

Article

Dermatological Knowledge and Image Analysis Performance of Large Language Models Based on Specialty Certificate Examination in Dermatology

Ka Siu Fan ^{1,*} and Ka Hay Fan ^{2,†} ¹ Department of Nutritional Sciences, University of Surrey, Guildford GU2 7XH, UK² Faculty of Medicine, Imperial College, London SW7 5NH, UK

* Correspondence: fankasiu@doctors.org.uk

† These authors contributed equally to this work.

Abstract: Large language models (LLMs) are trained using large datasets and may be applied to language-based tasks. Studies have demonstrated their ability to perform and pass postgraduate medical examinations, and with the increasingly sophisticated deep learning algorithms and incorporation of image-analysis capabilities, they may also be applied to the Specialty Certificate Examination (SCE) in Dermatology. The Dermatology SCE sample questions were used to assess the performance of five freely available and high-performance LLMs. The LLMs' performances were recorded by comparing their output on multiple-choice questions against the sample answers. One hundred questions, four of which included photographs, were entered into the LLMs. The responses were recorded and analysed, with the pass mark set at 77%. The accuracies for Claude-3.5 Sonnet, Copilot, Gemini, ChatGPT-4o, and Perplexity were 87, 88, 75, 90, and 87, respectively ($p = 0.023$). The LLMs were generally capable of interpreting and providing reasoned responses to clinical scenarios and clinical data. This continues to demonstrate the potential of LLMs in both medical education and clinical settings.



Citation: Fan, K.S.; Fan, K.H. Dermatological Knowledge and Image Analysis Performance of Large Language Models Based on Specialty Certificate Examination in Dermatology. *Dermato* **2024**, *4*, 124–135. <https://doi.org/10.3390/dermato4040013>

Academic Editor: Thilo Gambichler

Received: 5 August 2024

Revised: 27 September 2024

Accepted: 28 September 2024

Published: 30 September 2024



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

Keywords: artificial intelligence; dermatology; certification; education; large language model

1. Introduction

Artificial intelligence (AI) is being progressively integrated into healthcare to streamline workflow and assist with clinical decision-making [1,2]. Among these are large language models (LLMs), which are now increasingly popular among both healthcare professionals and the public alike. These models utilise deep learning algorithms and multi-layered neural networks to learn from large amounts of text and generate human-like conversations and responses. Many models are now available to the public, with little to no upfront costs. Commercially available models, like ChatGPT, Gemini, and Perplexity, all provide free access to their base models while more powerful models require subscriptions. Open-sourced alternatives, including Mixtral and Llama, are often less powerful but can be used freely under open-access licenses. Together, the explosive interests in LLMs and artificial intelligence have driven rapid advancements that offer great potential to improve healthcare.

Despite being trained on large amounts of data from the Internet, which included academic resources, their role in education and medical practice continues to be investigated [3,4]. These models may be used by both laypersons and professionals to produce insight into medical problems and potentially replace search engines which merely index and curate resources. While the use and implications of AI in real-world clinical settings remain controversial, the breadth and depth of its training data provides great potential to augment medical education [5]. Within education, the use of chatbots may provide real-time feedback and answers to medical vignettes and scenarios. Additionally, it provides

the interactivity vital to learning experience without the resource-intensive requirements of traditional learning settings like expert-led lectures [6]. For practicing clinicians, generative AI may be used to better understand real-life cases, as well as to stay up to date on less-encountered or rarer presentations. Additionally, with recent advances, these models can now analyse images and may provide a resource-efficient means to further the professional development and training of dermatologists [7].

The growing use of generative AI in medical education must be cautioned against its actual ability to perform. While medical educators are typically experienced in their fields and understand evidence-based medicine and are held to set standards by governing bodies, generative AIs are not. This raises an important concern: are the training datasets of appropriate quality to output accurate and unbiased responses? To begin answering this, studies have examined its ability to sit and pass various postgraduate medical and surgical examinations. These studies evaluated different levels of medical knowledge, where it was able to reach the pass threshold in the United States Medical Licensing Examination (USMLE) and the intercollegiate Membership of the Royal College of Surgeons (MRCS) and Membership of the Royal College of Physicians (MRCP) examinations of the United Kingdom (UK) [8–15]. Its applications appear consistent across exams that are based on multiple-choice or single-best-answer question formats.

The Specialty Certificate Examination (SCE) in Dermatology is a postgraduate exam for dermatologists in the UK. This is one of the requirements to qualify as a dermatology specialist. The majority of questions related to general dermatology, skin oncology, and paediatric genetics. The SCE is an exam with 200 multiple-choice questions, spread between 14 categories [16]. With the most recent pass mark being set at 77.4%, the overall pass rate was 50.6% [17]. The performance of LLMs in the Dermatology SCE was previously examined, demonstrating marked differences between ChatGPT models, with ChatGPT-3.5 scoring 63% and ChatGPT-4 scoring 90% [18]. With the increased availability and capability of commercial LLMs, this study aimed to evaluate and compare the performance of different models.

While the role of ChatGPT in basic and clinical sciences has been evaluated, there remains debate about the differences in the abilities of other common LLMs. The evidence to support the use of generative AI and LLMs as a diagnostic and learning aid in medical education remains scarce. This study aimed to evaluate the performance of various LLMs in the multiple-choice Specialty Certificate Examination in Dermatology.

2. Materials and Methods

2.1. Question Bank and LLM

This study was conducted in August 2024. To evaluate the performance of LLMs, the SCE in Dermatology sample questions and multiple-choice answers were extracted from the Membership of the Royal Colleges of Physicians of the United Kingdom website [19]. All questions provided as part of the sample exam were retrieved and were considered an accurate representation of the actual SCE in Dermatology exam for this study.

Five web-based LLMs were used in this study: Claude-3.5 Sonnet (Anthropic, San Francisco, CA, USA), Copilot (Microsoft, Redmond, WA, USA), Gemini (Google, Mountain View, CA, USA), ChatGPT-4o (OpenAI, San Francisco, CA, USA), and Perplexity (Perplexity AI, San Francisco, CA, USA) [20–24]. These models were selected as they were all available in most countries and provided free access with limitations. These were considered to be representative of the LLMs likely to be known and chosen by users. Another model, MetaAI, is a powerful LLM but its accessibility remains limited to a handful of countries and was therefore not included. These LLMs process prompts and synthesise an output to answer the users' query.

All LLMs were given a prompt of "Please select the most appropriate option for the following questions:" before being provided the questions. Each of the 100 questions and their respective five multiple-choice answers were inputted into LLMs individually. The responses of each LLM were recorded and compared against the standard answer.

Four multiple-choice options included photographic reference material and images were uploaded in their original resolution alongside the clinical text/question in the same prompt. No additional input or feedback was provided to guide future responses of the LLMs. All data collection was performed in July 2024. The workflow of the study is illustrated in Figure 1.

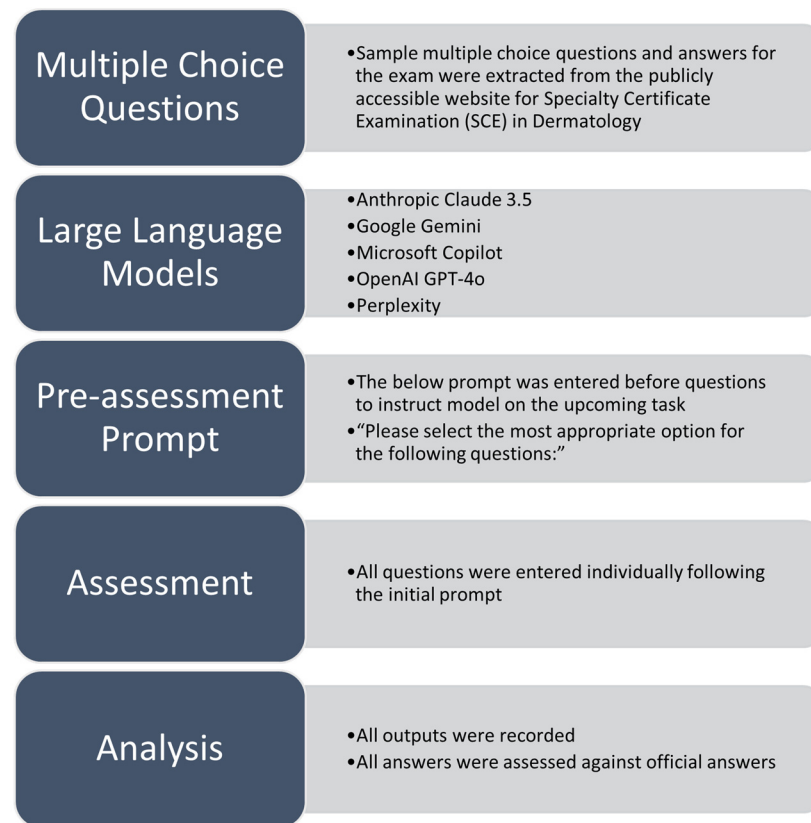


Figure 1. The workflow and procedure of the study.

2.2. Data Analysis

Data were collected and analysed using Excel® (Microsoft, Redmond, WA, USA) and SPSS version 25 (SPSS, New York, NY, USA). The Chi-Squared test was used to compare the performance of LLMs, with statistical significance defined as $p < 0.05$. Boundaries of pass marks were based on historical data, which ranged from 70% to 77%, which this study referred to as lower and upper pass mark boundaries [17,18].

2.3. Ethical Statement

No patient data were used in this study. This study was performed in accordance with the ethical standards of the 1964 Declaration of Helsinki and its later amendments.

3. Results

All 100 sample questions and multiple-choice answers were used in the evaluation of all five LLMs. A response was retrieved from each LLM without additional prompting or response regeneration. The accuracies for Claude-3.5 Sonnet, Copilot, Gemini, ChatGPT-4o, and Perplexity were 87, 88, 75, 90, and 87, respectively (Figure 2). There were statistically significant differences in the performance between the LLMs ($p = 0.023$). All five LLMs passed the minimum pass mark of 70%, with three surpassing the 2023 pass mark of 77%. ChatGPT-4o scored the highest, closely followed by Copilot, Claude-3.5 Sonnet, and Perplexity. Figure 3 and Table 1 demonstrate the breakdown by question category. A total of 56 questions were answered correctly by all five LLMs. All questions were correctly answered by at least one LLM. Four image-based questions were assessed, with

two to five LLMs answering correctly for each of the questions. Only Claude and Copilot provided a caveat on the potential for AI-generated content being incorrect. Justification was often provided by LLMs to show justification for their decision-making without additional prompting (Figure 4).

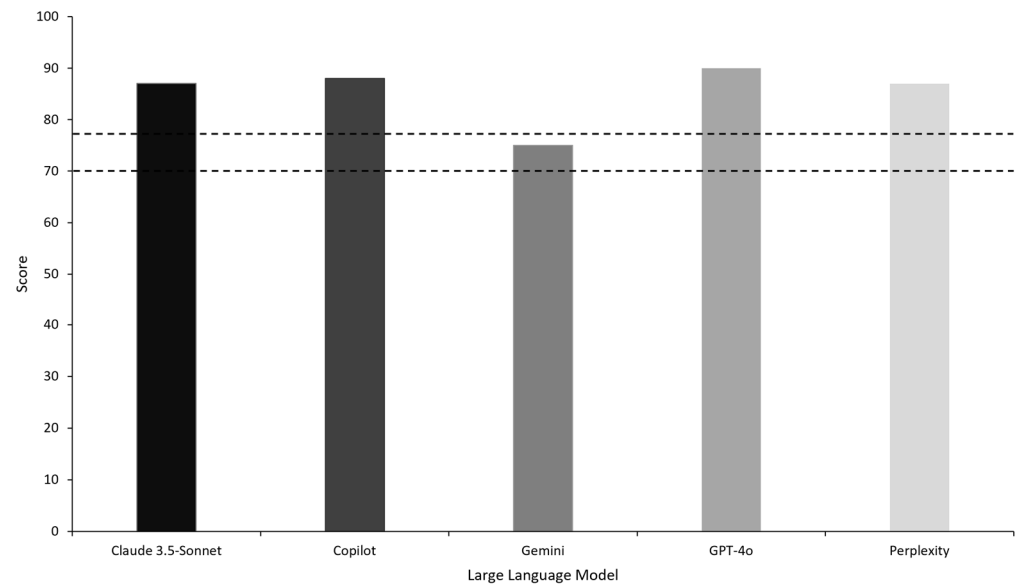


Figure 2. The overall performance of each large language model, with dotted lines representing the lower and upper thresholds of historical cut-off percentages for passing the exam.

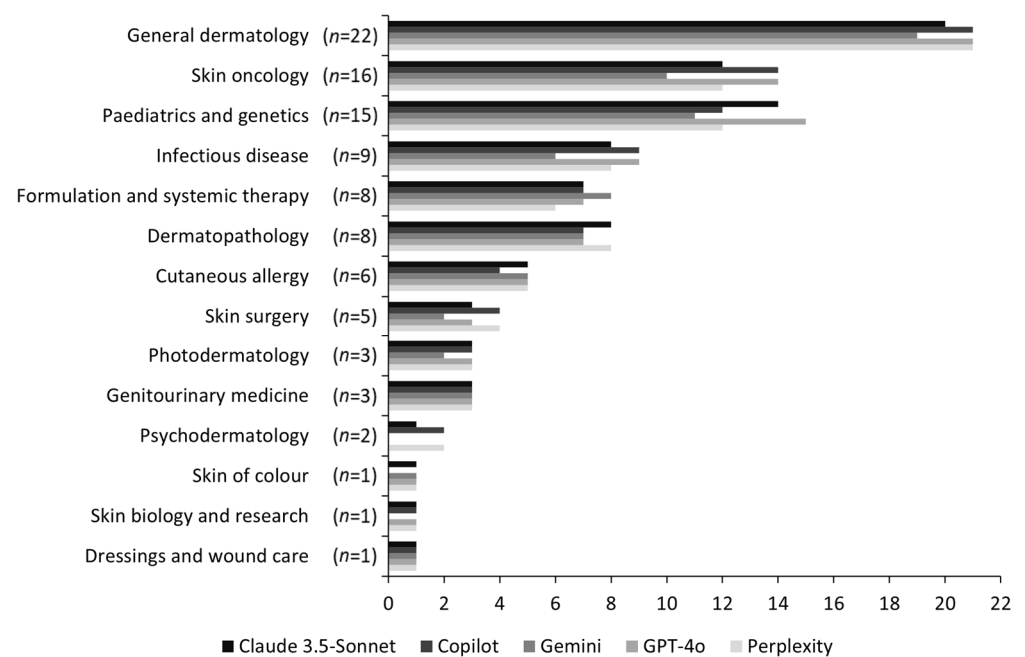


Figure 3. The number of correct answers from each large language model, grouped by category of question.

<p>The most appropriate next step in this patient's management is:</p> <p>E. prednisolone 0.5 mg/kg/day</p> <p>Explanation:</p> <p>Pemphigoid gestationis is a rare autoimmune blistering disorder that occurs during pregnancy or immediately postpartum. In this case, the patient has a severe presentation with widespread urticated eruption, bullae, and excoriations that has not responded to potent topical corticosteroids.</p> <p>Systemic corticosteroids are the first-line treatment for moderate to severe pemphigoid gestationis, especially when topical treatments have failed. Prednisolone at a dose of 0.5 mg/kg/day is typically effective in controlling the disease while minimizing risks to the fetus.</p> <p>Let's consider why the other options are less appropriate:</p> <p>A. Azathioprine: While it can be used in pemphigoid gestationis, it's generally considered as a steroid-sparing agent for long-term management, not as first-line therapy. It also has a slower onset of action.</p> <p>B. Immediate induction of labour: This is too drastic at this stage. While pemphigoid gestationis can worsen around the time of delivery, inducing labor at 35 weeks without trying systemic treatment first is not justified.</p> <p>C. Intravenous immunoglobulin: This is typically reserved for very severe or refractory cases and is not the first-line treatment after failure of topical therapy.</p> <p>D. Oxytetracycline: This antibiotic is not a standard treatment for pemphigoid gestationis and is generally avoided in late pregnancy due to potential effects on fetal bone development.</p> <p>Therefore, prednisolone 0.5 mg/kg/day is the most appropriate next step in managing this patient's condition.</p>	<p>The most appropriate next step in the management of this patient with pemphigoid gestationis is:</p> <p>E. prednisolone 0.5 mg/kg/day</p> <p>Rationale:</p> <ol style="list-style-type: none"> Pemphigoid Gestationis is an autoimmune blistering disorder that occurs during pregnancy, characterised by a pruritic, urticated eruption with bullae formation. It often requires systemic treatment, particularly when potent topical corticosteroids are ineffective. Systemic Corticosteroids such as prednisolone are the mainstay of treatment for pemphigoid gestationis, especially when topical treatments have failed. The dosage of 0.5 mg/kg/day is appropriate for managing this condition and reducing inflammation and pruritus. Other Options: <ul style="list-style-type: none"> A. Azathioprine 2.5 mg/kg/day: Azathioprine can be used as a steroid-sparing agent, but it is not the first-line treatment for acute management. B. Immediate induction of labour: This is not indicated unless there are obstetric complications or foetal distress. Managing the disease medically is preferred. C. Intravenous immunoglobulin 2 g/kg: This is reserved for refractory cases that do not respond to systemic corticosteroids or immunosuppressants. D. Oxytetracycline 500 mg twice daily: This is not a treatment for pemphigoid gestationis and is generally used for other dermatological conditions like bullous pemphigoid, but not as a first-line therapy. <p>Therefore, prednisolone 0.5 mg/kg/day is the most appropriate next step in the management of this patient's condition.</p>
<p>Copilot</p> <p>The most appropriate next step in the management of pemphigoid gestationis, especially when potent topical corticosteroids have proven ineffective, is prednisolone 0.5 mg/kg/day (Option E). Systemic corticosteroids are often necessary in more extensive cases to control the disease ^{1, 2}.</p> <p>Would you like more information on this condition or its management?</p> <p>Learn more 1 2 3 4 5</p> <p>The correct answer is:</p> <p>C. Intravenous immunoglobulin 2 g/kg</p> <p>Here's why:</p> <ul style="list-style-type: none"> Pemphigoid gestationis is an autoimmune blistering disease that occurs during pregnancy or the postpartum period. It is characterized by an itchy, blistering rash that can be severe and difficult to treat. In this case, the patient has already tried a potent topical corticosteroid, which has not been effective. This suggests that she needs a stronger treatment. Intravenous immunoglobulin (IVIg) is a powerful immunosuppressant that can be very effective in treating pemphigoid gestationis. It works by blocking the immune system's attack on the skin. IVIg is usually given in a series of infusions over several days. It can help to improve the rash and reduce itching. The other options are not appropriate: <ul style="list-style-type: none"> Azathioprine is another immunosuppressant, but it takes longer to work and can have more side effects than IVIg. Immediate induction of labor is not necessary, as the baby is not in danger. Oxytetracycline is an antibiotic that would not be effective in treating pemphigoid gestationis. Prednisolone is a systemic corticosteroid that can be used to treat pemphigoid gestationis. However, IVIg is usually preferred as it is less likely to have side effects on the baby. <p>It is important to note that the management of pemphigoid gestationis should be supervised by a dermatologist or other specialist. The patient may also need to be monitored closely during her pregnancy to ensure that her baby is healthy.</p>	<p>Answer</p> <p>Based on the information provided, the most appropriate next step in managing this patient's pemphigoid gestationis is:</p> <p>E. prednisolone 0.5 mg/kg/day</p> <p>Here's the rationale:</p> <ol style="list-style-type: none"> The patient has been diagnosed with pemphigoid gestationis, a rare autoimmune blistering disorder of pregnancy, confirmed by skin biopsy and immunofluorescence ¹. She has a widespread urticated eruption with numerous bullae and excoriations, indicating a moderate to severe case that has not responded to potent topical corticosteroids ¹. Systemic corticosteroids are the mainstay of treatment for moderate to severe pemphigoid gestationis, especially during pregnancy ^{2, 4}. The recommended starting dose for oral corticosteroids in pemphigoid gestationis is 0.5 mg/kg/day of prednisone or prednisolone ^{1, 2}. This dose can provide significant symptom clearance and can be increased if there is an inadequate response ¹. Systemic corticosteroids are preferred over other oral corticosteroids during pregnancy because they are metabolized by placental enzymes, resulting in lower levels reaching the fetus ¹. While other options like azathioprine or intravenous immunoglobulin can be considered as steroid-sparing agents, they are typically reserved for cases that are refractory to systemic corticosteroids or when corticosteroids are contraindicated ³. Immediate induction of labor is not necessary unless there are other obstetric indications, as pemphigoid gestationis itself does not typically require early delivery ⁵. Oxytetracycline is not a standard treatment for pemphigoid gestationis and is generally avoided during pregnancy due to potential risks to the fetus. <p>Therefore, starting oral prednisolone at 0.5 mg/kg/day is the most appropriate next step in managing this patient's pemphigoid gestationis, given its severity and lack of response to topical treatment.</p>

Figure 4. Example of responses provided by each large language model. The subpanels illustrate the following models: Anthropic Claude-3.5 (A), Microsoft Copilot (B), Google Gemini (C), Open AI ChatGPT-4o (D), and Perplexity(E).

Table 1. The number and proportion of correct responses of each large language model by question category.

Topic	Questions	GPT-4o	Gemini	Claude 3.5-Sonnet	Copilot	Perplexity
General dermatology	22	21 (95%)	19 (86%)	20 (91%)	21 (95%)	21 (95%)
Skin oncology	16	14 (88%)	10 (63%)	12 (75%)	14 (88%)	12 (75%)
Paediatrics and genetics	15	15 (100%)	11 (73%)	14 (93%)	12 (80%)	12 (80%)
Infectious disease	9	9 (100%)	6 (67%)	8 (89%)	9 (100%)	8 (89%)
Formulation and systemic therapy	8	7 (88%)	8 (100%)	7 (88%)	7 (88%)	6 (75%)
Dermatopathology	8	7 (88%)	7 (88%)	8 (100%)	7 (88%)	8 (100%)
Cutaneous allergy	6	5 (83%)	5 (83%)	5 (83%)	4 (67%)	5 (83%)
Skin surgery	5	3 (60%)	2 (40%)	3 (60%)	4 (80%)	4 (80%)
Photodermatology	3	3 (100%)	2 (67%)	3 (100%)	3 (100%)	3 (100%)
Genitourinary medicine	3	3 (100%)	3 (100%)	3 (100%)	3 (100%)	3 (100%)
Psychodermatology	2	0 (0%)	0 (0%)	1 (50%)	2 (100%)	2 (100%)
Skin of colour	1	1 (100%)	1 (100%)	1 (100%)	0 (0%)	1 (100%)
Skin biology and research	1	1 (100%)	0 (0%)	1 (100%)	1 (100%)	1 (100%)
Dressings and wound care	1	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)

4. Discussion

The overall performance of the five LLMs was satisfactory in answering sample SCE in Dermatology questions. The LLMs were able to pass the sample exam, scoring within the historical pass cut-offs of 70–77% [17,18]. GPT-4o and Copilot were both clearly above the pass mark, whereas Gemini scored within the range of different historical pass marks. Unlike the previous study, which omitted several questions, this utilised the full set of sample questions. The sample questions, however, may not be representative of the real exam as the exam questions in circulation are not publicly available. Similarly, with the advancements in LLMs to now include image analysis capability, questions involving clinical or histological photographs no longer have to be excluded [25]. The findings of this study were more encouraging than those identified by Passby and Joh, where ChatGPT achieved ~60 and ~90% accuracy in similar dermatological exams [18,26]. Additionally, the study by Nicikowski showcased the performance of both ChatGPT-3.5 and ChatGPT-4 against real candidates in the Polish dermatological exam, demonstrating large leaps in capability between LLM models (45.7% vs. 69.8% accuracy) [27]. Their overall findings also demonstrated that the most powerful models at the time were generally able to pass the exam but underperform compared to human candidates. In comparison, the findings in this study generally match the performance pattern of UK dermatology trainees, where questions were answered correctly between 70 and 90% of the time for most question categories [17]. Though it is not possible to compare these different exams directly, baselining the performance with the pass mark of the exam provides a point of reference to understand how quickly LLM models advance and to demonstrate their abilities to support the use of LLMs in medical education.

Current literature evaluating the medical output of LLMs focuses on ChatGPT. Studies have highlighted its variable performance in medical exams while others identified its pitfalls in producing factually inaccurate and poor clinical advice [8,28–32]. All five LLMs examined were able to meet the lower boundaries of the pass mark, however, Gemini does not exceed the upper boundary. Compared to previous iterations of ChatGPT, all of these LLMs show a considerable capacity to process and respond to medical problems and scenarios [18]. ChatGPT remains at the forefront of LLMs among both the general public and medical professionals. It showcased strong performance in the USMLE, in-service examinations, and board examinations, demonstrating robust performance across various specialities, including medical, surgical, and radiology alike [12,13,33–35]. These studies reported its performance to be at the level of third-year medical students and first-year residents. Other models like Gemini, Claude, and Perplexity were rarely studied, but they do demonstrate some variability between studies [36–39]. In studies where multiple LLMs were examined, ChatGPT and Claude were generally shown to outperform Bard (now Gemini) [38–40]. Notably, LLM performance was not universally transferrable across

languages, where performance was also highly variable in clinical examinations in Chinese, Japanese, and Peruvian languages [40–42]. It appears that LLMs generally experience a performance or accuracy drop when used in other languages [26,42]. Regardless, there is currently a lack of inclusion of other LLM models and this study provided an encompassing snapshot to compare the abilities and performance of these models.

The design and core goals of these LLMs do vary and may explain the variations observed in this study. ChatGPT was developed by Open AI to be a general purpose model and is similar to Google's Gemini AI which was designed to handle complex tasks, including coding and logical reasoning [43]. On the other hand, Copilot was designed by Microsoft to enhance productivity when integrated with other Microsoft Office applications. While GPT, Claude, and Copilot all performed similarly in this study and likely passed the exam based on previous cut-offs, Gemini stood out with slightly poorer performance. The model architecture of the LLMs themselves will determine how the training data are used to learn patterns [11]. ChatGPT and Copilot operate with the Generative Pre-trained Transformer whereas Gemini uses the Language Model for Dialogue Application (LaMDA; Google, Mountain View, CA, USA) and Pathways Language Model (PaLM2; Google, Mountain View, CA, USA) [11]. Perplexity differs from the other models as it is designed as an 'answer engine', which incorporates a large language model into a search engine. Additionally, although each model boasts the vastness and breadth of their training data, the exact datasets are not shared. The use of different datasets and fine-tuning can significantly change how the AI models decipher relationships and draw conclusions. These would also include variations in the amount and quality of medical information used in training, as well as the different degrees of bias. However, it is important to note that the poorer performance does not necessarily translate across all fields of biomedical knowledge as it previously demonstrated reasonable performance in analysing laboratory data [44].

Another explanation is that different models are better at different types of tasks. The implications of this may be observed when comparing the different types/categories of question, where general dermatology questions ($n = 22$) were answered well by all five models but skin surgery had much more variable performance. Although not many questions were in this category ($n = 5$) and will require larger studies to assess, these types of questions may require more critical thinking processes. Rather than simple factual recall, these question types often call for the ability to draw from clinical experience and synthesise an appropriate and justifiable course of action. While LLMs are effective at identifying relationships between datapoints and excel at factual recall, more distinctly human traits such as critical thinking and creativity are not easily achieved by AI yet [45].

LLMs are examples of generative artificial intelligence, and rely on extensive training on large datasets. Drawing on data and text, LLMs can extract and construct billions of parameters to generate realistic and coherent texts. Its major advantage is its ability to generate conversational responses. Through its architecture, it can understand natural language input from users to mimic human interactions. The performance variation between models and iterations primarily lies in the number of parameters the LLMs were trained to process. This refers to the variables and weights derived by the model based on its dataset, where an increasing number of parameters should allow the processing and output of more complex output. Parameters essentially determine the behaviour of the models themselves and the number of parameters is often not disclosed for LLMs, but the later iterations are likely in the range of 100s of billions [46]. It is important to note that while more parameters may potentially translate to more capable models, it may also cause over-fitting, leading to poor generalisation in response to unseen inputs. Regardless, the increasing amount of computing power required to train and operate these high-performance models means that consumer-grade, locally-run, open-sourced LLMs, such as Llama and Vicuna, are often less powerful. These are often in the range of 5–50 billion parameters, with training on less extensive datasets, reducing their overall applicability outside of what they were specifically trained to perform. The leaps between LLM versions can be seen in how ChatGPT-4 significantly outperforms its predecessors,

such as version 3.5 [10,18]. These models assessed in this study are available in most geographic regions with little-to-no cost. As such, these models represent the same models that current medical learners and educators may be using.

The additional inclusion of image-based questions also offered useful insight into the current state of image analysis capabilities of the latest iterative of LLMs. Currently, AI-driven image recognition tools are already being implemented within the National Health Service of the United Kingdom [47]. These have shown promising results in detecting malignant lesions as part of the triaging process. However, these are models specifically designed and trained for dermatoscopy images and are not available outside of selected organisations. The assessment in this study evaluates the currently available models that can be accessed by all medical professionals at little-to-no upfront costs. Additionally, the architecture of LLMs differs and does not place significant emphasis on image analysis. While only four questions included images in this sample exam, this still provides a preliminary assessment, as previous iterations of LLMs were not able to process images. The overall performances were satisfactory, ranging from two to five LLMs being able to answer each of these questions. The same prompts when entered without the accompanying images resulted in LLM responses that requested the image. Without the images, LLMs attempted to offer a description of each of the available multiple-choice answers, including the epidemiology and macroscopic description of the lesion. This suggests that current iterations of LLMs are at least able to incorporate image analysis as part of their reasoning. However, the low number of questions involving images means that it is not possible to infer their overall performance as image analysis tools. Karampinis et al. previously explored the ability of ChatGPT-3.5 to analyse dermatoscopy images and diagnose various malignancies and cancer mimickers [7]. Though challenged by more ambiguous and complex presentations, human assessors generally agreed with the diagnoses and explanations. Specifically, this study also evaluated the use of dermatopathological language in the LLMs' responses, which was shown to be variable at times. Regardless, their findings are also supported by this study, where LLMs can incorporate dermatoscopic images with scenarios, descriptions, and other clinical information into their analysis process. Similar studies have also been completed in ophthalmology, which demonstrated variable performance [36]. Nonetheless, the findings of this study add to the ability of these AI modalities to compute and process multiple-choice and scenario-based questions.

Within medicine, AI and LLMs can be applied in a variety of areas, including summarising patient encounters, procuring evidence to support clinical decision-making, or being used to formulate diagnoses [48]. Their application to the typically structured and standardised medical documentation and discharge summaries means that with appropriate prompts, generative AI can create summaries very efficiently [49]. Similarly, prompts with lists of symptoms can be processed by LLMs to provide a list of differentials and investigations, applicable to both acute and chronic diseases alike [50]. Other implementations of AI include the efficient and accurate processing of radiological and macroscopic dermatoscopy images [47,51].

The role and long-term impact of AI in healthcare remain to be determined. To yield these tools appropriately, whether to facilitate medical education for professionals or patients, clinicians should appreciate both its benefits and its pitfalls. When instructed appropriately with prompt engineering, generative AI can provide justification and explanation to facilitate the education and clinical understanding of doctors. Similarly, these may potentially be used to construct revision aids and question sets to facilitate factual recall. Together, these uses can provide a valuable bridge between traditional medical education and technology. Despite advancements in training datasets, fine-tuning, and architectures, misinformation and hallucinations still represent an important risk to consider [52]. These likely originated from the indiscriminate use of unfiltered data, including satirical or biased content, which can be rectified in newer iterations. Unlike clinical knowledge and constructing causal relationships between data, LLMs may also face challenges in achieving the empathy and judgement expected by people. While the empathy demonstrated by LLMs

remains to be fully realised and may possess clinical knowledge and understanding of factual relationships, LLMs have yet to replicate the humanity of healthcare [53–55]. Finally, the intensive training of AI and LLMs may also be prone to over-fitting, where the models begin to learn the answers to a particular question, rather than operating based on a generalised knowledge set. This means that for the model to be trained effectively, overfitting should also be considered and mitigated appropriately with various techniques [56].

Although this study is the first to compare the performance of popular LLMs in the Dermatology SCE, it has several limitations. This includes the rapid development of LLMs and the non-transparency of the dataset used in training each of the LLMs. Like similar studies, the findings may not be representative of newer LLMs as improvements in the available dataset for training and architectural developments may significantly alter their capabilities. This was observed when drawing comparisons between different versions of ChatGPT in different studies [10,18]. None of the tested LLMs were specifically trained, fine-tuned, or designed to answer medical questions; however, usage by other users on similar questions may feed into their training data. This was not apparent as the LLMs still answered some questions incorrectly, suggesting that even if previously fed these question–answer pairs, they did not fully determine the response outputs in this study. Additionally, the selection of commercial LLMs was based on their ability to operate at high levels using a significant amount of computing power that is typically not available to consumer-grade hardware; therefore, findings may not necessarily be replicable in local LLMs, even with appropriate training. Additionally, some LLMs provided detailed explanations and reasoning for their answers, but this was not reviewed nor considered in its performance. Finally, the inclusion of only the sample question bank may not always be representative of the level of difficulty expected of the true exam. Despite their limitations, LLMs continue to be promising medical education aids and their performance will likely further improve with the training and inclusion of medical research articles and texts.

5. Conclusions

The findings of this study demonstrated that commercially available LLMs can perform adequately in the Specialty Certificate Examination (SCE) in Dermatology. The performance of LLMs did differ significantly between models, but all models met the lower bounds of the historical pass marks of the exam. Future work should consider the integration of LLMs to improve medical education, evaluate their reasoning process, and correlate their efficacy with exam performance.

Author Contributions: Conceptualization, K.S.F. and K.H.F.; methodology, K.S.F. and K.H.F.; validation, K.S.F. and K.H.F.; formal analysis, K.S.F. and K.H.F.; investigation, K.S.F. and K.H.F.; data curation, K.S.F. and K.H.F.; writing—original draft preparation, K.S.F. and K.H.F.; writing—review and editing, K.S.F. and K.H.F.; visualization, K.S.F. and K.H.F.; supervision, K.S.F. and K.H.F.; project administration, K.S.F. and K.H.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Original dataset is publicly available.

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

References

1. Han, E.R.; Yeo, S.; Kim, M.J.; Lee, Y.H.; Park, K.H.; Roh, H. Medical education trends for future physicians in the era of advanced technology and artificial intelligence: An integrative review. *BMC Med. Educ.* **2019**, *19*, 1–15. [[CrossRef](#)] [[PubMed](#)]
2. Mogali, S.R. Initial impressions of ChatGPT for anatomy education. *Anat. Sci. Educ.* **2024**, *17*, 444–447. [[CrossRef](#)] [[PubMed](#)]

3. Abd-alrazaq, A.; AlSaad, R.; Alhuwail, D.; Ahmed, A.; Healy, P.M.; Latifi, S.; Aziz, S.; Damseh, R.; Alrazak, S.A.; Sheikh, J. Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions. *JMIR Med. Educ.* **2023**, *9*, e48291. [CrossRef] [PubMed]
4. Shamil, E.; Jaafar, M.; Fan, K.S.; Ko, T.K.; Schuster-Bruce, J.; Eynon-Lewis, N.; Andrews, P. The use of large language models like ChatGPT on delivering patient information relating to surgery. *Facial Plast. Surg.* **2024**. Available online: <https://www.thieme-connect.de/products/ejournals/abstract/10.1055/a-2413-3529> (accessed on 20 September 2024).
5. Gerke, S.; Minssen, T.; Cohen, G. Ethical and legal challenges of artificial intelligence-driven healthcare. In *Artificial Intelligence in Healthcare*; Academic Press: Cambridge, MA, USA, 2020; pp. 295–336. Available online: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7332220/> (accessed on 26 September 2024).
6. Kobayashi, K. Interactivity: A Potential Determinant of Learning by Preparing to Teach and Teaching. *Front. Psychol.* **2019**, *9*, 2755. [CrossRef] [PubMed]
7. Karampinis, E.; Toli, O.; Georgopoulou, K.-E.; Kampra, E.; Spyridonidou, C.; Schulze, A.-V.R.; Zafiriou, E. Can Artificial Intelligence “Hold” a Dermoscope?—The Evaluation of an Artificial Intelligence Chatbot to Translate the Dermoscopic Language. *Diagnostics* **2024**, *14*, 1165. [CrossRef] [PubMed]
8. Sumbal, A.; Sumbal, R.; Amir, A. Can ChatGPT-3.5 Pass a Medical Exam? A Systematic Review of ChatGPT’s Performance in Academic Testing. *J. Med. Educ. Curric. Dev.* **2024**, *11*, 23821205241238641. [CrossRef]
9. Safranek, C.W.; Sidamon-Eristoff, A.E.; Gilson, A.; Chartash, D. The Role of Large Language Models in Medical Education: Applications and Implications. *JMIR Med. Educ.* **2023**, *9*, e50945. [CrossRef]
10. Chan, J.; Dong, T.; Angelini, G.D. The performance of large language models in intercollegiate Membership of the Royal College of Surgeons examination. *Ann. R. Coll. Surg. Engl.* **2024**. [CrossRef]
11. Rossettini, G.; Rodeghiero, L.; Corradi, F.; Cook, C.; Pillastrini, P.; Turolla, A.; Castellini, G.; Chiappinotto, S.; Gianola, S.; Palese, A. Comparative accuracy of ChatGPT-4, Microsoft Copilot and Google Gemini in the Italian entrance test for healthcare sciences degrees: A cross-sectional study. *BMC Med. Educ.* **2024**, *24*, 694. [CrossRef]
12. Gilson, A.; Safranek, C.W.; Huang, T.; Socrates, V.; Chi, L.; Taylor, R.A.; Chartash, D. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med. Educ.* **2023**, *9*, e45312. [CrossRef]
13. Bhayana, R.; Krishna, S.; Bleakney, R.R. Performance of ChatGPT on a Radiology Board-style Examination: Insights into Current Strengths and Limitations. *Radiology* **2023**, *307*, 230582. [CrossRef]
14. Antaki, F.; Touma, S.; Milad, D.; El-Khoury, J.; Duval, R. Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of Its Successes and Shortcomings. *Ophthalmol. Sci.* **2023**, *3*, 100324. [CrossRef]
15. Vij, O.; Calver, H.; Myall, N.; Dey, M.; Kouranloo, K. Evaluating the competency of ChatGPT in MRCP Part 1 and a systematic literature review of its capabilities in postgraduate medical assessments. *PLoS ONE* **2024**, *19*, e0307372. [CrossRef] [PubMed]
16. General Medical Council. Dermatology Curriculum. 2023. Available online: <https://www.gmc-uk.org/education/standards-guidance-and-curricula/curricula/dermatology-curriculum> (accessed on 1 August 2024).
17. Membership of the Royal Colleges of Physicians of the United Kingdom. Specialty Certificate Examination (SCE) in Dermatology 2023 Selected Examination Metrics. 2024. Available online: https://www.thefederation.uk/sites/default/files/2024-02/Dermatology%20results%20report%202023_Liliana%20Chis.pdf (accessed on 1 August 2024).
18. Passby, L.; Jenko, N.; Wernham, A. Performance of ChatGPT on Specialty Certificate Examination in Dermatology multiple-choice questions. *Clin. Exp. Dermatol.* **2024**, *49*, 722–727. [CrossRef] [PubMed]
19. Membership of the Royal Colleges of Physicians of the United Kingdom. Dermatology | The Federation. Available online: <https://www.thefederation.uk/examinations/specialty-certificate-examinations/specialties/dermatology> (accessed on 1 August 2024).
20. OpenAI. GPT-4. Available online: <https://openai.com/gpt-4> (accessed on 1 August 2024).
21. Google. Gemini Models. Available online: <https://ai.google.dev/gemini-api/docs/models/gemini> (accessed on 1 August 2024).
22. Anthropic. Introducing Claude. Available online: <https://www.anthropic.com/news/introducing-claude> (accessed on 1 August 2024).
23. Microsoft. Microsoft Copilot | Microsoft AI. Available online: <https://www.microsoft.com/en-us/microsoft-copilot> (accessed on 1 August 2024).
24. Perplexity Frequently Asked Questions. Available online: <https://www.perplexity.ai/hub/faq> (accessed on 26 September 2024).
25. Hou, W.; Ji, Z. GPT-4V exhibits human-like performance in biomedical image classification. *bioRxiv* **2024**. Available online: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10802384/> (accessed on 26 September 2024).
26. Joh, H.C.; Kim, M.H.; Ko, J.Y.; Kim, J.S.; Jue, M.S. Evaluating the Performance of ChatGPT in Dermatology Specialty Certificate Examination-style Questions: A Comparative Analysis between English and Korean Language Settings. *Indian J. Dermatol.* **2024**, *69*, 338. [CrossRef] [PubMed]
27. Nicikowski, J.; Szczepański, M.; Miedziaszczyk, M.; Kudliński, B. The potential of ChatGPT in medicine: An example analysis of nephrology specialty exams in Poland. *Clin. Kidney J.* **2024**, *17*, 193. [CrossRef]
28. Meyer, A.; Riese, J.; Streichert, T. Comparison of the Performance of GPT-3.5 and GPT-4 With That of Medical Students on the Written German Medical Licensing Examination: Observational Study. *JMIR Med. Educ.* **2024**, *10*, e50965. [CrossRef]

29. Birkett, L.; Fowler, T.; Pullen, S. Performance of ChatGPT on a primary FRCA multiple choice question bank. *Br. J. Anaesth.* **2023**, *131*, e34–e35. [\[CrossRef\]](#)
30. Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S.S.; Wei, J.; Chung, H.W.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; Payne, P.; et al. Large language models encode clinical knowledge. *Nature* **2023**, *620*, 172. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Sallam, M.; Al-Salahat, K. Below average ChatGPT performance in medical microbiology exam compared to university students. *Front. Educ.* **2023**, *8*, 1333415. [\[CrossRef\]](#)
32. Shamil, E.; Ko, T.K.; Fan, K.S.; Schuster-Bruce, J.; Jaafar, M.; Khwaja, S.; Eynon-Lewis, N.; D'Souza, A.R.; Andrews, P. Assessing the quality and readability of online patient information: ENT UK patient information e-leaflets vs responses by a Generative Artificial Intelligence. *Facial Plast. Surg.* **2024**. Available online: <https://www.thieme-connect.de/products/ejournals/abstract/10.1055/a-2413-3675> (accessed on 20 September 2024).
33. Humar, P.; Asaad, M.; Bengur, F.B.; Nguyen, V. ChatGPT Is Equivalent to First-Year Plastic Surgery Residents: Evaluation of ChatGPT on the Plastic Surgery In-Service Examination. *Aesthetic Surg. J.* **2023**, *43*, NP1085–NP1089. [\[CrossRef\]](#) [\[PubMed\]](#)
34. Kung, T.H.; Cheatham, M.; Medenilla, A.; Sillos, C.; De Leon, L.; Elepaño, C.; Madriaga, M.; Aggabao, R.; Diaz-Candido, G.; Maningo, J.; et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digit. Heal.* **2023**, *2*, e0000198. [\[CrossRef\]](#)
35. Ali, R.; Tang, O.Y.; Connolly, I.D.; Sullivan, P.L.Z.; Shin, J.H.; Fridley, J.S.; Asaad, W.F.; Cielo, D.; Oyelese, A.A.; Doberstein, C.E.; et al. Performance of ChatGPT and GPT-4 on Neurosurgery Written Board Examinations. *Neurosurgery* **2023**, *93*, 1353–1365. [\[CrossRef\]](#) [\[PubMed\]](#)
36. Masalkhi, M.; Ong, J.; Waisberg, E.; Lee, A.G. Google DeepMind's gemini AI versus ChatGPT: A comparative analysis in ophthalmology. *Eye* **2024**, *38*, 1412. [\[CrossRef\]](#)
37. Bahir, D.; Zur, O.; Attal, L.; Nujeidat, Z.; Knaanie, A.; Pikkil, J.; Mimouni, M.; Plopsky, G. Gemini AI vs. ChatGPT: A comprehensive examination alongside ophthalmology residents in medical knowledge. *Graefes Arch. Clin. Exp. Ophthalmol.* **2024**, *1–10*. [\[CrossRef\]](#)
38. Morreel, S.; Verhoeven, V.; Mathysen, D. Microsoft Bing outperforms five other generative artificial intelligence chatbots in the Antwerp University multiple choice medical license exam. *PLOS Digit. Heal.* **2024**, *3*, e0000349. [\[CrossRef\]](#) [\[PubMed\]](#)
39. Uppalapati, V.K.; Nag, D.S. A Comparative Analysis of AI Models in Complex Medical Decision-Making Scenarios: Evaluating ChatGPT, Claude AI, Bard, and Perplexity. *Cureus* **2024**, *16*, e52485. [\[CrossRef\]](#) [\[PubMed\]](#)
40. Torres-Zegarra, B.C.; Rios-Garcia, W.; Ñaña-Cordova, A.M.; Arteaga-Cisneros, K.F.; Chalco, X.C.B.; Ordoñez, M.A.B.; Rios, C.J.G.; Godoy, C.A.R.; Quezada, K.L.T.P.; Gutierrez-Arratia, J.D.; et al. Performance of ChatGPT, Bard, Claude, and Bing on the Peruvian National Licensing Medical Examination: A cross-sectional study. *J. Educ. Eval. Health Prof.* **2023**, *20*, 30. [\[CrossRef\]](#)
41. Yu, P.; Fang, C.; Liu, X.; Fu, W.; Ling, J.; Yan, Z.; Jiang, Y.; Cao, Z.; Wu, M.; Chen, Z.; et al. Performance of ChatGPT on the Chinese Postgraduate Examination for Clinical Medicine: Survey Study. *JMIR Med. Educ.* **2024**, *10*, e48514. [\[CrossRef\]](#)
42. Noda, M.; Ueno, T.; Koshu, R.; Takaso, Y.; Shimada, M.D.; Saito, C.; Sugimoto, H.; Fushiki, H.; Ito, M.; Nomura, A.; et al. Performance of GPT-4V in Answering the Japanese Otolaryngology Board Certification Examination Questions: Evaluation Study. *JMIR Med. Educ.* **2024**, *10*, e57054. [\[CrossRef\]](#) [\[PubMed\]](#)
43. Alhur, A. Redefining Healthcare With Artificial Intelligence (AI): The Contributions of ChatGPT, Gemini, and Co-pilot. *Cureus* **2024**, *16*, e57795. [\[CrossRef\]](#) [\[PubMed\]](#)
44. Kaftan, A.N.; Hussain, M.K.; Naser, F.H. Response accuracy of ChatGPT 3.5 Copilot and Gemini in interpreting biochemical laboratory data a pilot study. *Sci. Rep.* **2024**, *14*, 8233. [\[CrossRef\]](#) [\[PubMed\]](#)
45. Amisha Malik, P.; Pathania, M.; Rathaur, V.K. Overview of artificial intelligence in medicine. *J. Fam. Med. Prim. Care* **2019**, *8*, 2328–2331. [\[CrossRef\]](#) [\[PubMed\]](#)
46. De Angelis, L.; Baglivo, F.; Arzilli, G.; Privitera, G.P.; Ferragina, P.; Tozzi, A.E.; Rizzo, C. ChatGPT and the rise of large language models: The new AI-driven infodemic threat in public health. *Front. Public Heal.* **2023**, *11*, 1166120. Available online: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10166793/> (accessed on 26 September 2024).
47. Thomas, L.; Hyde, C.; Mullarkey, D.; Greenhalgh, J.; Kalsi, D.; Ko, J. Real-world post-deployment performance of a novel machine learning-based digital health technology for skin lesion assessment and suggestions for post-market surveillance. *Front. Med.* **2023**, *10*, 1264846. [\[CrossRef\]](#)
48. Fan, K.S. *Advances in Large Language Models (LLMs) and Artificial Intelligence (AI)*; AtCAD: London, UK, 2024; Available online: <https://atomicademia.com/articles/implications-of-large-language-models-in-medical-education.122/> (accessed on 20 September 2024).
49. Patel, S.B.; Lam, K. ChatGPT: The future of discharge summaries? *Lancet Digit. Heal.* **2023**, *5*, e107–e108. [\[CrossRef\]](#)
50. Kumar, Y.; Koul, A.; Singla, R.; Ijaz, M.F. Artificial intelligence in disease diagnosis: A systematic literature review, synthesizing framework and future research agenda. *J. Ambient. Intell. Humaniz. Comput.* **2023**, *14*, 8459–8486. [\[CrossRef\]](#)
51. Hosny, A.; Parmar, C.; Quackenbush, J.; Schwartz, L.H.; Aerts, H.J.W.L. Artificial intelligence in radiology. *Nat. Rev. Cancer* **2018**, *18*, 500. [\[CrossRef\]](#) [\[PubMed\]](#)
52. Walker, H.L.; Ghani, S.; Kuemmerli, C.; Nebiker, C.A.; Müller, B.P.; Raptis, D.A.; Staubli, S.M. Reliability of Medical Information Provided by ChatGPT: Assessment Against Clinical Guidelines and Patient Information Quality Instrument. *J. Med. Internet Res.* **2023**, *25*, 1–14. [\[CrossRef\]](#)

53. Howe, P.D.L.; Fay, N.; Saletta, M.; Hovy, E. ChatGPT's advice is perceived as better than that of professional advice columnists. *Front. Psychol.* **2023**, *14*, 1281255. [[CrossRef](#)] [[PubMed](#)]
54. Elyoseph, Z.; Hadar-Shoval, D.; Asraf, K.; Lvovsky, M. ChatGPT outperforms humans in emotional awareness evaluations. *Front. Psychol.* **2023**, *14*, 1199058. [[CrossRef](#)] [[PubMed](#)]
55. Jeffrey, D. Empathy, sympathy and compassion in healthcare: Is there a problem? Is there a difference? Does it matter? *J. R. Soc. Med.* **2016**, *109*, 446–452. [[CrossRef](#)] [[PubMed](#)]
56. Charilaou, P.; Battat, R. Machine learning models and over-fitting considerations. *World J. Gastroenterol.* **2022**, *28*, 605. [[CrossRef](#)] [[PubMed](#)]

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