgery"
OR "ChatGPT surgery"
OR "Large language models surgical exam"
OR "ChatGPT surgical exam"
OR "large language models surgical examination"
OR "ChatGPT surgical examination"
OR "large language models medical test"
OR "large language models surgical test"
OR "ChatGPT medical test"
OR "ChatGPT surgical test"
OR "Large language models surgical license"
OR "ChatGPT surgical license"

Reviews, letters, non-peer reviewed articles, conference abstracts and proceedings were excluded from the analysis.

### 2.2 Data extraction

Identified articles were screened by title and abstract, followed by full-text review, data extraction, and review of references. Two reviewers (AM and KG) independently screened titles and abstracts for relevance. The sample, phenomenon of interest, design, evaluation, and research type (SPIDER) tool was used to structure qualitative research questions (Cooke et al. 2012). In case of insufficient information, the corresponding authors of the articles concerned were contacted for further details. References were checked to retrieve further studies.

### 2.3 Data analysis

Since the studies concerned many medical examinations, they were subdivided into three distinct groups: National Qualifying Examinations, Medical Specialty Examinations, and other studies. For each group, a table was prepared to visually present the data of the studies. The SPIDER tool was applied to the studies of each group, reporting: the number of questions of the examinations (Sample), the name of the examination (Phenomenon of Interest), the LLM, datasets, and prompt engineering technique (Design), the passing score and results (Evaluation), and whether the study was qualitative or quantitative (Research type).

### 2.4 Risk of bias

By considering the nature of the review, a bias analysis according to the Cochrane tool for assessing risk of bias was not applicable. The bias was rated in terms of memory retention of LLM, overlap between test and training data of LLMs, management of missing data, and type of funding (e.g. private and/or public).

## 3 Research questions

By using the SPIDER tool, the following research questions were formulated to serve as a roadmap for the scientific investigation, ensuring a thorough analysis of the published literature and of the performances of major LLMs like ChatGPT, GPT-4, and Bard on the FRS test.

RQ1: What are the medical examinations where LLMs were applied? How do different LLMs compare on the same exam? What are the performances of these LLMs on other medical examinations?
RQ2: Which type of prompt engineering techniques were used to improve the reasoning of LLMs?
RQ3: Do ChatGPT, GPT-4, and Bard pass the FRS test on cognitive skills? What is their consistency in confirming performances in subsequent attempts? How do their scores vary over multiple releases?

RQ4: What is the variability of ChatGPT, GPT-4, and Bard not only in terms of the overall score but also in terms of how many times all the FRS questions were answered correctly and erroneously?
RQ5: What are the main challenges of LLMs in the different stages of medical education, e.g., preparation for medical examinations?

## 4 Applications of LLMs in medical examinations

### 4.1 Results of the literature search

The database search retrieved 2393 results. After title and abstract screening, the full texts of 106 records were screened, but only 57 were found eligible for inclusion. A total of 45 studies were retrieved for full-text analysis, including 10 additional studies after references check. A list of the excluded articles from the 106 screened ones along with the reason for exclusion is provided in Online Appendix B. The flowchart based on the PRISMA statement is shown in Fig. 1 (Page et al. 2021).

The 45 studies included in the review comprised 16 on national qualifying examinations (Fang et al. 2023; Gilson et al. 2023; Han et al. 2023; Jang et al. 2023; Kasai et al. 2023; Kung et al. 2023; Shama et al. 2023; Singhal et al. 2023a; Singhal et al. 2023b; Nori et al. 2023; Taira et al. 2023; Takagi et al. 2023; Thirunavukarasu et al. 2023; Toma et al. 2023; Tu 2023; Wu et al. 2023), four on neurology and neurosurgery (Ali et al. 2023a; Ali et al. 2023b; Giannos et al. 2023a; Hopkins et al. 2023), three on orthopedics (Cuthbert et al. 2023; Saad et al. 2023; Lum et al. 2023), two on anesthesiology (Angel et al. 2024; Shay et al. 2023), ophthalmology (Antaki et al. 2023; Mihalache et al. 2023), general surgery (Beaulieu-Jones et al. 2024; Oh et al. 2023), and radiology (Bhayana et al. 2023; Huang et al. 2023). The others included one study on examinations each on the following specialty: emergency medicine, family medicine, clinical informatics, cardiology, urology, gynecology, general practitioners, dermatology, gastroenterology, otolaryngology, and pharmacy (Kumah-Crystal et al. 2023; Hoch et al. 2023; Huynh et al. 2023; Li et al. 2023d; Liu et al. 2023c; Passby et al. 2023; Skalidis et al. 2023; Smith et al. 2023a; Suchman et al. 2023; Wang et al. 2023c; Weng et al. 2023). One study concerned admission to university while two concerned exams at a single institution (Giannos et al. 2023b; Huh et al. 2023; Strong et al. 2023).

### 4.2 Prompt engineering strategies

The following prompt engineering strategies were applied in the reviewed studies.

- Few-shot: the model is given a few demonstrations of the task at inference time as conditioning (Brown et al. 2020).
- One-shot: similar to few-shot but with one demonstration (Brown et al. 2020).
- Zero-shot: similar to few-shot but with a natural language description of the task instead of any examples (Brown et al. 2020).
- Chain of thought: demonstrations of intermediate natural language reasoning steps are provided in the exemplars for few-shot prompting: (Wei et al. 2022).

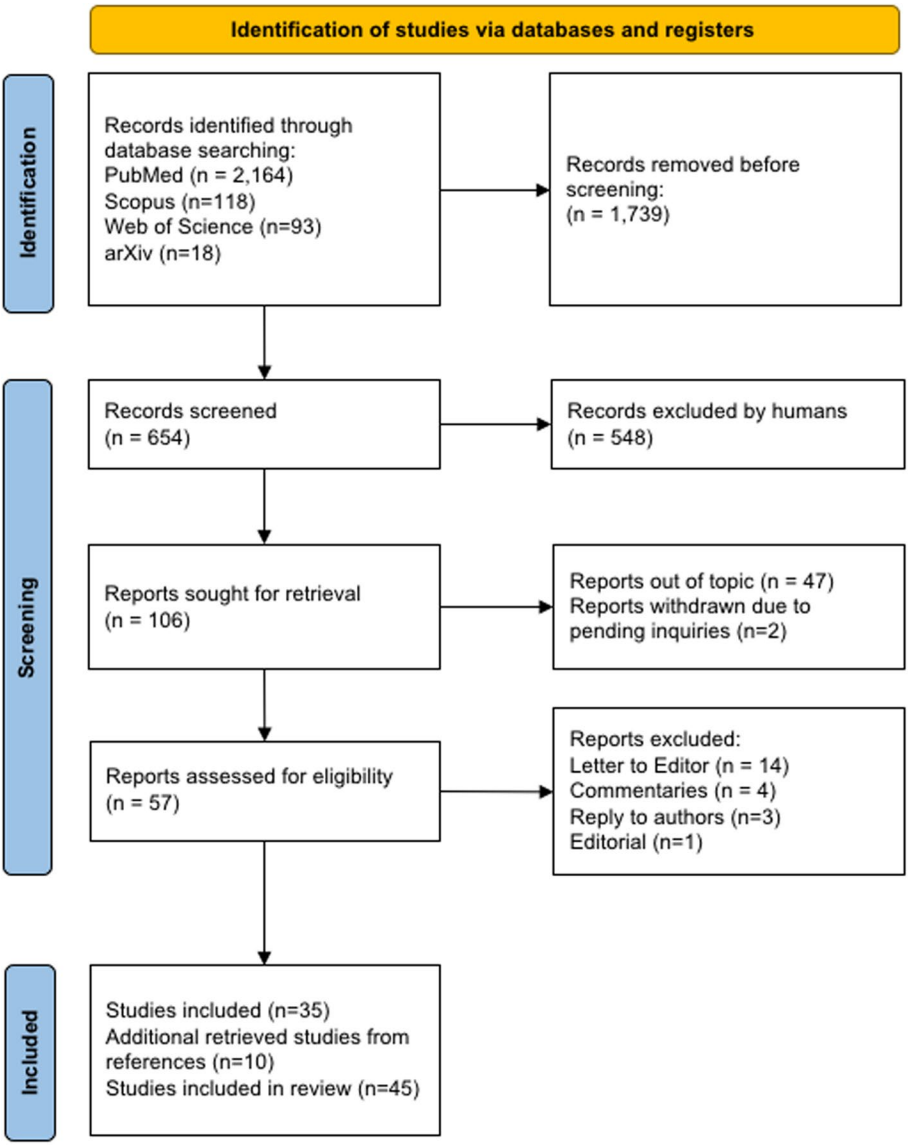


**Fig. 1** Flow chart of the study selection process according to the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) (Page et al. 2021)

– Self-consistency: an LLM is first prompted with a set of chain-of-thought exemplars. Then, a set of outputs from the LLM, generating a diverse set of reasoning paths, is sampled. Finally, the most consistent answer is chosen among the generated outputs (Wang et al 2022).
– Ensemble refinement: in the first stage an LLM is prompted with a set of chain-of-thought exemplars to generate a set of output, similar to self-consistency. In this case, each output involves an explanation of the answer. During the second stage, the LLM is

**Table 2** Studies on National Qualifying examinations

| Study | Sample (number of questions) | Phenomenon of interest | Design | | | Evaluation | | Research Type |
|---|---|---|---|---|---|---|---|---|
| | | | LLM | Dataset | Prompt engineering technique | Passing score | Results | |
| Kung et al. (2023) | 350 | USMLE | ChatGPT | Publicly-available test questions were from the June 2022 edition | Open end questions, removing all answer choices, and adding a variable lead-in interrogative phrase<br>Multiple choice questions, single answer without forced justification<br>Multiple choice single answer with forced justification | 60.0% | Method 1:<br>45.4%–75.0% (Step 1),<br>54.1%–61.5% (Step 2), and 61.5%–68.8% (Step 3)<br>Method 2:<br>36.1%–55.8% (Step 1),<br>56.9%–59.1% (Step 2), and 55.7%–61.3% (Step 3)<br>Method 3:<br>41.2%–64.5% (Step 1),<br>49.5%–52.4% (Step 2), and 59.8%–65.2% (Step 3) | Qualitative |
| Gilson et al. (2023) | 220 | USMLE | ChatGPT, GPT-3, InstructGPT | 100 from AMBOSS and 120 from National Board of Medical Examiners | | 60.0% | ChatGPT: 44.0%–64.4% (Step 1), 42.0%–57.8% (Step 2); InstructGPT: 36.0%–51.7% (Step 1), 35.0%–52.9% (Step 2); GPT–3: 20.0%-25.3% (Step 1), 17.0–18.6% (Step 2) | Qualitative |
| Sharma et al. (2023) | 114 | USMLE | ChatGPT | Two question banks: United States Medical Licensing Exam and AMBOSS | | 60.0% | 58.2% (logical questions)—60.0% (critical reasoning) | Quantitative |
| Singhal et al. (2023a) | 1,273 | USMLE | Med-PaLM | USMLE style from MedQA question bank | Few-shot, chain of thought, and self-consistency | 60.0% | 67.6% | Qualitative |
| Singhal et al. (2023b) | 1,273 | USMLE | Med-PaLM 2 | USMLE style from MedQA question bank | Ensemble refinement | 60.0% | 86.5% | Qualitative |
| Tu et al. (2023) | 1,273 | USMLE | Med-PaLM M | USMLE style from MedQA question bank | Few-shot | 60.0% | 69.7% | Qualitative |
| Nori et al. (2023) | 376 + 1,273 | USMLE | GPT-4 | Official test questions were from the June 2022 – July 2023 edition<br>+ USMLE style from MedQA question bank | Zero-shot, few-shot (5 shot) | 60.0% | Official USMLE: Released model: 84.0% (zero shot)—86.6% (five shot); Base model: 88.3% (zero shot) – 87.8% (five shot)<br>MedQA: Released model: 78.9% (zero shot)—81.4% (five shot); Base model: 84.4% (zero shot)—86.1% (five shot) | Qualitative |
| Toma et al. (2023) | | USMLE | Clinical Camel | USMLE style from MedQA question bank | | 60.0% | 53.2% (Step 1), 51.4% (Step 2), and 58.2% (Step 3) | Quantitative |

**Table 2** (continued)

| Study | Sample (number of questions) | Phenomenon of interest | Design | | | Evaluation | | Research Type |
|---|---|---|---|---|---|---|---|---|
| | | | LLM | Dataset | Prompt engineering technique | Passing score | Results | |
| Han et al. (2023) | 374 | USMLE | MedAlpaca | USMLE style from MedQA question bank | Zero-shot | 60.0% | 47.3% (Step 1), 47.7% (Step 2), and 60.2% (Step 3) | Quantitative |
| Wu et al. 2023 | 1,273 | USMLE | PMC-LLaMA | USMLE style from MedQA question bank | | 60.0% | 44.7% | Quantitative |
| Thirunavukarasu et al. (2023) | 674 | Applied Knowledge Test of the Royal College of General Practitioners | ChatGPT | Three online question banks for preparation to the exam | | 70.4% | 41.9%–61.6% (Trial 1); 45.2%–62.1% (Trial 2) | Quantitative |
| Takagi et al. (2023) | 254 (Questions in Japanese) | Japanese Medical Licensing Examination | ChatGPT, GPT-4 | Real exam (2023 edition) | Inserting a brief preamble specifying that the question was single-choice | 80.0% (on the essential knowledge questions) | Essential knowledge: 55.1% (ChatGPT), 87.2% (GPT-4) | Quantitative |

**Table 2** (continued)

| Study | Sample (number of questions) | Phenomenon of interest | Design | | | Evaluation | | | Research Type |
|---|---|---|---|---|---|---|---|---|---|
| | | | LLM | Dataset | Prompt engineering technique | Passing score | Results | | |
| Kasai et al. (2023) | 400 (Questions in Japanese) | National Medical Practitioners Qualifying Examination in Japan | GPT-3, ChatGPT (gpt-3.5-turbo), GPT-4 | Real exam (2018–2023 editions) | Three in-context examples | 70.0% (general sections)—80.0% (requires sections) | | GPT-3: 43.0%–53.8% on (required sections), 34.8%–39.5% (general sections); ChatGPT: 50.0%–71.5% on (required sections), 47.4%–54.9% (general sections); GPT-4: 80.5%–86.5% on (required sections), 72.6%–76.8% (general sections) | Qualitative |
| Taira et al. (2023) | 240 | National Nursing Licensing Examinations in Japan | ChatGPT | Real exam (2019–2023 editions) | | 80.0% (basic knowledge), and 60.0% (general questions) | | 75.1% (basic knowledge questions), and 64.5% (general questions) | Quantitative |

conditioned on the original prompt, question, and the concatenated generations, and is prompted to produce a refined explanation and answer. The second stage is performed multiple times. Finally, a plurality vote over the generated answers is conducted to determine the final answer (Singhal et. 2023b).

### 4.3 Studies on national qualifying examinations

The published studies on National Qualifying Examinations are reported in Table 2.

Ten studies concerned USMLE, one the Applied Knowledge Test of the Membership of the Royal College of General Practitioners, one the Japanese Medical Licensing Examination, one the National Medical Practitioners Qualifying Examination in Japan, one the National Nursing Licensing Examinations in Japan, one the Korean National Licensing Examination for Korean Medicine Doctors, and one the Chinese National Medical Licensing Examination.

Three studies concerned ChatGPT, one GPT-3, one InstructGPT, one Med-PaLM, one Med-PaLM 2, one Med-PaLM M, one GPT-4, one Clinical Camel, one Med-Alpaca, and one PMC-LLaMA. Two studies used original questions from a real edition of the USMLE examination (Kung et al. 2023; Nori et al. 2023), while the others used online question banks. Four of these used 1,273 USMLE-style questions from the MedQA dataset (Nori et al. 2023; Singhal et al. 2023a; Singhal et al. 2023b; Tu et al. 2023). As a consequence, the number of questions varied among studies, from 114 to 1,649 (Nori et al. 2023; Shama et al. 2023). The passing score of USMLE varies over the years but was generally close to 60.0%. In the study by Gilson et al., only ChatGPT met this threshold in contrast with GPT-3 and InstructGPT. The study by Kung et al. investigated three different strategies to prompt USMLE questions: (i) as open-end questions; (ii) multiple-choice questions; and (iii) multiple choice with forced justification, i.e., by asking ChatGPT to provide the rationale on the response. With the first method ChatGPT passed all steps, with the second method only step 3, and with the third method both step 1 and step 3.

One study on ChatGPT reached 60.0% on 45 USMLE questions concerning critical reasoning (Shama et al. 2023). Prompt engineering techniques have demonstrated success to increase LLMs performances. Med-PaLM was the first LLM passing USMLE thanks few shots, chain of thought, and self-consistency (Singhal et al. 2023a). It reached 67.6% of correct answers and was beaten by Med-PaLM 2 reaching 86.5% thanks to ensemble refinement (Singhal et al. 2023b). Med-PaLM M, which are LLMs for medicine, outperformed Med-PaLM, a multimodal LLM for medicine, achieving 69.7% using the few-shot technique. GPT-4 used zero-shot and five-shot prompt engineering strategies in two different configurations, namely the base and released model, with the latter aligned with safety (Nori et al. 2023). The base GPT-4 model was able to reach 88.3% vs. 86.6% for the released one with five-shot prompting on the real USMLE exam (Nori et al. 2023). Scores on USMLE-like questions provided by MedQA were in line with Med-PaLM 2, i.e., 86.1% for the based GPT-4 model and 81.4% for the released one. None of the LLMs based on LLaMA passed USMLE, except MedAlpaca on Step 3. This LLM used the zero-shot technique (Wu et al. 2023). In the study by Kung et al., the USMLE questions were prompted to ChatGPT in different modalities: open-end, multiple-choice questions without and with forced justification.

In the United Kingdom general practitioners must pass the Applied Knowledge Test of the Membership of the Royal College of General Practitioners to complete their training.

**Table 2** (continued)

| Study | Sample (number of questions) | Phenomenon of interest | Design | | | Evaluation | | | Research Type |
|---|---|---|---|---|---|---|---|---|---|
| | | | LLM | Dataset | Prompt engineering technique | Passing score | Results | | |
| Jang et al. (2023) | 340 (Questions in Korean) | Korean National Licensing Examination for Korean Medicine Doctors | ChatGPT GPT-4 | Real exam (2022 edition) | Inserting a brief preamble about the type of examination | 60.0% | | 42.1% (ChatGPT) 57.3% (GPT-4) | Qualitative |
| Fang et al. (2023) | 600 (Questions in Chinese) | Chinese National Medical Licensing Examination | GPT-4 | Online question bank | Questions were formatted by deleting all the choices and adding a variable lead-in imperative or interrogative phrase | 60.0% | | 73.7% | Qualitative |

List of abbreviations: United States Medical Licensing Exam (USMLE)

ChatGPT was tested in two different trials on questions from online question banks. However, it did not meet the passing threshold in either trial (Thirunavukarasu et al. 2023).

The Japanese Medical Licensing Examination is a mandatory exam for certifying medical practitioners in Japan. A study on 254 questions from the real 2023 edition has shown that GPT-4 passed it successfully in contrast with ChatGPT (Takagi et al. 2023). The National Medical Practitioners Qualifying Examination in Japan is taken by sixth-year medical students. It consists of a compulsory and a general part, for a total of 400 questions. Kasai et al. assessed GPT-3, ChatGPT (in the gpt-3.5-turbo configuration), and GPT-4 on six real editions (from 2018 to 2023) with questions in Japanese. All questions were prompted with three in-context examples. GPT-4 passed both parts in all editions, ChatGPT partly the compulsory part, while GPT-3 did not pass any part (Kasai et al. 2023). In a study on five real editions (from 2019 to 2023) of the National Nursing Licensing Examinations in Japan, ChatGPT passed the part on general questions (Taira et al. 2023). Both ChatGPT and GPT-4 did not pass the 2022 edition of the Korean National Licensing Examination for Korean Medicine Doctors after prompting questions in Korean, with a brief preamble about the type of examination (Jang et al. 2023). In contrast, GPT-4 passed the Chinese National Medical Licensing Examination with questions in Chinese from an online question bank (Fang et al. 2023). Questions were reformatted by deleting all the choices and adding a variable lead-in imperative or interrogative phrase, as in the study by Kung et al. (2023)

### 4.4 Studies on medical specialties examinations

The published studies on medical specialties examinations are shown in Table 3.

LLMs were tested on neurology/neurosurgery in four studies. GPT-4 passed both the UK Specialty Certificate examination and the written part of the American Board of Neurological Surgery, while ChatGPT only the latter (Ali et al. 2023b; Giannos et al. 2023a). Concerning the oral part of the American Board of Neurological Surgery, GPT-4 outperformed ChatGPT and Bard. In none of the three studies on orthopedics, either GPT-4 or ChatGPT reached the passing score. Concerning the American Board of Anesthesiology examination only GPT-4 reached the threshold, in contrast with GPT-3, ChatGPT, and Bard (Angel et al. 2024; Shay et al. 2023). The study on American Academy of Ophthalmology's Basic and Clinical Science Course did not specify the passing score (Antaki et al. 2023). ChatGPT did not pass the Ophthalmic Knowledge Assessment Program examination (Mihalache et al. 2023).

Since the passing score was not specified in a study on the American Board of Surgery Qualifying Exam and another on the Korean general surgery board exams, it was not possible to know whether or not LLMs passed them (Beaulieu-Jones et al. 2024; Oh et al. 2023).

It is interesting to note that none of the published studies on neurology, orthopedics, anesthesiology, ophthalmology, and general surgery assessed LLMs on real examinations but using either online question banks or mock of actual exams (Table 3). Two studies on neurology, one on ophthalmology, two on surgery, and one on gynecology did not specify the passing score (Table 3).

Both ChatGPT and GPT-4 passed the real version of the American College of Radiology Radiation Oncology in-training (TXIT) examination, while ChatGPT scored slightly below the threshold of the Canadian Royal College Examination in Radiology using online question banks (Bhayana et al. 2023; Huang et al. 2023). The remaining published studies

**Table 3** Studies on medical specialties examinations

| Study | Sample (Number of questions) | Phenomenon of interest | Design | | | Evaluation | | Research Type |
|---|---|---|---|---|---|---|---|---|
| | | | LLM | Dataset | Prompt engineering technique | Passing score | Results | |
| Giannos et al. (2023a) | 69 | UK Specialty Certificate Examination (SCE) in Neurology | ChatGPT, GPT-4 | Pool-Specialty Certificate Examination Neurology Web Question bank | | 58.0% | 57.0% (ChatGPT), 64.0% (GPT-4) | Quantitative |
| Ali et al. (2023a) | 149 | American Board of Neurological Surgery oral board examination | ChatGPT, GPT-4, Bard | Self-Assessment Neurosurgery Exam (SANS) Indications Exam for oral board preparation | | | 62.4% (ChatGPT), 82.6% (GPT-4), 44.2% (Bard) | Quantitative |
| Ali et al. (2023b) | 500 | American Board of Neurological Surgery written board examination | ChatGPT, GPT-4 | Mock of the real exam | | 69.0% | 73.4% (ChatGPT), 83.4% (GPT-4) | Quantitative |
| Hopkins et al. (2023) | 618 | American Board of Neurological Surgery written board examination | ChatGPT | Surrogate questions for preparation to the exam | | | 53.2% | Quantitative |
| Cuthbert et al. (2023) | 134 | FRCS examination in Trauma and Orthopaedic Surgery | ChatGPT | United Kingdom and Ireland In-Training Examination (UKITE) in Trauma and Orthopaedic Surgery as mock examination | | 65.8% | 35.8% | Quantitative |

**Table 3** (continued)

| Study | Sample (Number of questions) | Phenomenon of interest | Design | | | Evaluation | | Research Type |
|---|---|---|---|---|---|---|---|---|
| | | | LLM | Dataset | Prompt engineering technique | Passing score | Results | |
| Saad et al. (2023) | 240 | Orthopaedic (FRCS Orth) Part A examination | GPT-4 | Question bank with mock questions | | 70.0% | 67.5% | Quantitative |
| Lum et al. (2023) | 207 | American Board of Orthopaedic Surgery examination | ChatGPT | One online question bank for preparation to the exam | | 63.0% (in 2020) | 47.0% | Quantitative |
| Angel et al. (2024) | 130 | American Board of Anesthesiology examination | GPT-3, GPT-4, Bard | Two question banks: 60 sample questions provided on the ABA website, and 70 from the book "Anesthesia Review: 1000 Questions and Answers to Blast the BASICS and Ace the ADVANCED" | | 75.0% | GPT-4: 78.3% (basic section), 80.0% (advanced section)%; GPT-3: 58.3% (basic section), 50.0% (advanced section); Bard: 46.7% (basic section), 45.7% (advanced section) | Quantitative |
| Shay et al. (2023) | 1,321 | American Board of Anesthesiology written examinations | ChatGPT | Board examination preparation book, Anesthesiology Examination and Board Review | | 60.0%-70.0% | 56.2% | Quantitative |

**Table 3** (continued)

| Study | Sample (Number of questions) | Phenomenon of interest | Design | | | Evaluation | | Research Type |
|---|---|---|---|---|---|---|---|---|
| | | | LLM | Dataset | Prompt engineering technique | Passing score | Results | |
| Antaki et al. (2023) | 520 | Ophthalmic Knowledge Assessment Program examination | ChatGPT | Two question banks: American Academy of Ophthalmology's Basic and Clinical Science Course (BCSC) Self-Assessment Program and the OphthoQuestions online question bank | Zero shot | | 55.8% on BCSC; 42.7% on OphthoQuestions | Quantitative |
| Mihalache et al. (2023) | 125 | Ophthalmic Knowledge Assessment Program examination | ChatGPT | Question bank: OphthoQuestions | First trial: multiple choice questions<br>Second trial: open end by removal of multiple choice options | 62.5% | 46.4% (first trial/ multiple choice); 58.4% (second trial/ open ended) | Qualitative |
| Beaulieu-Jones et al. (2024) | 112 | American Board of Surgery Qualifying Exam | GPT-4 | Question bank for the preparation (Data-B) | Open end by removing all answer choices<br>Multiple choice single answer without forced justification | | 66.1% (open ended)—67.9% (multiple choice) | Qualitative |

**Table 3** (continued)

| Study | Sample (Number of questions) | Phenomenon of interest | Design | | | Evaluation | | Research Type |
|---|---|---|---|---|---|---|---|---|
| | | | LLM | Dataset | Prompt engineering technique | Passing score | Results | |
| Oh et al. (2023) | 280 (Questions in Korean) | Korean general surgery board exams | ChatGPT, GPT-4 | Question bank with question recalled by examinees who took the real exam | | | 46.8% (ChatGPT), 76.4% (GPT-4) | Quantitative |
| Bhayana et al. (2023) | 150 | Canadian Royal College Examination in Radiology | ChatGPT | Two question banks: 5 from the Canadian Royal College website, and 145 selected to match the Canadian Royal College examination | | 70.0% | 69.0% | Qualitative |
| Huang et al. (2023) | 293 | American College of Radiology Radiation Oncology in-training (TXIT) examination | ChatGPT, GPT-4 | Real exam (2021 edition) | | 60.0% | 63.6% (ChatGPT), 74.6 (GPT-4) | Qualitative |
| Smith et al. (2023a, b) | 240 | Australian College of Emergency Medicine examination | ChatGPT, GPT-4, Bard, Bing | Official practice questions exclusively accessible to registered trainees | | 63.3% | 48.8% (ChatGPT), 75.8 (GPT-4), 65.8% (Bard), 68.3% (Bing) | Quantitative |

**Table 3** (continued)

| Study | Sample (Number of questions) | Phenomenon of interest | Design | | | Evaluation | | Research Type |
|---|---|---|---|---|---|---|---|---|
| | | | LLM | Dataset | Prompt engineering technique | Passing score | Results | |
| Liu et al. (2023c) | 165 (Questions in Chinese) | Chinese Clinical Medicine Entrance Examination | ChatGPT | Online question bank | Inserting a brief preamble specifying that the question was either multi-choice or single-choice | 43.0% | 51.2% | Qualitative |
| Kumah-Crystal et al. (2023) | 254 | Clinical Informatics Board examination | ChatGPT | Mankowitz's Clinical Informatics Board Review book | Inserting a brief preamble requesting justification | 60.0% | 74.0% | Quantitative |
| Weng et al. (2023) | 125 (Questions in Chinese) | Family Medicine Board Examination | ChatGPT | Real exam (2022 edition) | Inserting a brief preamble specifying that the question was either multi-choice or single-choice | 60.0% | 41.6% | Qualitative |
| Skalidis et al. (2023) | 362 | European Exam in Core Cardiology | ChatGPT | 68 from European Society of Cardiology, 150 from Braunwald's Heart Disease Review and Assessment, and 144 from StudyPRN | | 60.0% | 58.8% | Quantitative |

**Table 3** (continued)

| Study | Sample (Number of questions) | Phenomenon of interest | Design | | | Evaluation | | Research Type |
|---|---|---|---|---|---|---|---|---|
| | | | LLM | Dataset | Prompt engineering technique | Passing score | Results | |
| Huyn et al. (2023) | 135 | American Urological Association Self-Assessment Study Program examination | ChatGPT | Real exam (2022 edition restricted to registered users) | Open end questions<br>Multiple choice questions | 50.0% | 26.7% (open-ended questions), 28.2% (multiple-choice question) | Qualitative |
| Li et al. (2023c) | 28–35 | Objective structured clinical examinations by Obstetrical and Gynaecological Society of Singapore | ChatGPT | Mock questions for preparation to the exam | | | 77.2% | Qualitative |
| Passby et al. (2023) | 84 | To simulate Specialty Certificate Examination in Dermatology | ChatGPT, GPT-4 | One online question bank for preparation to the exam | | 70.0%-70.2% | 63.1% (ChatGPT), 90.5% (GPT-4) | Quantitative |
| Suchman et al. (2023) | 455 | American College of Gastroenterology self-assessment tests | ChatGPT, GPT-4 | Real exam (2021–2022 editions) | | 70% | 65.1% (ChatGPT), 62.4% (GPT-4) | Quantitative |
| Hoch et al. (2023) | 2,576 | Preparation for German otolaryngology board certification | ChatGPT | One online question bank for preparation to the exam | Inserting a brief preamble specifying that the question was either multi-choice or single-choice | 60.0% | 57.0% | Quantitative |

**Table 3** (continued)

| Study | Sample (Number of questions) | Phenomenon of interest | Design | | | Evaluation | | Research Type |
|---|---|---|---|---|---|---|---|---|
| | | | LLM | Dataset | Prompt engineering technique | Passing score | Results | |
| Wang et al. (2023c) | 450 (Questions in Chinese and English) | Taiwanese Pharmacist Licensing Examination | ChatGPT | Real exam (2023 edition) | | 60.0% | First stage: 54.4% (in Chinese), 56.9% (in English); second stage: 53.8% (in Chinese), 67.6% (in English) | Quantitative |

*ABA* American Board of Anesthesiology, *FRCS* Fellowship of the Royal College of Surgeons, *OKAP* Ophthalmic Knowledge Assessment Program

included one report for each medical specialty. Question banks or mock-up versions were used in six studies, while questions on real examinations were used for the Australian College of Emergency Medicine, Family Medicine Board Examination, American Urological Association Self-Assessment Study Program examination, American College of Gastroenterology self-assessment tests, and Taiwanese Pharmacist Licensing Examination (Table 3). GPT-4, Bard, and Bing passed the Australian College of Emergency Medicine exam in contrast with ChatGPT which did not meet the threshold (Smith et al. 2023a, b). ChatGPT did not pass the Family Medicine Board Examination (with questions in Chinese), and the American Urological Association Self-Assessment Study Program examination (Huyhn et al. 2023; Weng et al. 2023). Neither ChatGPT nor GPT-4 passed the American College of Gastroenterology self-assessment tests (Suchman et al. 2023). On the Taiwanese Pharmacist Licensing Examination ChatGPT met the threshold only on the part on pharmaceutical laws (questions in English) but not in pharmacology (questions in both Chinese and English) (Wang et al. 2023c).

Different prompt engineering strategies were applied to medical specialty examinations. Zero-shot and formatting questions both open-end and multiple choice were used in the studies for the Ophthalmic Knowledge Assessment Program examination (Antaki et al. 2023; Mihalache et al. 2023). In the study on preparation for the American Board of Surgery Qualifying Exam, all questions were prompted as both open-end and multiple-choice single answer without forced justification, as done in previous studies on USMLE and Chinese National Medical Licensing Examination (Beaulieu-Jones et al. 2024; Fang et al. 2023; Kung et al. 2023). A brief preamble specifying that the question was either multi-choice or single-choice was used in the study on the Chinese Clinical Medicine Entrance Examination, Family Medicine Board Examination, and Preparation for German otolaryngology board certification (Hoch et al. 2023; Liu et al. 2023c; Weng et al. 2023). A brief preamble requesting justification for the generated responses was applied to questions for the Clinical Informatics Board examination (Kumah-Crystal et al. 2023). Questions were prompted as both open end and multiple choice on the study on the American Urological Association Self-Assessment Study Program examination (Huyhn et al. 2023). Overall, GPT-4 outperformed ChatGPT in all examinations except the American College of Gastroenterology self-assessment test (Suchman et al. 2023). GPT-4 scored higher than Bard on the American Board of Neurological Surgery oral board examination, American Board of Anesthesiology examination, and Australian College of Emergency Medicine examination (Ali et al. 2023a; Angel et al. 2024; Smith et al. 2023a, b).

### 4.5 Other studies

The remaining reviewed studies are reported in Table 4.

They concerned the UK BioMedical Admissions Test, the Clinical reasoning exams administered to pre-clerkship medical students at Stanford University, and a parasitology exam at Hallym University (South Korea), as shown in Table 4. In all three studies, the LLMs were assessed on real exams. However, in two of them the passing score was not specified (Giannos et al. 2023b; Huh et al. 2023; Strong et al. 2023). ChatGPT score slightly below the passing score of the Clinical reasoning exams at Stanford University, while for the other two studies the threshold was not specified (Table 4). No prompt engineering strategies were applied to these studies (Table 4).

**Table 4** Studies on other examinations

| Study | Sample (Number of questions) | Phenomenon of interest | Design | | | Evaluation | | Research Type |
|---|---|---|---|---|---|---|---|---|
| | | | LLM | Dataset | Prompt engineering technique | Passing score | Results | |
| Giannos et al. (2023b) | 180 | BioMedical Admissions Test | ChatGPT | Publicly available and official material (2019–2022 editions) | | No official threshold. It varies among institutions | 50%–66% (on Sect. 1), and 5%–45% (on Sect. 2) | Quantitative |
| Strong et al. (2023) | 64 | Clinical reasoning exams administered to pre-clerkship medical students at Stanford University | ChatGPT | Real exam | | 70.0% | 69.0% | Quantitative |
| Huh et al. (2023) | 79 | Parasitology examination at Hallym University | ChatGPT | Real examination | | | 60.8% | Quantitative |

### 4.6 Analysis of bias

The analysis of bias is reported in Table 5.

To reduce memory retention bias, a new chat session of ChatGPT was started for each question in two studies on USMLE, one on the Korean National Licensing Examination for Korean Medicine Doctors, one on the Chinese National Medical Licensing Examination, two on orthopedics examination, one on ChatGPT on ophthalmology, one on the Chinese Clinical Medicine Entrance Examination, and one on American Urological Association Self-Assessment Study Program examination (Table 5). A new chat session of GPT-4 was started for each question in one study on the Korean National Licensing Examination for Korean Medicine Doctors, and one on the Chinese National Medical Licensing Examination (Fang et al. 2023; Jang et al. 2023). A new chat session of ChatGPT and GPT-4 was started after five questions in one study on the American College of Radiology Radiation Oncology in-training examination (Huang et al. 2023). The other studies did not specify whether the queue of chats of the LLM was cleared or not.

The high score of GPT-4, Med-PaLM, and Med-PaLM 2 on USMLE suggests the hypothesis of memorization effect, which can arise when benchmark data are included in an LLM training set. However, specific tests using the Memorisation effects Levenshtein detector (MELD) method revealed no evidence of memory effect of GPT-4 on USMLE questions (Nori et al. 2023). No overlap was found between USMLE like questions of MultiMedQA and Med-PaLM training set, while little overlap was observed on MCQs (Singhal et al. 2023a). An overlap ranging from 0.9 to 11.15% was found between USMLE like questions of MedQA and Med-PaLM 2 (Singhal et al. 2023b). The other studies did not check overlap between test set and training set of the LLMs. Finally, a source of bias could be funding from private companies developing LLMs which occurred in four qualitative studies (Nori et al. 2023; Singhal et al. 2023a; Singhal et al. 2023b; Tu et al. 2023).

## 5 Comparative analysis on fundamentals of robotic surgery

### 5.1 Description of FRS

FRS was developed through four consensus conferences to establish the tasks, metrics, and curriculum content by involving 66 subject matters experts (surgeons, psychologists, psychometricians, engineers, simulation experts, and medical educators) from the Department of Defense and Veterans Administration, the American Board of Surgery, and 14 international surgical societies (Satava et al. 2020). FRS was designed to be agnostic to any particular platform and therefore it applies to basic robotic skills independent to the platform used. The FRS section on cognitive skills consists of online modules with educational videos and text. At the end of this part, trainees must pass successfully a questionnaire before progressing to the technical skills part consisting of a set of tasks of increasing difficulty on a virtual reality simulator (Satava et al. 2020). Learners must reach proficiency in each task before moving to the next.

**Table 5** Analysis of bias in the reviewed studies

| | Memory retention bias | Overlap between training and test data used by LLM | Funding | | | | |
|---|---|---|---|---|---|---|---|
| | | | Private | Public | Private and public | Not reported | Not received |
| Kung et al. (2023) | | | | | | X | |
| Gilson et al. (2023) | | | | | | X | |
| Sharma et al. (2023) | | | | | | X | |
| Singhal et al. (2023a) | A new chat was started after each question | | | | | | X |
| Singhal et al. (2023b) | | | | | | X | |
| Tu et al. (2023) | A new chat was started after each question | | | | | X | |
| Nori et al. (2023) | | No to minimal overlap between MultiMedQA and PaLM-Med training | X | | | | |
| Toma et al. (2023) | | 0.9%-11.15% overlap between MedQA and PaLM-Med 2 training | X | | | | |
| Han et al. (2023) | | | X | | | | |
| Wu et al. 2023 | | No overlap between USMLE questions and GPT-4 training data, assessed by Memorisation effects Levenshtein detector (MELD) method | X | | | | |
| Thirunavukarasu et al. (2023) | | | | | | X | |
| Takagi et al. (2023) | | | | | | X | |
| Kasai et al. (2023) | | | | | | X | |
| Taira et al. (2023) | | | | X | | | |
| Jang et al. (2023) | A new chat was started after each question | | | X | | | |
| Fang et al. (2023) | A new chat was started after each question | | | | | X | |

**Table 5** (continued)

| | Memory retention bias | Overlap between training and test data used by LLM | Funding | | | | |
|---|---|---|---|---|---|---|---|
| | | | Private | Public | Private and public | Not reported | Not received |
| Giannos et al. (2023a) | | | | | | X | |
| Ali et al. (2023a) | | | | | | | X |
| Ali et al. (2023b) | | | | | | X | |
| Hopkins et al. (2023) | | | | | | X | |
| Cuthbert et al. (2023) | A new chat was started after each question | | | | | | X |
| Saad et al. (2023) | | | | | | X | |
| Lum et al. (2023) | A new chat was started after each question | | | | | | X |
| Angel et al. (2024) | | | | X | | | |
| Shay et al. (2023) | | Memorisation effects Levenshtein detector (MELD) method | | | | | X |
| Antaki et al. (2023) | A new chat was started after each question | | | | | | X |
| Mihalache et al. (2023) | | | | | | | X |
| Beaulieu-Jones et al. (2024) | | | | | | X | |
| Oh et al. (2023) | | | X | | | | |
| Bhayana et al. (2023) | | | | | | X | |
| Huang et al. (2023) | A new chat was started after five questions | | | | | X | |
| Smith et al. (2023a, b) | | | | | | X | |
| Liu et al. (2023c) | A new chat was started after each question | | | | | | X |
| Kumah-Crystal et al. (2023) | | | | | | | X |

**Table 5** (continued)

| | Memory retention bias | Overlap between training and test data used by LLM | Funding | | | | |
|---|---|---|---|---|---|---|---|
| | | | Private | Public | Private and public | Not reported | Not received |
| Weng et al. (2023) | | | | | | X | |
| Skalidis et al. (2023) | | | | | | | X |
| Huyn et al. (2023) | A new chat was started after each question | | | | | X | |
| Li et al. (2023c) | | | | | | | X |
| Passby et al. (2023) | | Memorisation effects Levenshtein detector (MELD) method | | | | | X |
| Suchman et al. (2023) | | | | | | | X |
| Hoch et al. (2023) | | | | | | | X |
| Wang et al. (2023c) | | | | | | X | |
| Giannos et al. (2023b) | | | | | | X | |
| Strong et al. (2023) | | | | | | | X |
| Huh et al. (2023) | | | | | | | X |

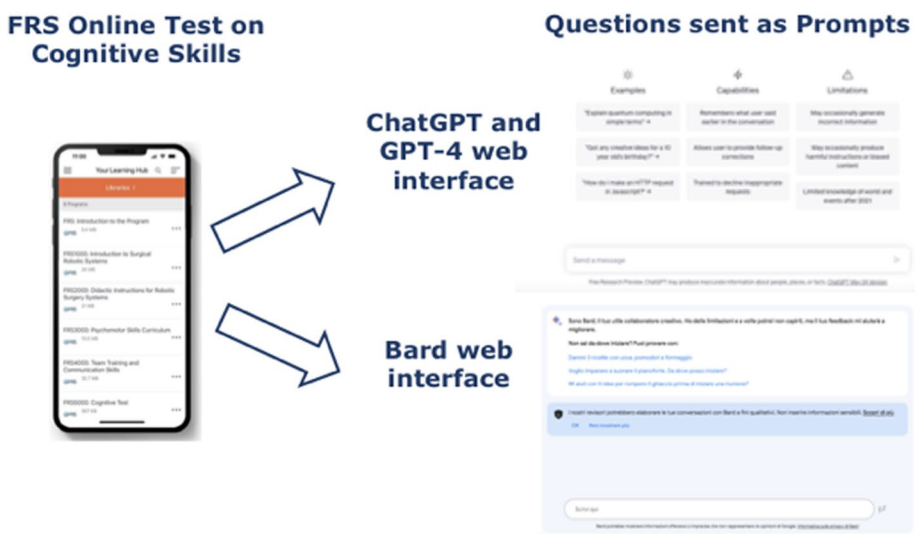


**Fig. 2** Protocol of experiments (all LLMs generated answers were stored in Word files)

### 5.2 Input source of questions

The FRS part on cognitive skills consists of four online modules providing basic knowledge. It starts with an introduction to surgical robotic systems, then moves on to didactic instructions on robotic surgery, psychomotor skills curriculum, and ends with team training and communication skills. It can be accessed after downloading an app available for iOS and Android devices.[9] At the end of the modules, the learners are required to pass a test consisting of 44 multiple choice questions (MCQs), each with four options, except one with three, and one with two. The breakdown of the questions is: i) introduction to surgical robotic systems (n = 18), ii) didactic instructions for robotic surgery (n = 13), iii) psychomotor skills curriculum (n = 6), and iv) team training and communication skills (n = 7). The proficiency level for passing the questionnaire is equivalent to 35 correct answers (79.5%) (Satava et al. 2020).

### 5.3 LLMs testing

The used protocol involved manually prompting ChatGPT, GPT-4, and Bard web interface with all the original MCQs of FRS (Fig. 2).

We chose this technique as it is the closest to human test-taking. Only one of the 44 MCQs included both text and an image. Even though in our study ChatGPT could only manage text information, this question was retained. The questions and answers obtained by ChatGPT, GPT-4, and Bard were saved in Microsoft Word files. All MCQs were analyzed manually to determine the selected response, which was marked as correct, incorrect,

[9] https://www.surgicalexcellence.org/frs-registration

and not selected option (i.e., when the LLM did not choose an answer). Not selected answers were considered incorrect.

Data were then imported into a Microsoft Excel spreadsheet. After FRS test was completed seven times with the January 30, 2023 release, a new version of ChatGPT became available online. The FRS test was then repeated seven times with the following 2023 versions: February 13, March 14, May 3, and May 24. Likewise, FRS test was submitted seven times to GPT-4 (March 14, 2023 releases) and Bard (July 13, 2023 release). After completing a full questionnaire, the queue of prompts and answers were cleared to avoid memory effect on subsequent trials. Statistically significant difference was tested with Kruskall-Wallis and Wilcoxon test ($p < 0.001$ for statistically significant difference).

### 5.4 Performances of the five different releases of ChatGPT

The results comparing the five releases of ChatGPT on FRS test are reported in Table 6.

The performances of ChatGPT over trials is shown in Fig. 3. On the first attempt (baseline) the third release achieved the highest score (79.5%) vs 77.3% of the fifth and 72.7% for the fourth, followed by 54.5% for first two releases. On baseline, only the third version reached the proficiency level for passing FRS test. The average score of correct answers is depicted in Fig. 3, improving slightly from the first (64.6%) to second release (65.6%), but more substantially by the third (75.0%) and fourth (78.9%) releases. Surprisingly the mean score of the fifth tested version of ChatGPT decreased to 72.7%.

Kruskal–Wallis tests confirmed statistically significant difference ($p < 0.001$). ChatGPT was not able to achieve the benchmark in any attempts with the first two and fifth releases but reached the proficiency level on two trials with the third release, and on three attempts with the fourth. ChatGPT answered correctly the question with both text and image three, six, two, and four times from the first to the fourth tested version, respectively. It always provided an erroneous response with the fifth version. In the second release it had the highest average rate of answers without choosing any option (14.6%), followed by the third (7.8%), first and fourth (6.5%), and fifth (3.2%) as shown in Fig. 4).

### 5.5 Performances of GPT-4

Scores of GPT-4 are shown in Table 7 and Fig. 3.

At the baseline, it always outperformed the five releases of ChatGPT (Table 6). GPT-4 successfully passed the FRS test on all seven attempts with an average of correct answers of 91.5%. It achieved a maximum of 95.4% in the sixth trial. It always performed better than all five tested versions of ChatGPT in each of the corresponding attempt, with the difference being statistically significant ($p < 0.001$). In sharp contrast with ChatGPT, it did not generate any response without choosing any option, except for the question containing both text and image (Fig. 5). In this case, it is always specified its inability to interpret images since this functionality was not available at the time of testing. Moreover, GPT-4 provided concise answers without additional explanations.

**Table 6** Comparison among different releases of ChatGPT on FRS test

| | Trial Number | Correct answers (score) (%) | Wrong answers (%) | No answer selected (%) | Answer to question with image | Correct answers after removing the question with image (score) (%) |
|---|---|---|---|---|---|---|
| ChatGPT (January 30, 2023 release) | 1 | 54.5 | 40.9 | 4.6 | Erroneous | 55.8 |
| | 2 | 61.3 | 36.4 | 2.3 | Erroneous | 62.8 |
| | 3 | 68.2 | 31.8 | 0.0 | Correct | 67.4 |
| | 4 | 61.3 | 34.1 | 4.6 | Erroneous | 62.8 |
| | 5 | 68.2 | 20.4 | 11.4 | Correct | 67.4 |
| | 6 | 72.7 | 18.2 | 9.1 | Erroneous | 74.4 |
| | 7 | 65.9 | 20.5 | 13.6 | Correct | 65.1 |
| ChatGPT (February 13, 2023 release) | 1 | 54.5 | 25.0 | 20.5 | Correct | 53.5 |
| | 2 | 65.9 | 22.7 | 11.4 | Correct | 65.1 |
| | 3 | 65.9 | 20.5 | 13.6 | Correct | 65.1 |
| | 4 | 72.7 | 15.9 | 11.4 | Erroneous | 74.4 |
| | 5 | 65.9 | 22.7 | 11.4 | Correct | 65.1 |
| | 6 | 72.7 | 9.1 | 18.2 | Correct | 72.1 |
| | 7 | 61.4 | 22.7 | 15.9 | Correct | 60.5 |
| ChatGPT (March 14, 2023 release) | 1 | 79.5 | 20.5 | 0.0 | Erroneous | 81.4 |
| | 2 | 75.0 | 18.2 | 6.8 | Erroneous | 76.7 |
| | 3 | 79.5 | 15.9 | 4.6 | Erroneous | 81.4 |
| | 4 | 70.4 | 11.4 | 18.2 | Correct | 69.8 |
| | 5 | 72.7 | 20.5 | 6.8 | Erroneous | 74.4 |
| | 6 | 72.7 | 15.9 | 11.4 | Erroneous | 74.4 |
| | 7 | 75.0 | 18.2 | 6.8 | Correct | 74.4 |

**Table 6** (continued)

| | Trial Number | Correct answers (score) (%) | Wrong answers (%) | No answer selected (%) | Answer to question with image | Correct answers after removing the question with image (score) (%) |
|---|---|---|---|---|---|---|
| ChatGPT (May 3, 2023 release) | 1 | 72.7 | 20.5 | 6.8 | Correct | 72.1 |
| | 2 | 77.2 | 11.4 | 11.4 | Correct | 76.7 |
| | 3 | 84.1 | 11.4 | 4.5 | Correct | 83.7 |
| | 4 | 84.1 | 11.4 | 4.5 | Correct | 83.7 |
| | 5 | 75.0 | 18.2 | 6.8 | Erroneous | 76.7 |
| | 6 | 81.8 | 13.7 | 4.5 | Erroneous | 83.7 |
| | 7 | 77.2 | 15.9 | 6.9 | Erroneous | 79.1 |
| ChatGPT (May 24, 2023 release) | 1 | 77.2 | 20 | 2.3 | It did not select any answer | 79.1 |
| | 2 | 75.0 | 25.0 | 0.0 | It did not select any answer | 76.7 |
| | 3 | 70.4 | 29.6 | 0.0 | It did not select any answer | 72.1 |
| | 4 | 72.7 | 27.3 | 0.0 | It did not select any answer | 74.4 |
| | 5 | 65.9 | 34.1 | 0.0 | It did not select any answer | 67.4 |
| | 6 | 72.7 | 25.0 | 2.3 | It did not select any answer | 74.4 |
| | 7 | 75.0 | 22.7 | 2.3 | It did not select any answer | 76.7 |

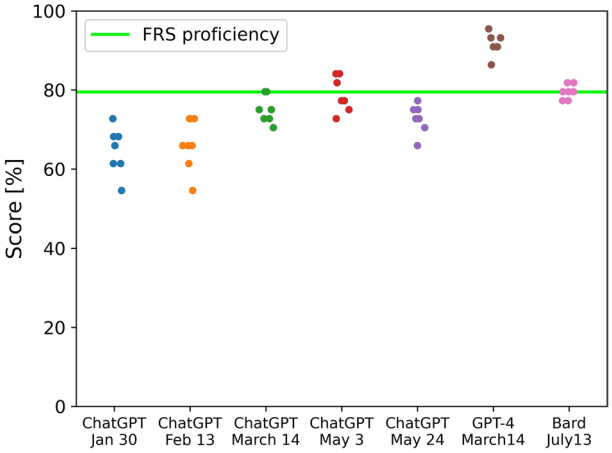


**Fig. 3** Comparison on trend of the five different releases of ChatGPT, GPT-4, and Bard on FRS test

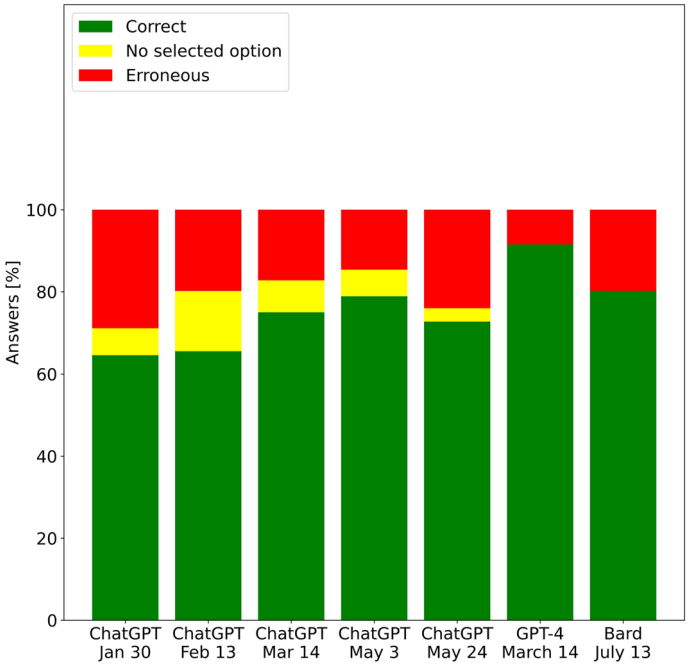


**Fig. 4** Rate of correct, erroneous, and answers without option selected for the five releases of ChatGPT, GPT-4, and Bard on the entire FRS test

## 5.6 Performances of Bard

Scores of Bard are reported in Table 8 and in Fig. 3.

Bard successfully passed FRS test on five out of seven attempts with an average of correct answers of 79.5%. It achieved a maximum of 81.8% in the third and sixth trials (Table 8). GPT-4 always performed better than Bard in each of the corresponding attempts,

**Table 7** Results of GPT-4 on FRS test

| | Trial Number | Correct answers (score) (%) | Wrong answers (%) | No answer selected | Answer to question with image | Correct answers after removing the question with image (score) (%) |
|---|---|---|---|---|---|---|
| GPT-4 (March 14, 2023 release) | 1 | 90.9 | 9.1 | 0 | Erroneous | 93.0 |
| | 2 | 93.2 | 6.8 | 0 | Erroneous | 95.3 |
| | 3 | 86.4 | 13.6 | 0 | Erroneous | 88.4 |
| | 4 | 93.2 | 6.8 | 0 | Erroneous | 95.4 |
| | 5 | 90.9 | 9.1 | 0 | Erroneous | 93.0 |
| | 6 | 95.4 | 4.6 | 0 | Erroneous | 97.7 |
| | 7 | 90.9 | 9.1 | 0 | Erroneous | 93.0 |

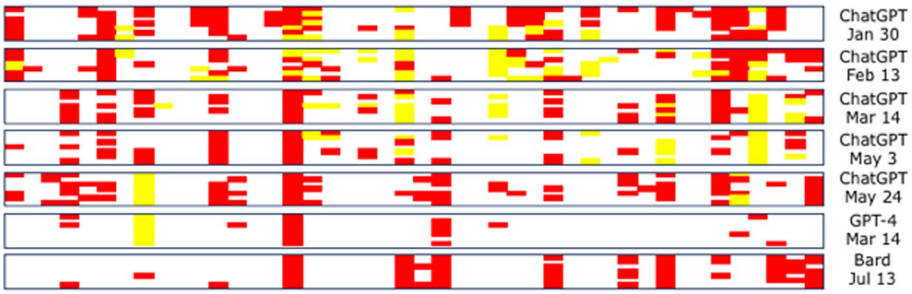


**Fig. 5** Raw marks for each attempt of all tested LLMs with correct answers in white, erroneous in red, and not selected answers in yellow

but without being statistically significant ($p = 0.002$). Bard always provided explanations for the response it generated and was able to answer correctly the question with text and image in six out of seven trials.

### 5.7 Consistency over multiple attempts and releases

In addition to showing variability in the overall score among attempts and in the case of ChatGPT also among releases, the LLMs demonstrated variability in the number of times each question was answered correctly or not, as depicted in Fig. 5.

The first release of ChatGPT provided the correct answer to 17 questions (38.6%) in all seven trials, vs 13 (29.5%) of the second vs 25 (56.8%) of the third vs 23 (52.3%) of the fourth vs 21 (47.7%) of the fifth vs 35 for GPT-4 (79.5%) vs 30 for Bard (68.2%). ChatGPT generated an erroneous answer in all attempts in two questions (4.5%) of the second, third, and fourth version vs one (2.2%) of the fifth releases vs one (2.2%) of GPT-4 vs four of Bard (9.1%).

### 5.8 Answers related to the knowledge domain

The answers provided by ChatGPT, GPT-4, and Bard relating to the knowledge domain are shown in Table 9 and Figs. 6, 7, 8, 9.

GPT-4 outperformed all versions of ChatGPT in all domains by a substantial margin. It achieved the highest rate of correct answers on team training and communication skills (100.0%) vs. 85.7% for Bard, while ChatGPT ranged from 77.5 to 91.8%. On questions on the robotic system GPT-4 obtained 96.7% vs 73.0% for Bard, while the correct answers of ChatGPT ranged from 62.7 to 73.1%. On the clinical steps involved in a procedure of RAS, GPT-4 answered correctly on 84.6% of cases vs. 80.2% for Bard, whereas the correct answers of ChatGPT ranged from 53.8 to 79.1%. On psychomotor skills GPT-4 obtained 80.9% vs. 83.3% for Bard, while correct answers of ChatGPT ranged from 57.1 to 76.2%.

**Table 8** Results of Bard on FRS test

| | Trial Number | Correct answers (score) (%) | Wrong answers (%) | No answer selected | Answer to question with image | Correct answers after removing the question with image (score) (%) |
|---|---|---|---|---|---|---|
| Bard (July 13, 2023 release) | 1 | 79.5 | 20.5 | 0 | Correct | 79.1 |
| | 2 | 79.5 | 20.5 | 0 | Correct | 79.1 |
| | 3 | 81.8 | 18.2 | 0 | Correct | 81.4 |
| | 4 | 79.5 | 20.5 | 0 | Correct | 79.1 |
| | 5 | 77.3 | 23.7 | 0 | Erroneous | 79.1 |
| | 6 | 81.8 | 18.2 | 0 | Correct | 81.4 |
| | 7 | 77.3 | 22.7 | 0 | Correct | 76.7 |

**Table 9** Breakdown of questions topic

| | | | Correct answers (%) | Wrong answers (%) | It did not select any option (%) |
|---|---|---|---|---|---|
| Introduction to surgical robotic systems | ChatGPT | January 20, 2023 release | 67.5 | 8.7 | 23.8 |
| | | February 13, 2023 release | 62.7 | 13.5 | 23.8 |
| | | March 14, 2023 release | 69.8 | 11.9 | 18.3 |
| | | May 3, 2023 release | 74.6 | 14.3 | 11.1 |
| | | May 24, 2023 release | 75.4 | 22.2 | 2.4 |
| | GPT-4 | March 14, 2023 release | 96.8 | 3.2 | 0.0 |
| | Bard | July 13, 2023 release | 73.0 | 27.0 | 0.0 |
| Didactic instructions for robotic surgery | ChatGPT | January 20, 2023 release | 53.8 | 5.5 | 40.7 |
| | | February 13, 2023 release | 65.9 | 15.4 | 18.7 |
| | | March 14, 2023 release | 73.8 | 6.0 | 20.2 |
| | | May 3, 2023 release | 79.1 | 2.2 | 16.7 |
| | | May 24, 2023 release | 64.8 | 27.5 | 7.7 |
| | GPT-4 | March 14, 2023 release | 84.6 | 7.7 | 8.7 |
| | Bard | July 13, 2023 release | 80.2 | 19.8 | 0.0 |
| Psychomotor skills | ChatGPT | January 20, 2023 release | 64.3 | 4.8 | 30.9 |
| | | February 13, 2023 release | 57.1 | 16.7 | 26.2 |
| | | March 14, 2023 release | 76.2 | 2.4 | 21.4 |
| | | May 3, 2023 release | 76.2 | 0.0 | 23.8 |
| | | May 24, 2023 release | 69.0 | 31.0 | 0.0% |
| | GPT-4 | March 14, 2023 release | 80.9 | 19. | 0.0 |
| | Bard | July 13, 2023 release | 83.3 | 16.7 | 0.0 |

**Table 9** (continued)

| | | | Correct answers (%) | Wrong answers (%) | It did not select any option (%) |
|---|---|---|---|---|---|
| Team training and communication skills | ChatGPT | January 20, 2023 release | 77.5 | 4.1 | 18.4 |
| | | February 13, 2023 release | 77.5 | 14.3 | 8.2 |
| | | March 14, 2023 release | 83.7 | 6.1 | 10.2 |
| | | May 3, 2023 release | 91.8 | 8.2 | 0.0 |
| | | May 24, 2023 release | 85.7 | 14.3 | 0.0 |
| | GPT-4 | March 14, 2023 release | 100.0 | 0.0 | 0.0 |
| | Bard | July 13, 2023 release | 85.7 | 14.3 | 0.0 |

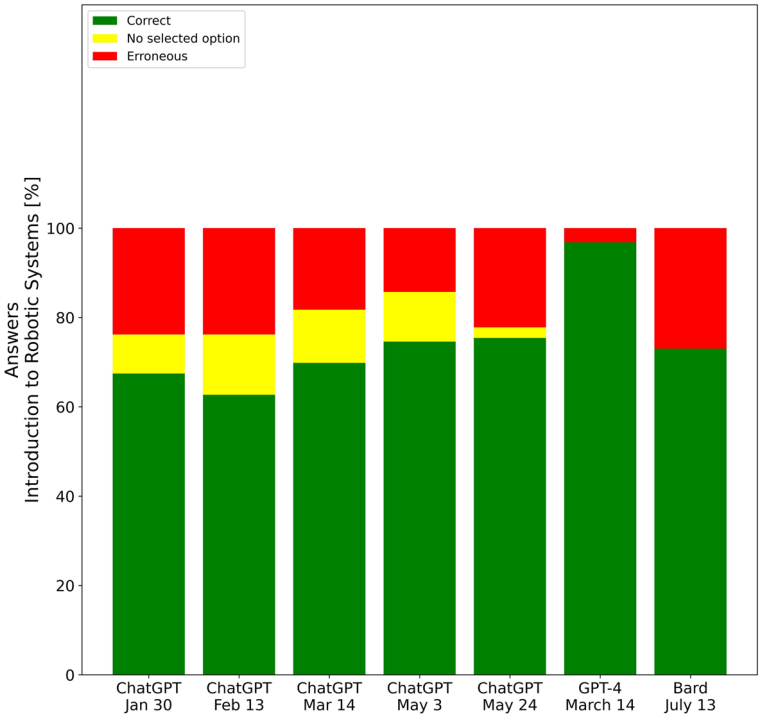


**Fig. 6** Rate of correct, erroneous, and answers without selected option for the three releases of ChatGPT, GPT-4, and Bard on domain specific questions of the FRS test (Introduction to surgical robotic systems)

## 6 Discussion

### 6.1 Main findings

Launched in November 2022, ChatGPT is a generic LLM trained on information available on the Internet until the end of 2021. It was released free to users for testing, and immediately generated a viral interest, reaching 100 million users after the first two months, representing the fastest hike in a consumer Internet app, before the launch of Threads, the microblogging app by Meta, in July 2023.[10] Since then, new LLMs have been launched by giants like Google and Meta, start-ups like Anthropic and Hippocratic AI, and research groups.

In this systematic review, a comprehensive search strategy on a wide array of search terms was performed on PubMed, Web of Science, Scopus, and arXiv. The choice of arXiv was motivated by the need the discover studies in a preprint format, missing in the other databases. Additionally, arXiv is used by an increasing number of research groups publishing their efforts in computer science applications, including LLMs. The literature search was strengthened by using the SPIDER tool, which was used also to formulate the research questions. The results of this review have shown that a generic LLMs like GPT-4 is capable

[10] https://www.zdnet.com/article/threads-hit-100-million-users-in-under-a-week-breaking-chatgpts-record/

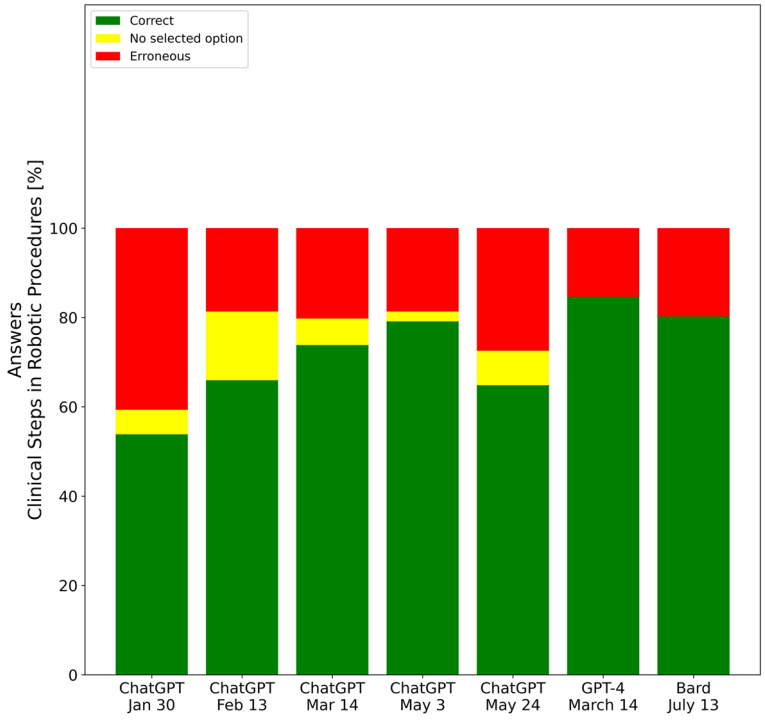


**Fig. 7** Rate of correct, erroneous, and answers without option selected for the three releases of ChatGPT, GPT-4, and Bard on domain-specific questions of the FRS test (Clinical steps in robot-assisted surgery procedures)

to pass national qualifying examinations like USMLE and others with questions in a different language from English using simple prompt engineering strategies like zero-shot and few-shot learning (Fang et al. 2023; Kasai et al. 2023; Nori et al. 2023; Takagi et al. 2023). It achieved a score slightly below the threshold only for the Korean National Licensing Examination for Korean Medicine Doctors (Jang et al. 2023). Novel LLMs designed specifically for the healthcare domain, namely Med-PaLM, and Med-PaLM-2 passed USMLE thanks to refined prompt engineering techniques like chain of thought, self-consistency, and ensemble refinement (Singhal et al. 2023a; Singhal et al. 2023b).

In the present study, we reported the performance over time of three different LLMs on a test for surgical education: ChatGPT and GPT-4 by OpenAI, and Bard by Google. They were assessed on the standardized cognitive questionnaire of FRS, consisting of four knowledge domains, which has been adopted by an increasing number of surgical training and education centers in the United States and the European Union. In total, in the present study a total of 2,156 answers, generated by LLMs, were analyzed. Like the study by Antaki et al. on the Ophthalmic Knowledge Assessment Program examination we prompted questions in the original form since this technique is the closest to human test-taking.

As in the recent study by Jang et al. on the Korean National Licensing Examination for Korean Medicine Doctors we assessed LLMs on multiple attempts to evaluate their consistency. In contrast with the study by Jang et al., we performed seven trials instead

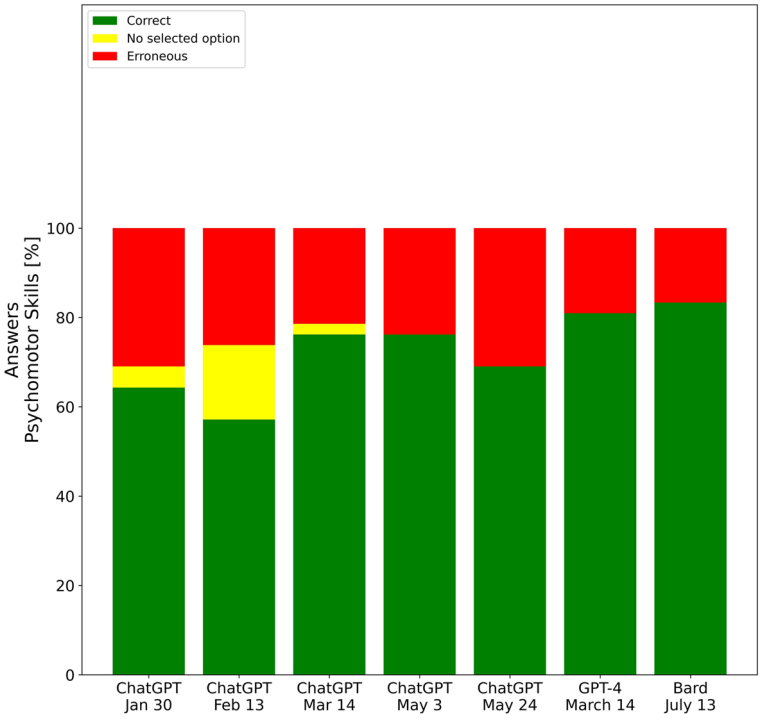


**Fig. 8** Rate of correct, erroneous, and answers without selected option for the three releases of ChatGPT, GPT-4, and Bard on domain-specific questions of the FRS test (Psychomotor skills)

of five. Like the study by Mihalache et al. on the Ophthalmic Knowledge Assessment Program examination we tested several ChatGPT releases (in our case five instead of two).

Although it is expected that LLMs improve performances over time, there is no published evidence quantifying progress/ improvement of LLMs on surgery, and, to a broader extent, on the medical domain. In fact, in the study of Mihalache et al., the questions were prompted in different modalities between releases, namely as MCQs with the first and as open end with the second one (Mihalache et al. 2023).

Our findings demonstrated that the mean performance of ChatGPT on the FRS test improved from 64.6% to 78.9% from the first to the fourth tested release, but unexpectedly dropped to 72.7% with the fifth version. In particular, ChatGPT was unable to pass the FRS test in any of the seven trials with the first two versions and the fifth one. In contrast, with the third and fourth releases it passed the FRS test, although not consistently.

The results of the present study confirmed that GPT-4 outperformed ChatGPT and Bard in every attempt, thus in agreement with the results of our systematic review (Ali et al. 2023a; Ali et al. 2023b; Angel et al. 2024; Giannos et al. 2023a; Huang et al. 2023; Jang et al. 2023; Kasai et al. 2023; Oh et al. 2023; Passby et al. 2023; Smith et al. 2023a, b). The difference of the performance between ChatGPT and GPT-4 is in agreement with the reports included in our systematic review on the National Medical Practitioners Qualifying Examination in Japan, Korean National Licensing Examination for Korean Medicine Doctors, the UK Specialty Certificate Examination in Neurology, the American Board of Neurological Surgery board examination (both oral and written part), the Korean

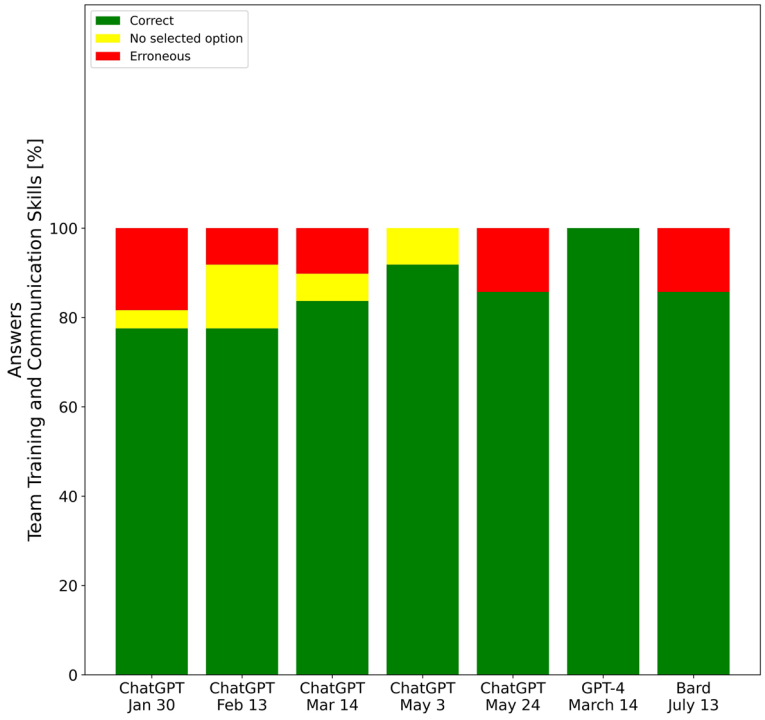


**Fig. 9** Rate of correct, erroneous, and answers without option selected for the three releases of ChatGPT, GPT-4, and Bard on domain-specific questions of the FRS test (Team training and communication skills)

general surgery board exams, the American College of Radiology Radiation Oncology in-training examination, the Australian College of Emergency Medicine examination, and the Specialty Certificate Examination in Dermatology (Ali et al. 2023a; Ali et al. 2023b; Giannos et al. 2023a; Huang et al. 2023; Jang et al. 2023; Kasai et al. 2023; Oh et al. 2023; Passby et al. 2023; Smith et al. 2023a, b).

Of the four domains on cognitive skills of FRS, all LLMs achieved the highest score on team training and communication skills, probably because this is a topic with a large amount of information publicly available online. Bard was the only one of the tested LLMs able to answer multimodal questions with mixed text and images. Although GPT-4 is equipped with the same functionality, unfortunately, it was not available at the time of testing.

LLMs learn the statistical patterns of language on massive datasets of online text. They may produce errors and misleading information, especially for technical topics on which they have been trained only with small datasets (Stokel-Walker et al. 2023). In this regard, the present study has identified questions that were consistently answered erroneously, e.g., ChatGPT provided an incorrect response in 24 out of 35 attempts (68.6%) to one MCQ, indicating that the current RAS systems are semi-autonomous, performing independently part of an operation while the surgeon performs a different part. We believe that this may depend on some bias within the training dataset. It may well refer to one study published in 2016 reporting a prototype of an autonomous surgical robot performing anastomosis on animal tissue (Shademan et al. 2016).

However, this robot has never been adopted for clinical practice on humans. GPT-4 answered wrongly in all seven attempts on a question on errors committed during a knot tying task. Bard provided an erroneous response on two questions on the ergonomics of the surgeon console, and one on communication skills in all seven trials.

Unfortunately, none of these LLMs provide references to support the generated answers, despite the consensus that the sources of information should always be verifiable (Stokel-Walker et al. 2023).

Even though the FRS test contains knowledge available before 2021 and ChatGPT was trained with data until 2021, the findings of the present study demonstrated that three 2023 releases of ChatGPT were unable to pass the FRS test in any trial, while it reached the benchmark only with the third and fourth version.

At present, there are no studies on LLMs, focused on the biomedical domain, on the FRS test. Recent evidence has shown that LLMs designed specifically for healthcare like Med-PaLM, Med-PaLM 2, and Med-PaLM M outperformed PubMedGPT on USMLE (Singhal et al. 2023a; Singhal et al. 2023b; Tu et al. 2023).

Overall, by considering that ChatGPT, GPT-4, and Bard are considered generic LLMs, we believe that their scores, observed by the present study on FRS, represent a remarkable result. The impressive performances of LLMs on competency examinations may contribute to the perception that artificial intelligence forays in healthcare will eventually devalue human intelligence (Nori et al. 2023). Additionally, the LLMs growing prowess may influence decisions of about medicine as a career path, and, for medical students, their choice of specialty (Nori et al. 2023). According to a recent survey among 32 medical schools in the United States, artificial intelligence had a negative impact on the choice of radiology as a career path among medical students (Reeder et al. 2022).

### 6.2 Limitations

We acknowledge some limitations in the present work. In the systematic review published studies in non-English languages were excluded. The paucity of the studies on the same examination prevented to compare their results, except those on USMLE. Additionally, since official questions of medical examinations are not generally freely available, most studies used databases with surrogate questions.

The most important limitation on the comparative study on FRS test was the inability to compare the scores of surgical trainees with the performances of LLMs. Secondly, the questions of the FRS test were submitted in the original form, without prompt engineering strategies to elicit some form of reasoning, to increase the probability of LLMs to generate the correct answer. However, we selected this technique as it is the closest to human test-taking. We are aware of the importance of prompt engineering to guide LLMs to generate more accurate output text. In a future study, we will investigate the role of prompt engineering to help trainees in the preparation of surgical examinations.

### 6.3 Open challenges

The hype behind LLMs has led to unwarranted speculations on their potential to transform medical education at different stages, including preparation for medical examinations. Firstly, they may be used to conduct needs assessments where they may help teachers to identify content gaps in education (Abd-Alrazaq et al. 2023). Secondly, they may develop

measurable learning objectives and tailor the curriculum to meet the diverse needs of trainees. Thirdly, they may help instructors in preparing teaching materials (e.g., written simulated case reports, and contents of lectures) (Lee 2023a). A recent study reported positively on the application of ChatGPT to simulate standardized patients (Liu et al. 2023d), supporting the belief that in the future LLMs may be helpful in designing clinical scenarios for surgical simulations by integrating medical imaging, electronic health records, and virtual reality contents. Furthermore, LLMs can play the role of tutors by providing trainees with real-time and customized feedback, identifying areas of strength and weakness, and offering targeted suggestions for improvement (Abd-Alrazaq et al. 2023; Lee et al. 2023a). This could be helpful in the self-study phase before taking a real examination. Alternatively, LLMs may be employed as mentors to explain difficult topics in simple terms, thus streamlining the education process for struggling trainees (Abd-Alrazaq et al. 2023).

However, there remain some challenges that need to be addressed, the most critical being ensuring the accuracy and reliability of the information generated by LLMs. Since they predict the probability distribution of text, the risk of getting misleading answers is significant, as highlighted by the present study. Due to its non-deterministic nature, the text generated by LLMs can change over time, thus leading to confusion in some scenarios, e.g., students obtaining a different response to the same question over time, or students within the same class obtaining a different response to the same question asked at the same time.

The present study has identified different responses to the same questions of the FRS curriculum, especially for ChatGPT. In some instances, they were even contradictory. As a result, ChatGPT was unable to confirm proficiency on the FRS test after achieving it once. In contrast, Bard and GPT-4 showed a lower variability on FRS testing. Currently, the present study indicates that they do not possess the required reliability to act as mentors of trainees in complex subjects exemplified by surgery, despite the huge information freely accessible on the Internet, which might have been used to train LLMs. Although the improvement of LLMs performances, as demonstrated by our research, supports the belief that their potential in the medical education sector is vast, human expertise and guidance remain essential. In essence, the take home message from the present study is that human experts should always check and scrutinize the artificial intelligence generated content before approval to ensure the highest efficacy and reliability before the integration of LLMs within future surgical education.

The introduction of LLMs in healthcare education might replicate the path traced by simulation, which, after initial skepticism as in the case of surgery, has become an integral part. Simulation has represented a paradigm shift in the training of healthcare professionals allowing trainees to practice and enact errors in a risk-free environment for patient safety. For instance, in surgical simulation, trainees are allowed to commit errors and learn from them in sharp contrast to actual surgery (Gallagher and O'Sullivan 2012). This is the main strength of surgical simulation and is the main reason for becoming an integral component of surgical training programs. However, it is imperative to validate surgical simulation in order to demonstrate their effectiveness (Zevin et al. 2012). Likewise, LLMs may become a new tool in the armamentarium of the next generation of medical trainees in the different stages of the learning process, including preparation to real examinations, provided that their validity is demonstrated. Guidelines on the use of LLMs in medical education should therefore be developed to ensure safety, reliability, efficacy, and privacy protection.

## 7 Conclusions

In this work, the authors presented the first systematic review on LLMs on medical examinations. The results have shown that GPT-4 passed by a large margin several national qualifying examinations including USMLE and others with questions in Chinese and Japanese using zero shot and few shot learning. Med-PaLM 2 obtained similar scores on USMLE using a more refined prompt engineering approach like ensemble refinement. GPT-4 outperformed ChatGPT and Bard on several medical specialties examinations, namely the National Medical Practitioners Qualifying Examination in Japan, the Korean National Licensing Examination for Korean Medicine Doctors, UK Specialty Certificate Examination in Neurology, American Board of Neurological Surgery board examination, the American Board of Anesthesiology examination, Korean general surgery board exams, Australian College of Emergency Medicine examination, and the Specialty Certificate Examination in Dermatology.

Our findings on FRS tests have shown that performances of ChatGPT improved from the initial release, although this trend was reversed with the latest tested version. GPT-4 showed impressive performance in passing FRS test outperforming ChatGPT and Bard in all seven trials. The 95.4% of correct answers to FRS questionnaire represent the highest score by GPT-4 in a high-stake examination in surgery. In comparison Bard reached 81.8% as maximum score on FRS test.

Hence, it seems more than likely that LLMs will continue to improve their performance in medical examinations. In addition to collecting larger datasets with more updated information and integrating a search with the latest available data on the Internet, research should focus on improving RLHF to reduce the risk of generating harmful output, and on prompt engineering to improve reasoning capabilities of LLMs to solve challenging unmet needs in healthcare.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s10462-024-10849-5.

**Author Contribution** Study concept and design: A.M., K.G., P.C., L.M., R.S., A.C. Acquisition of data: A.M. Analysis and interpretation of data: A.M., K.G., P.C., L.M., R.S., A.C. Drafting of the manuscript: A.M., K.G., P.C., L.M., R.S., A.C. Critical revision of the manuscript for important intellectual content: A.M., K.G., P.C., L.M., R.S., A.C. Statistical analysis: A.M. Obtaining funding: P.C., L.M. Ownership of data: A.M.

**Funding** Open access funding provided by Politecnico di Milano within the CRUI-CARE Agreement. This work was supported by FAIR (Future of Artificial Intelligence Research) project, funded within the PNRR-PE program by Italian Ministry of University.

## Declarations

**Competing interests** The other authors have no competing interests to declare that are relevant to the content of this article. The authors have no relevant financial interests to disclose.

## References


Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Heal PM, Latifi S, Aziz S, Damseh R, Alabed Alrazak S, Sheikh J (2023) Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 9:e48291

Abi-Rafeh J et al (2023) Complications following facelift and neck lift: implementation and assessment of large language model and artificial intelligence (ChatGPT) performance across 16 simulated patient presentations. Aesthetic Plast Surg 47(6):2407–2414. https://doi.org/10.1007/s00266-023-03538-1

Agarwal M et al (2023) Analysing the applicability of ChatGPT, bard, and Bing to generate reasoning-based multiple-choice questions in medical physiology. Cureus 15(6):e40977–e40977. https://doi.org/10.7759/cureus.40977

Alanzi TM (2023) Impact of ChatGPT on teleconsultants in healthcare: perceptions of healthcare experts in Saudi Arabia. J Multidiscip Healthc 16:2309–2321. https://doi.org/10.2147/JMDH.S419847

Alayrac, J, Donahue J, Luc P (2022) Flamingo: a visual language model for fewshot learning. arXiv:2204.14198. https://doi.org/10.48550/arXiv.2204.14198

Ali R et al (2023a) Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. Neurosurgery 93(6):1353–1365. https://doi.org/10.1227/neu.0000000000002632

Ali R et al (2023b) Performance of ChatGPT, GPT-4, and google bard on a neurosurgery oral boards preparation question bank. Neurosurgery 93(5):1090–1098. https://doi.org/10.1227/neu.0000000000002551

Ali K et al (2024) ChatGPT-A double-edged sword for healthcare education? Implications for assessments of dental students. Eur J Dent Educ 28(1):206–211. https://doi.org/10.1111/eje.12937

Almazyad M et al (2023) Enhancing expert panel discussions in pediatric palliative care: innovative scenario development and summarization with ChatGPT-4. Cureus 15(4):e38249. https://doi.org/10.7759/cureus.38249

Alshami A et al (2023) Harnessing the power of ChatGPT for automating systematic review process: methodology, case study, limitations, and future directions. Systems 11(7):351. https://doi.org/10.3390/systems11070351

Altamimi I et al (2023) Snakebite advice and counseling from artificial intelligence: an acute venomous snakebite consultation with ChatGPT. Cureus 15(6):e40351. https://doi.org/10.7759/cureus.40351

Angel MC, Rinehart JB, Canneson MP, Baldi P (2024) Clinical knowledge and reasoning abilities of AI large language models in anesthesiology: a comparative study on the ABA exam. Anesth Analg. https://doi.org/10.1213/ANE.0000000000006892

Anghelescu A et al (2023) PRISMA systematic literature review, including with meta-analysis vs. Chatbot/GPT (AI) regarding current scientific data on the main effects of the calf blood deproteinized hemoderivative medicine (Actovegin) in ischemic stroke. Biomedicines 11(6):1623. https://doi.org/10.3390/biomedicines11061623

Antaki F, Touma S, Milad D, El-Khoury J, Duval R (2023) Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. Ophthalmol Sci 3(4):100324. https://doi.org/10.1016/j.xops.2023.100324

Ayoub M et al (2023) Mind + Machine: ChatGPT as a basic clinical decisions support tool. Cureus 15(8):e43690. https://doi.org/10.7759/cureus.43690

Ayoub NF et al (2024) Head-to-head comparison of ChatGPT versus google search for medical knowledge acquisition. Otolaryngol Head Neck Surg 170(6):1484–1491. https://doi.org/10.1002/ohn.465

Babl FE, Babl MP (2023) Generative artificial intelligence: Can ChatGPT write a quality abstract? Emerg Med Australas 35(5):809–811. https://doi.org/10.1111/1742-6723.14233

Bai Y, Kadavath S, Kundu S, et al (2022) Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073v1. https://doi.org/10.48550/arXiv.2212.08073

Beaulieu-Jones BR, Shah S, Berrigan MT, Marwaha JS, Lai SL, Brat GA (2024) Evaluating capabilities of large language models: performance of GPT4 on surgical knowledge assessments. Surgery 175(4):936–942. https://doi.org/10.1016/j.surg.2023.12.014

Bellinger JR et al (2024) BPPV information on google versus AI (ChatGPT). Otolaryngol Head Neck Surg 170(6):1504–1511. https://doi.org/10.1002/ohn.506

Bhayana R, Krishna S, Bleakney RR (2023) Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. Radiology 307(5):e230582. https://doi.org/10.1148/radiol.230582

Biswas S et al (2023) ChatGPT and the future of journal reviews: a feasibility study. Yale J Biol Med 96(3):415–420. https://doi.org/10.59249/SKDH9286

Bosbach WA et al (2023) Ability of ChatGPT to generate competent radiology reports for distal radius fracture by use of RSNA template items and integrated AO classifier. Curr Problems Diagnostic Radiol 53(1):102–110. https://doi.org/10.1067/j.cpradiol.2023.04.001

Brown T, Mann B, Ryder N, et al (2020) Language models are few-shot learners. arXiv:2005.14165. https://doi.org/10.48550/arXiv.2005.14165

Chiesa-Estomba CM, Lechien JR, Vaira LA, Brunet A, Cammaroto G, Mayo-Yanez M, Sanchez-Barrueco A, Saga-Gutierrez C (2024) Exploring the potential of Chat-GPT as a supportive tool for sialendoscopy clinical decision making and patient information support. Eur Arch Otorhinolaryngol 281(4):2081–2086. https://doi.org/10.1007/s00405-023-08104-8

Chung P et al (2023) Case scenario generators for trauma surgery simulation utilizing autoregressive language models. Artif Intell Med 144:102635. https://doi.org/10.1016/j.artmed.2023.102635

Cobbe K, Kosaraju V, Bavarian M, et al (2021) Training verifiers to solve math word problems. arXiv:2110.14168. https://doi.org/10.48550/arXiv.2110.14168

Cocci A, Pezzoli M, Lo Re M, Russo GI, Asmundo MG, Fode M, Cacciamani G, Cimino S, Minervini A, Durukan E (2023) Quality of information and appropriateness of ChatGPT outputs for urology patients. Prostate Cancer Prostatic Dis 27(1):103–108. https://doi.org/10.1038/s41391-023-00705-y

Cooke A, Smith D, Booth A (2012) Beyond PICO: the SPIDER tool for qualitative evidence synthesis. Qual Health Res 22(10):1435–1443. https://doi.org/10.1177/1049732312452938

Cuthbert R, Simpson AI (2023) Artificial intelligence in orthopaedics: can Chat Generative Pre-trained Transformer (ChatGPT) pass Section 1 of the Fellowship of the Royal College of Surgeons (Trauma & Orthopaedics) examination? Postgrad Med J. https://doi.org/10.1093/postmj/qgad053

Dhanvijay AKD et al (2023) Performance of large language models (ChatGPT, Bing search, and google bard) in solving case vignettes in physiology. Cureus 15(8):e42972. https://doi.org/10.7759/cureus.42972

Driess D et al (2023) PaLM-E: an embodied multimodal language model. arXiv:2303.03378. https://doi.org/10.48550/arXiv.2303.03378

Fang C et al (2023) How does ChatGPT4 preform on non-English national medical licensing examination? An evaluation in Chinese language. PLOS Digit Health 2(12):e0000397. https://doi.org/10.1371/journal.pdig.0000397

Gabriel J et al (2023) The utility of the ChatGPT artificial intelligence tool for patient education and enquiry in robotic radical prostatectomy. Int Urol Nephrol 55(11):2717–2732. https://doi.org/10.1007/s11255-023-03729-4

Gallagher AG, O'Sullivan GC (2012) Fundamentals of surgical simulation. Springer, Cham

Gao CA et al (2023) Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digit Med 6(1):75. https://doi.org/10.1038/s41746-023-00819-6

Gebrael G, Sahu KK, Chigarira B, Tripathi N, Mathew Thomas V, Sayegh N, Maughan BL, Agarwal N, Swami U, Li H (2023) Enhancing triage efficiency and accuracy in emergency rooms for patients with metastatic prostate cancer: a retrospective analysis of artificial intelligence-assisted triage using ChatGPT 4.0. Cancers (basel). 15(14):3717. https://doi.org/10.3390/cancers15143717

Giannos P (2023a) Evaluating the limits of AI in medical specialisation: ChatGPT's performance on the UK neurology specialty certificate examination. BMJ Neurol Open 5(1):e000451. https://doi.org/10.1136/bmjno-2023-000451

Giannos P, Delardas O (2023b) Performance of ChatGPT on UK standardized admission tests: insights from the BMAT, TMUA, LNAT, and TSA examinations. JMIR Med Educ 9:e47737. https://doi.org/10.2196/47737

Gilson A, Safranek CW, Huang T et al (2023) How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 9:e45312. https://doi.org/10.2196/45312

Haemmerli J et al (2023) ChatGPT in glioma adjuvant therapy decision making: ready to assume the role of a doctor in the tumour board? BMJ Health Care Inform 30(1):e100775. https://doi.org/10.1136/bmjhci-2023-100775

Han T, et al (2023) MedAlpaca -- An Open-Source Collection of Medical Conversational AI Models and Training Data. arXiv: 2304.08247. https://doi.org/10.48550/arXiv.2304.08247

Hatamizadeh A, Tang Y, Nath V, et al (2021) UNETR: Transformers for 3D Medical Image Segmentation. arXiv:2103.10504v3. https://doi.org/10.48550/arXiv.2103.10504

Hoch CC, Wollenberg B, Lüers JC, Knoedler S, Knoedler L, Frank K, Cotofana S, Alfertshofer M (2023) ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification preparation questions. Eur Arch Otorhinolaryngol 280(9):4271–4278. https://doi.org/10.1007/s00405-023-08051-4

Holmes J, Liu Z, Zhang L, Ding Y, Sio TT, McGee LA, Ashman JB, Li X, Liu T, Shen J, Liu W (2023) Evaluating large language models on a highly-specialized topic, radiation oncology physics. Front Oncol 13:1219326. https://doi.org/10.3389/fonc.2023.1219326

Hopkins BS, Nguyen VN, Dallas J, Texakalidis P, Yang M, Renn A, Guerra G, Kashif Z, Cheok S, Zada G, Mack WJ (2023) ChatGPT versus the neurosurgical written boards: a comparative analysis of artificial intelligence/machine learning performance on neurosurgical board-style questions. J Neurosurg. https://doi.org/10.3171/2023.2.JNS23419

Hsu HY et al (2023) Examining real-world medication consultations and drug-herb interactions: ChatGPT performance evaluation. JMIR Med Educ 9:e48433. https://doi.org/10.2196/48433

Huang Y et al (2023) Benchmarking ChatGPT-4 on ACR radiation oncology in-training (TXIT) exam and red journal gray zone cases: potentials and challenges for AI-assisted medical education and decision making in radiation oncology. Front Oncol 13:1265024. https://doi.org/10.3389/fonc.2023.1265024

Huh S (2023) Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. J Educ Eval Health Prof 20:1. https://doi.org/10.3352/jeehp.2023.20.1

Huynh LM, Bonebrake BT, Schultis K, Quach A, Deibert CM (2023) New artificial intelligence ChatGPT performs poorly on the 2022 self-assessment study program for urology. Urol Pract 10(4):409–415. https://doi.org/10.1097/UPJ.0000000000000406

Jacob J (2023) ChatGPT: friend or foe?-Utility in trauma triage. Indian J Crit Care Med 27(8):563–566. https://doi.org/10.5005/jp-journals-10071-24498

Jang D et al (2023) Exploring the Potential of Large Language models in Traditional Korean Medicine: A Foundation Model Approach to Culturally-Adapted Healthcare. arXiv:2303.17807. https://doi.org/10.48550/arXiv.2303.17807

Kaarre J, et al (2023) Exploring the potential of ChatGPT as a supplementary tool for providing orthopaedic information. Knee Surg Sports Traumatol Arthrosc. 31(11):5190–5198. https.//doi.org/https://doi.org/10.1007/s00167-023-07529-2

Kao HJ et al (2023) Assessing ChatGPT's capacity for clinical decision support in pediatrics: a comparative study with pediatricians using KIDMAP of Rasch analysis. Medicine (baltimore) 102(25):e34068. https://doi.org/10.1097/MD.0000000000034068

Karakas C et al (2023) Leveraging ChatGPT in the pediatric neurology clinic: practical considerations for use to improve efficiency and outcomes. Pediatr Neurol 148:157–163. https://doi.org/10.1016/j.pediatrneurol.2023.08.035

Kasai J et al (2023) Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations. arXiv:2303.18027. https://doi.org/10.48550/arXiv.2303.18027

Kington RS, Arnesen S, Chou WYS, Curry SJ, Lazer D, Villarruel A (2021) Identifying credible sources of health information in social media: Principles and attributes. NAM Perspect. https://doi.org/10.31478/202107a

Koga S, Martin NB, Dickson DW (2023) Evaluating the performance of large language models: ChatGPT and Google Bard in generating differential diagnoses in clinicopathological conferences of neurodegenerative disorders. Brain Pathol 34(3):e13207. https://doi.org/10.1111/bpa.13207

Koh SJQ et al (2023) Leveraging ChatGPT to aid patient education on coronary angiogram. Ann Acad Med Singap 52(7):374–377. https://doi.org/10.47102/annals-acadmedsg.2023138

Kumah-Crystal Y, Mankowitz S, Embi P, Lehmann CU (2023) ChatGPT and the clinical informatics board examination: the end of unproctored maintenance of certification? J Am Med Inform Assoc. https://doi.org/10.1093/jamia/ocad104

Kung TH, Cheatham M, Medenilla A et al (2023) Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198

Lebhar MS et al (2023) Dr. ChatGPT: utilizing artificial intelligence in surgical education. Cleft Palate Craniofac J. https://doi.org/10.1177/10556656231193966

Lee H (2023a) The rise of ChatGPT: exploring its potential in medical education. Anat Sci Educ. https://doi.org/10.1002/ase.2270

Lee H (2023b) Using ChatGPT AS A LEARNING TOOL IN ACUPUNCTURE EDUCATION: COMPARATIVE STUDY. JMIR Med Educ 9:e47427. https://doi.org/10.2196/47427

Lee P, Bubeck S, Petro J (2023) Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 388(13):1233–1239. https://doi.org/10.1056/NEJMsr2214184

Li SW, Kemp MW, Logan SJS, Dimri PS, Singh N, Mattar CNZ, Dashraath P, Ramlal H, Mahyuddin AP, Kanayan S, Carter SWD, Thain SP, Fee EL, Illanes SE, Choolani MA, National University of Singapore Obstetrics and Gynecology Artificial Intelligence (NUS OBGYN-AI) Collaborative Group (2023c) ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and gynecology. Am J Obstet Gynecol 229(2):172.e1-172.e12. https://doi.org/10.1016/j.ajog.2023.04.020

Li XL, Liang P (2021) Prefix-tuning: Optimizing continuous prompts for generation. arXiv:2101.00190. https://doi.org/10.48550/arXiv.2101.00190

Li J, Li S, Savarese S, Hoi S (2023b) BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv:2301.12597. https://doi.org/10.48550/arXiv.2301.12597

Li C (2023a) LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day. arXiv:2306.00890. https://doi.org/10.48550/arXiv.2306.00890

Li Y et al (2023d) ChatDoctor: a medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge. arXiv: 2303.14070v5. https://doi.org/10.48550/arXiv.2303.14070

Liévin V, Egeberg Hother C, Winther O (2022) Can large language models reason about medical questions? arXiv: 2207.08143. https://doi.org/10.48550/arXiv.2207.08143

Liu S et al (2023) Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc 30(7):1237–1245. https://doi.org/10.1093/jamia/ocad072

Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G (2023b) Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surveys 55(9):1–35. https://doi.org/10.1145/3560815

Liu X, Fang C, Wang J (2023c) Performance of ChatGPT on clinical medicine entrance examination for Chinese postgraduate in Chinese. medRxiv. https://doi.org/10.1101/2023.04.12.23288452

Liu X, Wu C, Lai R, Lin H, Xu Y, Lin Y, Zhang W (2023d) ChatGPT: when the artificial intelligence meets standardized patients in clinical training. J Transl Med 21(1):447. https://doi.org/10.1186/s12967-023-04314-0

Liu H et al (2023e) How good is ChatGPT for medication evidence synthesis? Stud Health Technol Inform 302:1062–1066. https://doi.org/10.3233/SHTI230347

Liu H (2023a) Visual Instruction Tuning. arXiv:2304.08485. https://doi.org/10.48550/arXiv.2304.08485

Lower K et al (2023) ChatGPT-4: Transforming Medical Education and Addressing Clinical Exposure Challenges in the Post-pandemic Era. Indian J Orthop 57(9):1527–1544. https://doi.org/10.1007/s43465-023-00967-7

Lum ZC (2023) Can Artificial Intelligence Pass the American Board of Orthopaedic Surgery Examination? Orthopaedic Residents Versus ChatGPT. Clin Orthop Relat Res 481(8):1623–1630

Lyons RJ et al (2023) Artificial intelligence chatbot performance in triage of ophthalmic conditions. Can J Ophthalmol. https://doi.org/10.1016/j.jcjo.2023.07.016

Lyu Q et al (2023) Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art 6(1):9. https://doi.org/10.1186/s42492-023-00136-5

Macdonald C et al (2023) Can ChatGPT draft a research article? An example of population-level vaccine effectiveness analysis. J Glob Health 13:01003. https://doi.org/10.7189/jogh.13.01003

Mihalache A, Popovic MM, Muni RH (2023) Performance of an artificial intelligence Chatbot in ophthalmic knowledge assessment. JAMA Ophthalmol 141(6):589–597. https://doi.org/10.1001/jamaophthalmol.2023.1144

Mohapatra DP et al (2023) Leveraging large language models (LLM) for the plastic surgery resident training: do they have a role? Indian J Plast Surg 56(5):413–420. https://doi.org/10.1055/s-0043-1772704

Mondal H et al (2023) Using ChatGPT for writing articles for patients' education for dermatological diseases: a pilot study. Indian Dermatol Online J 14(4):482–486. https://doi.org/10.4103/idoj.idoj_72_23

Nath S, Marie A, Ellershaw S, Korot E, Keane PA (2022) New meaning for NLP: the trials and tribulations of natural language processing with GPT-3 in ophthalmology. Br J Ophthalmol 106(7):889–892. https://doi.org/10.1136/bjophthalmol-2022-321141

Nazario-Johnson L, Zaki HA, Tung GA (2023) Use of large language models to predict neuroimaging. J Am Coll Radiol 20(10):1004–1009. https://doi.org/10.1016/j.jacr.2023.06.008

Nori H, King N, McKinney SM, Carignan D, Horvitz E (2023) Capabilities of gpt-4 on medical challenge problems. arXiv:2303.13375. https://doi.org/10.48550/arXiv.2303.13375

Oh N, Choi GS, Lee WY (2023) ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. Ann Surg Treat Res 104(5):269–273. https://doi.org/10.4174/astr.2023.104.5.269

OpenAI. GPT-4 Technical report (2023). arXiv:2303.08774. https://doi.org/10.48550/arXiv.2303.08774

Ouyang L, Wu J, Jiang X, et al (2022) Training language models to follow instructions with human feedback. arXiv:2203.02155. https://doi.org/10.48550/arXiv.2203.02155

Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD et al (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Int J Surg 88:105906

Passby L, Jenko N, Wernham A (2023) Performance of ChatGPT on dermatology specialty certificate examination multiple choice questions. Clin Exp Dermatol 2:llad197. https://doi.org/10.1093/ced/llad197

Reeder K, Lee H (2022) Impact of artificial intelligence on US medical students' choice of radiology. Clin Imaging 81:67–71. https://doi.org/10.1016/j.clinimag.2021.09.018

Rizwan A, Sadiq T (2023) The use of AI in diagnosing diseases and providing management plans: a consultation on cardiovascular disorders with ChatGPT. Cureus 15(8):e43106. https://doi.org/10.7759/cureus.43106

Saad A, Iyengar KP, Kurisunkal V, Botchu R (2023) Assessing ChatGPT's ability to pass the FRCS orthopaedic part a exam: a critical analysis. Surgeon. https://doi.org/10.1016/j.surge.2023.07.001

Sallam M et al (2023) ChatGPT applications in medical, dental, pharmacy, and public health education: a descriptive study highlighting the advantages and limitations. Narra J 3(1):e103. https://doi.org/10.52225/narra.v3i1.103

Sarbay İ et al (2023) Performance of emergency triage prediction of an open access natural language processing based chatbot application (ChatGPT): a preliminary, scenario-based cross-sectional study. Turk J Emerg Med 23(3):156–161. https://doi.org/10.4103/tjem.tjem_79_23

Satava RM, Stefanidis D, Levy JS et al (2020) Proving the effectiveness of the fundamentals of robotic surgery (FRS) skills curriculum: a single-blinded, multispecialty. Multi-Inst Randomized Control Trial Ann Surg 272(2):384–392. https://doi.org/10.1097/SLA.0000000000003220

Sevgi UT et al (2023) The role of an open artificial intelligence platform in modern neurosurgical education: a preliminary study. Neurosurg Rev 46(1):86. https://doi.org/10.1007/s10143-023-01998-2

Shademan A, Decker RS, Opfermann JD, Leonard S, Krieger A, Kim PC (2016) Supervised autonomous robotic soft tissue surgery. Sci Transl Med 8(337):337ra64. https://doi.org/10.1126/scitranslmed.aad9398

Sharma P (2023) Performance of ChatGPT on USMLE: unlocking the potential of large language models for AI-assisted medical education. arXiv: 2307.00112. https://doi.org/10.48550/arXiv.2307.00112

Shay D, Kumar B, Bellamy D, Palepu A, Dershwitz M, Walz JM, Schaefer MS, Beam A (2023) Assessment of ChatGPT success with specialty medical knowledge using anaesthesiology board examination practice questions. Br J Anaesth 131(2):e31–e34. https://doi.org/10.1016/j.bja.2023.04.017

Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, Moher D, Tugwell P, Welch V, Kristjansson E, Henry DA (2017) AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 358:j4008. https://doi.org/10.1136/bmj.j4008

Shihadeh J, Ackerman M, Troske A, Lawson N, Gonzalez E (2022) Brilliance bias in GPT-3. In 2022 IEEE Global Humanitarian Technology Conference (GHTC) (pp. 62–69). https://doi.org/10.1109/GHTC55712.2022.9910995

Singhal K, Azizi S, Tu T et al (2023) Large language models encode clinical knowledge. Nature. https://doi.org/10.1038/s41586-023-06291-2

Singhal K, et al (2023b) Towards Expert-Level Medical Question Answering with Large Language Models. arXiv:2305.09617. https://doi.org/10.48550/arXiv.2305.09617

Skalidis I, Cagnina A, Luangphiphat W, Mahendiran T, Muller O, Abbe E, Fournier S (2023) ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story? Eur Heart J Digit Health 4(3):279–281. https://doi.org/10.1093/ehjdh/ztad029

Smith J, Choi PM, Buntine P (2023a) Will code one day run a code? Performance of language models on ACEM primary examinations and implications. Emerg Med Australas. https://doi.org/10.1111/1742-6723.14280

Smith A et al (2023b) Old dog, new tricks? Exploring the potential functionalities of ChatGPT in supporting educational methods in social psychiatry. Int J Soc Psychiatry 69(8):1882–1889. https://doi.org/10.1177/00207640231178451

Stiennon N, Ouyang L, Wu J, Ziegler DM, Lowe R, Voss C, Radford A, Amodei D, and Christiano P (2022) Learning to summarize from human feedback. arXiv:2009.01325. https://doi.org/10.48550/arXiv.2009.01325

Stokel-Walker C, Van Noorden R (2023) What ChatGPT and generative AI mean for science. Nature 614(7947):214–216. https://doi.org/10.1038/d41586-023-00340-6

Strong E, DiGiammarino A, Weng Y, Basaviah P, Hosamani P, Kumar A, Nevins A, Kugler J, Hom J, Chen JH (2023) Performance of ChatGPT on free-response, clinical reasoning exams. medRxiv. https://doi.org/10.1101/2023.03.24.23287731

Suchman K, Garg S, Trindade AJ (2023) Chat generative pretrained transformer fails the multiple-choice American college of gastroenterology self-assessment test. Am J Gastroenterol. https://doi.org/10.14309/ajg.0000000000002320

Taira K, Itaya T, Hanada A (2023) Performance of the Large language model ChatGPT on the national nurse examinations in japan: evaluation study. JMIR Nurs 6:e47305. https://doi.org/10.2196/47305

Takagi S, Watari T, Erabi A, Sakaguchi K (2023) Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. JMIR Med Educ. 9:e48002. https://doi.org/10.2196/48002

Taylor R, Kardas M, Cucurull G, et al (2022) Galactica: A Large Language Model for Science. arXiv:2211.09085. https://doi.org/10.48550/arXiv.2211.09085

Thirunavukarasu AJ, Hassan R, Mahmood S, Sanghera R, Barzangi K, El Mukashfi M, Shah S (2023) Trialling a large language model (ChatGPT) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. JMIR Med Educ 9:e46599. https://doi.org/10.2196/46599

Toma A, et al (2023) Clinical Camel: An Open-Source Expert-Level Medical Language Model with Dialogue-Based Knowledge Encoding. arXiv: 2305.12031. https://doi.org/10.48550/arXiv.2305.12031

Totlis T et al (2023) The potential role of ChatGPT and artificial intelligence in anatomy education: a conversation with ChatGPT. Surg Radiol Anat 45(10):1321–1329. https://doi.org/10.1007/s00276-023-03229-1

Touvron H, Lavril T, Izacard G, et al (2023) LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971. https://doi.org/10.48550/arXiv.2302.13971

Tu T (2023) Towards Generalist Biomedical AI. arXiv:2307.14334. https://doi.org/10.48550/arXiv.2307.14334

Valentín-Bravo FJ et al (2023) Artificial Intelligence and new language models in Ophthalmology: complications of the use of silicone oil in vitreoretinal surgery. Arch Soc Esp Oftalmol (engl Ed) 98(5):298–303. https://doi.org/10.1016/j.oftale.2023.04.011

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, et al (2017) Attention is all you need. arXiv:1706.03762v5. https://doi.org/10.48550/arXiv.1706.03762

Wang YM, Shen HW, Chen TJ (2023) Performance of ChatGPT on the pharmacist licensing examination in Taiwan. J Chin Med Assoc 86(7):653–658. https://doi.org/10.1097/JCMA.0000000000000942

Wang X, et al (2022) Self-consistency improves chain of thought reasoning in language models. arXiv:2203.11171. https://doi.org/10.48550/arXiv.2203.11171

Wang H, et al (2023a) HuaTuo: Tuning LLaMA Model with Chinese Medical Knowledge. arXiv:2304.06975. https://doi.org/10.48550/arXiv.2304.06975

Wang S, et al (2023b) ChatCAD: Interactive Computer-Aided Diagnosis on Medical Image using Large Language Models. arXiv: 2302.07257. https://doi.org/10.48550/arXiv.2302.07257

Wei J, et al (2022) Chain of thought prompting elicits reasoning in large language models. arXiv:2201.11903. https://doi.org/10.48550/arXiv.2201.11903

Weng TL, Wang YM, Chang S, Chen TJ, Hwang SJ (2023) ChatGPT failed Taiwan's family medicine board exam. J Chin Med Assoc. https://doi.org/10.1097/JCMA.0000000000000946

Wu J et al (2024) The application of ChatGPT in medicine: a scoping review and bibliometric analysis. J Multidiscip Healthc 17:1681–1692. https://doi.org/10.2147/JMDH.S463128

Wu C, et al (2023) PMC-LLaMA: Further Finetuning LLaMA on Medical Papers. arXiv: 2304.14454. https://doi.org/10.48550/arXiv.2304.14454

Xie Y et al (2023) Evaluation of the artificial intelligence Chatbot on breast reconstruction and its efficacy in surgical research: a case study. Aesthetic Plast Surg 47(6):2360–2369. https://doi.org/10.1007/s00266-023-03443-7

Xie Y et al (2024) Investigating the impact of innovative AI chatbot on post-pandemic medical education and clinical assistance: a comprehensive analysis. ANZ J Surg 94(1–2):68–77. https://doi.org/10.1111/ans.18666

Xiong H, et al (2023) DoctorGLM: Fine-tuning your Chinese Doctor is not a Herculean Task. arXiv:2304.01097v2. https://doi.org/10.48550/arXiv.2304.01097

Zevin B, Levy JS, Satava RM, Grantcharov TP (2012) A consensus-based framework for design, validation, and implementation of simulation-based training curricula in surgery. J Am Coll Surg 215(4):580-586.e3. https://doi.org/10.1016/j.jamcollsurg.2012.05.035
Zhou Z (2023) Evaluation of ChatGPT's capabilities in medical report generation. Cureus. 15(4):e37589.

## Authors and Affiliations

**Andrea Moglia[1] · Konstantinos Georgiou[2] · Pietro Cerveri[1,3] · Luca Mainardi[1] · Richard M. Satava[4] · Alfred Cuschieri[5,6]**

✉ Andrea Moglia
andrea.moglia@polimi.it

[1] Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Milan, Italy

[2] 1st Propaedeutic Surgical Unit, Athens Medical School, Hippocrateion Athens General Hospital, National and Kapodistrian University of Athens, Athens, Greece

[3] Department of Industrial and Information Engineering, University of Pavia, Pavia, Italy

[4] Department of Surgery, University of Washington Medical Center, Seattle, WA, USA

[5] Scuola Superiore Sant'Anna of Pisa, Pisa, Italy

[6] Institute for Medical Science and Technology, University of Dundee, Dundee, UK

https://doi.org/10.7759/cureus.37589