Chirurgia Narządów Ruchu i Ortopedia Polska – Polish Orthopaedics
ISSN 0009-479X, e-ISSN 2956-4719

Evaluation of the effectiveness of artificial intelligence models in the Polish State Specialization Exam in orthopedics and traumatology of the musculoskeletal system

Ocena skuteczności modeli sztucznej inteligencji w polskim Państwowym Egzaminie Specjalizacyjnym z ortopedii i traumatologii narządu ruchu

Alicja Główka[1*], Monika Hylinska[1*], Filip Robakowski[1*], Piotr Czarnecki[2]

[1] Student Scientific Club of Hand Surgery, University of Medical Sciences, Poznan, Poland
[2] Department of Traumatology, Orthopaedics and Hand Surgery, University of Medical Sciences, Poznań, Poland

* These authors contributed equally to the creation of this work.

ORIGINAL ARTICLE

Chir Narzadow Ruchu Ortop Pol, 2026; 91(1): 43-51

DOI: 10.31139/chnriop.2026.91.1.6

Level of evidence: II

Received: 19.02.2026
Accepted: 27.02.2026
Ahead of print: 09.03.2026
Published: 11.03.2026

Correspondence:
Alicja Główka MS
Student Scientific Club of Hand Surgery, Department of Traumatology, Orthopaedics and Hand Surgery,
Str. 28 Czerwca 1956 r. No 135/147,
61-545 Poznań, Poland
e-mail: alicjaa.glowka@gmail.com

Cite this article:
Główka A, Hylinska M, Robakowski F, Czarnecki P. Evaluation of the performance of artificial intelligence models on the Polish State Specialization Examination in orthopedics and traumatology of the musculoskeletal system. Chir Narzadow Ruchu Ortop Pol, 2026; 91(1): 43-51.
DOI: 10.31139/chnriop.2026.91.1.6

OPEN ACCESS
CC BY NC ND

Abstract

Introduction. The rapid development of language models such as ChatGPT and Google Gemini, and their growing interest in medicine, has prompted an analysis of their potential in the context of specialty examinations. Their use in orthopaedics and traumatology, which require extensive theoretical and clinical knowledge, is particularly significant. This study aimed to evaluate whether ChatGPT-5 and Google Gemini could pass the Polish orthopaedics and traumatology specialisation examination using official questions from the Centre for Medical Examinations (CEM, Poland).
Materials and methods. We analysed 468 validated questions from four Polish State Specialization Exams in orthopaedics and traumatology of the musculoskeletal system (2024–2025) and classified them into different categories. ChatGPT-5 and Gemini 2.5 answered all items in standardised, independent sessions. Accuracy was compared with the official answer key, inter-rater agreement assessed using Cohen's kappa, and group differences evaluated with Chi-square/Fisher's tests, Mann–Whitney U tests, and t-tests ($\alpha = 0.05$).
Results. ChatGPT-5.0 achieved consistently high accuracy across question types, cognitive domains, image versus text formats, and orthopaedic subspecialties, outperforming Gemini 2.5 in all comparisons. In four examination sessions, ChatGPT-5.0 reached performance comparable to or exceeding that of physicians, although the best performance in most sessions was achieved by a physician. Overall, ChatGPT-5.0 attained 80.2% accuracy, significantly higher than both Gemini 2.5 (75.7%) and physicians (74.3%), whereas Gemini 2.5 showed greater variability across domains and lower mean performance.
Conclusions. Rapidly evolving AI models, such as ChatGPT-5.0, demonstrate high scores in orthopaedic examinations, highlighting their potential for education and clinical support. However, further research is required to evaluate their safety, limitations, and integration into clinical practice.

Key words: artificial intelligence, orthopaedics, natural language processing, education, specialty medical examination, medical education

Streszczenie

Wstęp: Dynamiczny rozwój modeli językowych, takich jak ChatGPT i Google Gemini oraz rosnące zainteresowanie ich zastosowaniem w medycynie stanowią podstawę do analizy ich potencjału w kontekście egzaminów specjalizacyjnych. Ich wykorzystanie w ortopedii i traumatologii, które wymagają rozległej wiedzy teoretycznej i klinicznej, jest szczególnie istotne. Celem niniejszego badania była ocena, czy ChatGPT-5 i Google Gemini są zdolne do zdania polskiego egzaminu specjalizacyjnego z ortopedii i traumatologii, wykorzystując oficjalne pytania pochodzące z Centrum Egzaminów Medycznych (CEM, Polska).
Materiał i metody: Przeanalizowano 468 zweryfikowanych pytań z czterech polskich państwowych egzaminów specjalizacyjnych (PES) z ortopedii i traumatologii narządu ruchu (2024–2025) i sklasyfikowano je do różnych kategorii. ChatGPT-5.0 oraz Gemini 2.5 udzieliły odpowiedzi na wszystkie pytania w ustandaryzowanych, niezależnych sesjach. Trafność odpowiedzi porównano z oficjalnym kluczem odpowiedzi, zgodność między oceniającymi oceniono przy użyciu współczynnika kappa Cohena, a różnice między grupami analizowano z zastosowaniem testu chi-kwadrat/testu dokładnego Fishera, testu U Manna–Whitneya oraz testu t-Studenta ($\alpha = 0,05$).
Wyniki: ChatGPT-5.0 osiągnął konsekwentnie wysoką trafność odpowiedzi we wszystkich typach pytań, domenach poznawczych, formatach (obrazy vs. tekst) oraz podspecjalizacjach ortopedycznych, przewyższając Gemini 2.5 we wszystkich porównaniach. W czterech sesjach egzaminacyjnych ChatGPT-5.0 uzyskał wyniki porównywalne lub wyższe niż lekarze, chociaż w większości sesji najwyższy wynik osiągnął lekarz. Średnio ChatGPT-5.0 uzyskał 80,2% poprawnych odpowiedzi, co stanowiło wynik istotnie wyższy zarówno w porównaniu z Gemini 2.5 (75,7%), jak i lekarzami (74,3%), podczas gdy Gemini 2.5 wykazywał większą zmienność między domenami oraz niższą średnią skuteczność.
Wnioski: Dynamicznie rozwijające się modele sztucznej inteligencji, takie jak ChatGPT-5.0, wykazują wysokie wyniki w egzaminach z ortopedii, co podkreśla ich potencjał w edukacji i wsparciu klinicznym. Konieczne są jednak dalsze badania w celu oceny ich bezpieczeństwa, ograniczeń oraz możliwości integracji z praktyką kliniczną.

Słowa kluczowe: sztuczna inteligencja, ortopedia, przetwarzanie języka naturalnego, edukacja, specjalistyczne egzaminy medyczne, kształcenie medyczne

Introduction

Artificial intelligence (AI) has advanced rapidly over the past decade, particularly after Google introduced the transformer architecture in 2017, which laid the foundation for modern large language models (LLMs) [1]. The public release of ChatGPT-3.5 by OpenAI in November 2022 marked the start of an unprecedented AI boom, with the application reaching over 100 million users in less than two months [2]. Since then, multimodal models such as GPT-4, GPT-5, and Google Gemini have become widely available to both individual and professional users. As access to these tools has grown, LLMs have found applications not only in education, programming, and data analysis but also in law, business, public administration, and healthcare [3, 4]. The public's use of generative AI has increased substantially in the last year. A 2025 cross-national study conducted in six countries (Argentina, Denmark, France, Japan, the UK, and the US) found that the proportion of individuals who had ever used a standalone generative AI system such as ChatGPT rose from 40% in 2024 to 61% in 2025, with weekly usage nearly doubling from 18% to 34%. That same report indicates that 29% of respondents trust ChatGPT more than Gemini (18%) [5].

In medicine, interest in LLMs is growing particularly rapidly. These models are being studied for their usefulness in analysing clinical text and images, document summarisation, supporting medical education, and automating administrative work. The latest generations of LLMs, such as ChatGPT-5 (OpenAI) and Google Gemini, are characterised by significantly more parameters, better multimodality, and more stable reasoning than earlier versions of these models [6–8]. Their testing also includes attempts to solve medical exams ranging from the United States Medical Licensing Examination (USMLE) to various European specialty exams. Studies have shown that GPT-4 models achieve scores above the passing threshold on the USMLE exam. In Poland, they have achieved scores above the passing threshold in specialisation exams in dermatology [4], pathology [7], vascular surgery [8], infectious diseases [9], ophthalmology [10], endocrinology [11], and psychiatry [12, 13].

Orthopaedics and traumatology require extensive knowledge in a range of areas, including anatomy, biomechanics, trauma surgery, orthopaedic oncology, and paediatrics. The specialisation exam in this field includes both theoretical questions and the interpretation of imaging studies, as well as complex clinical cases, making it an interesting tool for assessing the capabilities of AI systems. To date, no publications have analysed whether the latest LLMs, ChatGPT-5 and Google Gemini, can meet the requirements of the Polish specialisation exam in orthopaedics and traumatology.

Guiding principles and the study's objective

This study aims to evaluate whether the ChatGPT-5 and Google Gemini models can meet the requirements of the Polish State Specialization Exam (PSSE) in orthopaedics and traumatology, using official Centre for Medical Examinations (Centrum Edukacji Medycznej – CEM, Poland) examination questions.

Materials and methods

To assess the potential competence of the ChatGPT-5 and Gemini 2.0 language models in the context of the PES, a retrospective analysis of official examination sets from 2024–2025 was conducted. The PSSE in Orthopaedics is organised by the CEM twice a year – in spring and autumn sessions – and each exam contains 120 test questions of type A (with one correct answer) and type K (multiple-choice test with one correct set of statements), in accordance with the CEM regulations. Questions from the last four exam sessions were selected for analysis, resulting in an initial pool of 480 questions. Prior to analysis, one question formally cancelled by the CEM, and 11 questions considered obsolete by an independent orthopaedic specialist, after taking into account comments published on the CEM website and an assessment of their compliance with the current state of medical knowledge, were excluded. After selection, a final pool of 468 unique questions was obtained, which were then divided by the researchers thematically into one of 12 categories: Arthroscopy and intra-articular procedures; Basics of Orthopaedics and Traumatology; Congenital and genetic defects; Deformities of the limbs; Fractures and dislocations; Infections of the musculoskeletal system; Metabolic disorders and bone diseases; Nerve, soft-tissue, and tendon pathologies; Paediatric orthopaedics; Prosthetics; Spine; Bone and soft-tissue tumours.

In addition, each question was assigned to one of two cognitive categories: memory (M), involving the direct recall of previously learned information, and critical thinking (C), which required understanding clinical problems and reasoning. Each question was assigned a difficulty index from the official CEM database. It measures how easy a question is by comparing the number of correct answers in the highest- and lowest-scoring groups to the number of examinees in each group; high values indicate an easy item, while low values reflect a difficult one.

Each set of questions was presented by two independent researchers using the ChatGPT-5 and Gemini 2.0 models, using separate user accounts that had no prior exposure to PES materials or interaction with the models in this context. To avoid carryover effects from prior interactions, each

session was initiated in a new chat window and consisted of a maximum of 15 questions. The same introductory instruction in Polish was used for each question: "You are an expert in orthopaedics. Answer the following question as if you were taking a specialist exam, choosing only one correct answer. Answer briefly, without giving reasons. Then, on a scale of 0 to 100%, indicate how confident you are in your answer." The model's answers were recorded and evaluated for consistency with the official CEM answer key. Data were collected between 29 October and 4 November 2025. Independent repetition of the test by two researchers allowed for additional evaluation of the repeatability and consistency of the model's answers under independent trial conditions.

Statistical analyses

All statistical analyses were conducted using R (version 4.5.2, 31 October 2025), JASP (version 0.95.4), MedCalc (version 23.4.0), and Microsoft Excel. Inter-rater reliability was assessed using Cohen's kappa to quantify agreement between independent raters for each AI model. Differences in categorical accuracy across groups, question types, and categories were analysed with Chi-squared tests or Fisher's Exact Test when expected counts were low. The association between continuous variables (e.g., confidence, difficulty) and correctness was examined using the Mann–Whitney U test, with effect sizes reported as rank-biserial correlations. Odds ratios (OR) with 95% confidence intervals were calculated to evaluate the magnitude of group differences. Comparisons of AI models with physician performance were conducted using independent-samples t-tests. Statistical significance was considered at $p < 0.05$.

Results

Inter-rater Agreement for GPT-5.0 and Gemini 2.5

The agreement between two independent raters in evaluating GPT-5 responses was considerable. Out of 468 questions, the raters agreed on 375 cases (≈80.1%). Cohen's kappa was 0.764 (95% CI: 0.690–0.839), indicating substantial agreement (Landis & Koch, 1977) and reliable assessment of response correctness. For Gemini 2.5, agreement was slightly lower: 350 out of 468 cases (≈74.8%) were rated consistently, with a kappa of 0.663 (95% CI: 0.584–0.742), also representing substantial agreement. Overall, both models demonstrated consistent inter-rater agreement, with GPT-5 showing slightly greater consistency.

In a further analysis, the responses of one of the researchers were randomly evaluated and compared with those of the entire group (i.e., two researchers) for each AI model separately.

Relationship between response accuracy and the "A" vs "K" categories

The chi-square test was used for the analysis, and the results are presented in Table 1. For ChatGPT-5.0 (single rater), 81.2% of responses in group A were correct, compared with 77.2% in group K. The Chi-square analysis yielded $\chi^2 = 0.959$ ($p = 0.3275$), indicating no statistically significant difference between groups. The odds ratio (OR) was 1.28 (95% CI: 0.78–2.10), suggesting a slight, non-significant advantage for group A.

In the combined ratings analysis, 81.4% of responses in group A were correct versus 77.2% in group K, with $\chi^2 = 2.069$ ($p = 0.1503$) and OR = 1.29 (95% CI: 0.91–1.84), similarly indicating a small, non-significant difference between groups. These results suggest that ChatGPT-5.0 maintains consistent accuracy irrespective of group context.

For Gemini 2.5 (single rater), accuracy was identical in both groups (72.9%), with $\chi^2 = 0.0$ ($p = 0.9959$) and OR = 1.00 (95% CI: 0.63–1.60), indicating no difference between groups. In the combined ratings analysis, group A achieved 71.7% correct responses compared with 73.3% in group K. The Chi-square test indicated $\chi^2 = 0.222$ ($p = 0.015$), suggesting a statistically significant difference; however, the OR of 0.92 (95% CI: 0.66–1.29) indicates that the practical difference in accuracy between groups is minimal.

Overall, ChatGPT-5.0 demonstrated stable performance across groups, whereas Gemini 2.5 showed slight variability in combined ratings. Despite the statistical significance for

Table 1. Relationship between responseaccuracy and the "A" vs "K" categories

AI Model	Group	Correct	Incorrect	Total	(Chi²)	*p*	OR (95% CI)
ChatGPT-5.0 single rater	A	277 (81%)	64 (19%)	341	0,959	0,3275	1.28 (0.78–2.10)
	K	98 (77%)	29 (23%)	127			
ChatGPT-5.0 combined ratings	A	555 (81%)	127 (19%)	682	2,069	0,1503	1.29 (0.91–1.84)
	K	196 (77%)	58 (23%)	254			
Gemini 2.5 single rater	A	86 (73%)	32 (27%)	118	0	0,9959	1.00 (0.63–1.60)
	K	255 (73%)	95 (27%)	350			
Gemini 2.5 combined ratings	A	167 (72%)	66 (28%)	233	0,222	0,015	0.92 (0.66–1.29)
	K	515 (73%)	188 (27%)	703			

Table. 2. Relationship between response accuracy and the the "M (memory)" vs "C (critical thinking)" categories

AI Model	Group	Correct	Incorrect	Total	Chi^2	*p*	OR (95% CI)
ChatGPT-5.0 single rater	M	226 (79%)	61 (21%)	287	0.889	0.3458	0.80 (0.50–1.28)
	C	149 (82%)	32 (18%)	181			
ChatGPT-5.0 combined ratings	M	456 (79%)	118 (21%)	574	0.587	0.4435	0.88 (0.63–1.23)
	C	295 (82%)	67 (18%)	362			
Gemini 2.5 single rater	M	220 (77%)	67 (23%)	287	1.371	0.2416	1.29 (0.91–1.83)
	C	130 (72%)	51 (28%)	181			
Gemini 2.5 combined ratings	M	436 (76%)	138 (24%)	574	0.575	0.4484	1.12 (0.85–1.47)
	C	267 (74%)	95 (26%)	362			

Table. 3. Relationship between response accuracy and the "with image" vs "without image" categories

AI Model	Group	Correct (n, %)	Incorrect (n, %)	Total	χ^2	*p*	OR (95% CI)
ChatGPT-5.0 single rater	with image	16 (80%)	4 (20%)	20	0.000	0,9883	0.99 (0.33–3.06)
	without image	359 (80%)	89 (20%)	448			
ChatGPT-5.0 combined ratings	with image	30 (75%)	10 (25%)	40	0.721	0,028	0.73 (0.43–1.50)
	without image	721 (81%)	175 (19.%)	896			
Gemini 2.5 single rater	with image	13 (65%)	7 (35%)	20	1.059	0,3035	0.61 (0.39–2.53)
	without image	337 (75%)	111 (25%)	448			
Gemini 2.5 combined ratings	with image	28 (70%)	12 (30%)	40	0.582	0,025	0.76 (0.50–1.97)
	without image	675 (75%)	221 (25%)	896			

Gemini 2.5 in combined ratings, the effect size is small, and the confidence intervals include 1, indicating limited practical relevance.

Relationship between response accuracy and the "M (memory)" vs "C (critical thinking)" categories

The chi-square test was used for the analysis, and the results are presented in Table 2. For ChatGPT-5.0, single-rater analysis showed 79% correct responses for the Memory category and 82% correct for the Critical Thinking category. The difference between categories was not statistically significant (χ^2 = 0.889, p = 0.346; OR = 0.80, 95% CI: 0.50–1.28), indicating similar performance across categories.

Combined ratings yielded 79% correct for Memory and 82% for Critical Thinking, with no significant difference (χ^2 = 0.587, p = 0.444; OR = 0.88, 95% CI: 0.63–1.23), confirming consistent performance.

For Gemini 2.5, single-rater analysis showed 77% correct for Memory and 72% for Critical Thinking, with no significant difference (χ^2 = 1.371, p = 0.242; OR = 1.29, 95% CI: 0.91–1.83). Combined ratings also showed similar performance across categories, 76% correct for Memory vs. 74% for Critical Thinking (χ^2 = 0.575, p = 0.448; OR = 1.12, 95% CI: 0.85–1.47).

In summary, both models performed consistently in both the factual and analytical reasoning categories.

Relationship between response accuracy and the "with image" vs "without image" categories

The chi-square test was used for the analysis, and the results are presented in Table 3. For ChatGPT-5.0, single-rater accuracy was 80% for both image- and text-based questions (χ^2 = 0.000, df = 1, p = 0.9883, contingency coefficient C = 0.001, OR = 0.99, 95% CI: 0.33–3.06). Combined ratings showed 75% accuracy for image-based versus 80.5% for text-only questions (χ^2 = 0.721, df = 1, p = 0.028, C = 0.028, OR = 0.73, 95% CI: 0.43–1.50), indicating a slight but statistically detectable decrease in odds for image-based items.

For Gemini 2.5, single-rater accuracy was 65% for image-based and 75.2% for text-only questions (χ^2 = 1.059, df = 1, p = 0.3035, C = 0.048, OR = 0.61, 95% CI: 0.39–2.53), and combined ratings were 70% versus 75% (χ^2 = 0.582, df = 1, p = 0.025, C = 0.025, OR = 0.76, 95% CI: 0.50–1.97).

Overall, both models performed consistently across question types, and the low contingency coefficients indicate that the practical effect of including images on response correctness is minimal, despite occasional statistical significance.

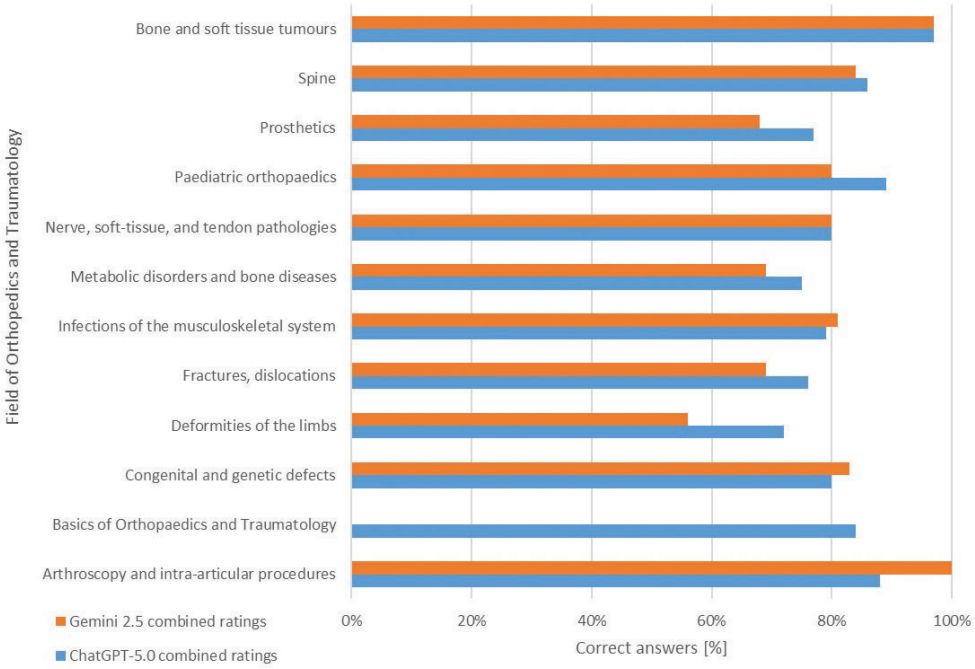


Fig. 1. Relationship between response accuracy and field of orthopeadics and traumatology

Relationship between response accuracy and the field of orthopaedics and traumatology

A Fisher's Exact Test was used to assess the relationship between question category within orthopaedics and traumatology and the correctness of responses generated by ChatGPT-5.0 and Gemini 2.5 (see Fig. 1). For ChatGPT-5.0 (single rater), the proportion of correct responses ranged from 75% for Metabolic Disorders and Bone Diseases to 100% for Bone and Soft Tissue Tumours, with no significant association between category and correctness ($p = 0.4913$). In the combined ratings, accuracy ranged from 72% for Deformities of the Limbs to 97% for Bone and Soft Tissue Tumours, showing a statistically significant association ($p = 0.0289$).

For Gemini 2.5, category-specific variation was more pronounced. Correct response rates ranged from 56% for Deformities of the Limbs to 100% for Arthroscopy and Intra-articular Procedures, and Fisher's test confirmed a significant association in both the single-rater analysis ($p = 0.0115$) and the combined evaluation ($p = 0.0001$). Overall, both models demonstrated category-dependent accuracy, with Gemini 2.5 showing greater variability across question types than ChatGPT-5.0.

Relationship between confidence levels and answer correctness

The relationship between model-reported confidence (0–1 scale) and answer correctness was evaluated for ChatGPT 5.0 and Gemini 2.5 using the Mann–Whitney U test. For ChatGPT 5.0, the single-rater analysis showed significantly higher confidence for correct answers than for incorrect ones ($U = 20{,}819$, $p < 0.001$), and the combined-rater analysis confirmed the same pattern ($U = 20{,}819$, $p < 0.001$) (see Fig. 2.). For Gemini 2.5, the single-rater analysis similarly demonstrated higher confidence for correct responses compared with incorrect responses ($U = 24{,}457$, $p < 0.001$), with the combined-rater analysis again supporting this finding ($U = 93{,}022$, $p < 0.001$) (see Fig. 3.).

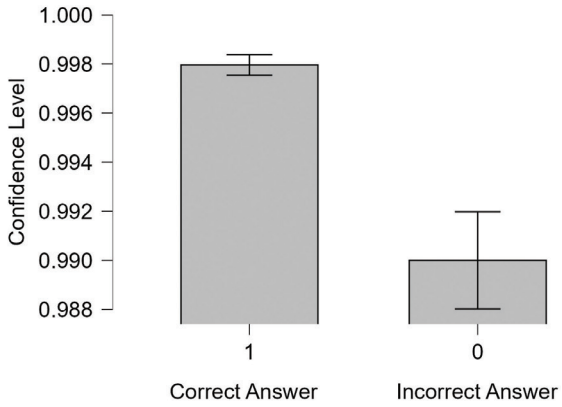


Fig. 2. Relationship between confidence levels and answer correctness for ChatGPT-5.0 combined raters

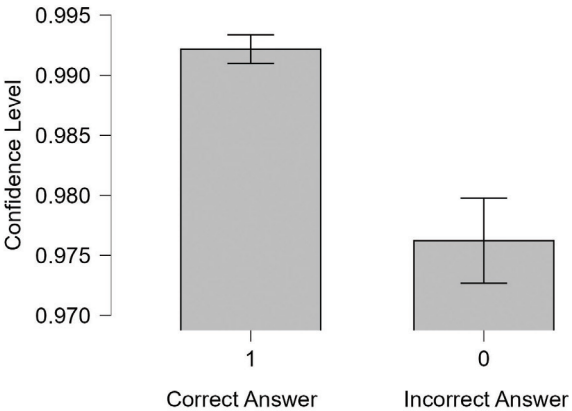


Fig. 3. Relationship between confidence levels and answer correctness for Gemini 2.5 combined raters

All differences were statistically significant, indicating that correct responses were systematically associated with higher confidence. However, median confidence values were very high across all groups (close to 1), meaning that the observed differences – although statistically robust – were subtle in magnitude and may not be easily noticeable without formal statistical testing.

Relationship between difficulty index and answer correctness

The relationship between question difficulty and answer correctness was examined using the Mann–Whitney U test for both ChatGPT-5.0 and Gemini, evaluated by one or two raters. For ChatGPT-5.0 (single rater), correct answers were associated with higher difficulty index values (indicating easier questions), and this difference was statistically significant ($U = 25{,}599$, $p < 0.001$). In the combined sample of ChatGPT-5.0 responses, correct answers again showed higher difficulty scores than incorrect ones ($U = 98{,}628$, $p < 0.001$).

Gemini demonstrated a comparable pattern: the single-rater analysis yielded $U = 11{,}830$, $p < 0.001$, while the aggregated two-rater evaluation produced $U = 49{,}966$, $p < 0.001$. Across both models and rater conditions, correct answers consistently corresponded to easier questions, with all differences reaching statistical significance.

Best performance of AI models and physicians

Figure 4. presents the maximum scores achieved by ChatGPT-5.0, Gemini 2.5, and physicians across four sessions. Physicians attained the highest scores in three sessions (Autumn 2025, Autumn 2024, and Spring 2024), demonstrating their peak performance. ChatGPT-5.0 outperformed both Gemini 2.5 and physicians in one session (Spring 2025), demonstrating that the model can occasionally reach or exceed human-level performance. Across all sessions, ChatGPT-5.0 consistently achieved higher maximum scores than Gemini 2.5, highlighting its relative advantage. Gemini 2.5 consistently had lower

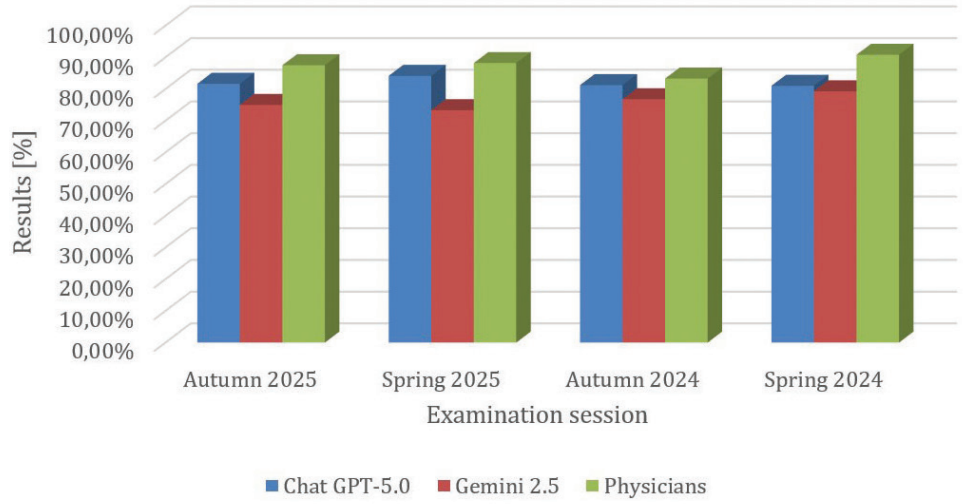


Fig. 4. Best performance of AI models and physicians

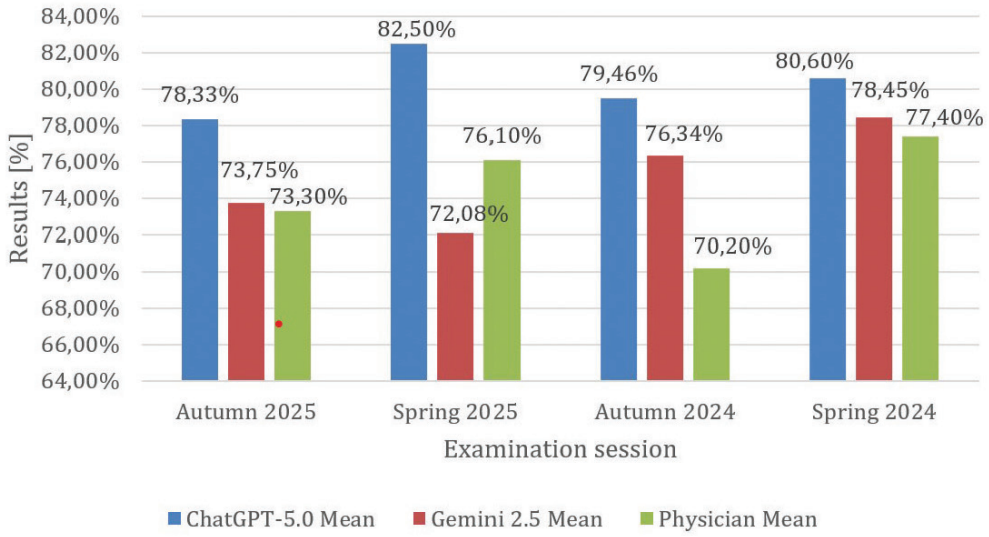


Fig. 5. Performance Comparison of ChatGPT-5.0, Gemini 2.5, and Physicians

maximum scores compared to both ChatGPT-5.0 and physicians, suggesting it is less competitive at its peak. These results indicate that while physicians generally maintain the highest top performance, ChatGPT-5.0 can rival human experts in specific instances.

Performance Comparison of ChatGPT-5.0, Gemini 2.5, and Physicians

Across all examination sessions, ChatGPT-5.0 consistently achieved higher mean scores than both Gemini 2.5 and the physicians (see Fig. 5.). In the Autumn 2025 session, ChatGPT-5.0 reached a mean accuracy of 78.33%, compared with 73.75% for Gemini 2.5 and 73.30% for physicians. The comparison showed no statistically significant difference between ChatGPT-5.0 and physicians ($p = 0.063$), Gemini 2.5 and physicians ($p = 0.944$), or ChatGPT-5.0 and Gemini 2.5 ($p = 0.186$).

In the Spring 2025 session, ChatGPT-5.0 again outperformed both comparators (82.50% vs 72.08% and 76.10%). The difference between ChatGPT-5.0 and physicians did not reach significance ($p = 0.062$), nor did the comparison between Gemini 2.5 and physicians ($p = 0.197$). However, ChatGPT-5.0 scored significantly higher than Gemini 2.5 ($p = 0.001$).

In the Autumn 2024 session, ChatGPT-5.0 achieved 79.46%, Gemini 2.5 achieved 76.34%, and physicians scored 70.20%. ChatGPT-5.0 performed significantly better than physicians (p = 0.046), while the difference between Gemini 2.5 and physicians was not significant ($p = 0.177$). The comparison between the two AI models also showed no significant difference (p = 0.107).

In the Spring 2024 session, ChatGPT-5.0 achieved a mean score of 80.60%, Gemini 2.5 scored 78.45%, and physicians scored 77.40%. None of the comparisons reached statistical significance (ChatGPT-5.0 vs physicians: $p = 0.286$; Gemini 2.5 vs physicians: p = 0.586; ChatGPT-5.0 vs Gemini 2.5: $p = 0.067$).

When results from all sessions were combined, ChatGPT-5.0 demonstrated the highest overall accuracy (80.22%), outperforming both Gemini 2.5 (75.65%) and physicians (74.25%). The difference was statistically significant for ChatGPT-5.0 vs physicians ($p = 0.018$) and ChatGPT-5.0 vs Gemini 2.5 ($p = 0.009$). No significant difference was found between Gemini 2.5 and physicians in the overall analysis ($p = 0.274$).

Discussion

Language programs are currently being improved at a rapid pace, and research should continue to confirm their increasing usefulness in medical education and examination settings. Our results indicate that the latest generations of LLMs are capable of effectively handling complex theoretical problems, achieving outcomes comparable to – or even exceeding – the average performance of examinees. Such an advantage had already been demonstrated by the ChatGPT-4.0 model in the Turkish orthopaedic specialty examination [14]. Kung et al. demonstrated that GPT-4 achieved 73.6% correct answers on a typical Orthopaedic In-Training Examination (OITE) question set, corresponding to the performance level of a postgraduate year 5 (PGY-5) resident and surpassing the passing threshold for the written portion of the American Board of Orthopaedic Surgery (ABOS) examination (67%) [15].

The broader literature confirms that performance has steadily improved across LLM generations. A systematic review by Zong et al. showed that GPT-4 consistently outperformed older models across 198 medical examinations in multiple countries and languages [16]. Similar findings were reported by Singhal et al., who demonstrated that Med-PaLM 2 achieved 86.5 percent accuracy on the MedQA dataset [17]. In Poland, Pokrywka et al. analysed 297 official PES questions; results showed that GPT-3.5 failed all specialty exams, whereas GPT-4 passed approximately 75 percent of them [18]. Earlier studies in pathology and anatomy also revealed substantial jumps in accuracy from GPT-3.5 to GPT-4, with improvements from roughly 45–64 percent to more than 90 percent on some tasks [19, 20]. We can expect that future generations of artificial intelligence will answer examination questions with progressively higher accuracy.

Contrary to earlier reports [21], our article shows that LLMs did not demonstrate significant differences in answer accuracy depending on the complexity and cognitive demands of the questions. It suggests that higher-order reasoning performance in LLMs strongly depends on the specific model and examination format and has markedly improved in the newest generations.

Our study showed that there were no significant practical differences in response accuracy with and without images for both AI models. These results differ from previous studies – Takagi et al. (2024) showed that the multimodal GPT-4 performs significantly worse on image-based questions from the Japanese medical exam than on text-based questions [22]. Currently, LLMs are getting better at answering questions that include images. This suggests that nowadays they increasingly use not only pattern recognition but also integrated multimodal reasoning [23, 24].

At the subcategory level, both AI models showed some variability, but Gemini 2.5's accuracy appeared more sensitive to differences between domains. This pattern resembles earlier findings for GPT-4, suggesting that certain models may still be influenced by the cognitive complexity of specific orthopaedic subfields [15, 25]. In contrast, ChatGPT-5.0 demonstrated more stable performance across categories, indicating greater suitability as a broadly applicable educational tool.

It should be noted that the high performance of LLMs may partly reflect data leakage, where publicly available exam materials unintentionally appear in training datasets. As highlighted by Kaneko & Baldwin (2024) [26], such contamination can inflate scores and distort assessments of true model ability. The results of our study reflect the performance of artificial intelligence models with access to comprehensive global knowledge, including online resources. For a more balanced and informative comparison, it would be valuable to assess physicians against AI models restricted to closed medical knowledge databases, or alternatively, to allow physicians to use online resources as well. Therefore, comparisons between human and AI performance should be interpreted with caution.

It is worth noting that outdated questions were excluded from the AI models' question pool, so that, like physicians during the exam, the models responded only to questions that were current at the time. However, this introduces a difference in the sets of questions answered, and although the reported results are expressed as percentages, the non-identical question sets may be a source of some inequality in the comparisons.

The analysis revealed that raters disagreed on approximately one in four questions for the Gemini 2.5 model and on about one in five questions for ChatGPT-5.0, even when presented with identical questions. These results indicate that, while language models demonstrate considerable potential, there remains scope for improvement before they can be regarded as fully reliable and useful tools for educational purposes, where near-perfect reproducibility is required.

From a practical standpoint, our findings highlight the educational value of LLMs in orthopaedic training. These models can support exam preparation by generating questions, simulating clinical scenarios, and providing automated feedback, with prior research indicating strong multilingual performance [27, 28]. Nevertheless, despite their ability to pass written exams, LLMs lack practical skills, procedural judgement, and clinical responsibility. Their role should therefore remain supportive, complementing traditional training and not replacing medical professionals [28, 29].

Conclusions

Our study demonstrated that current artificial intelligence models, particularly ChatGPT-5.0, achieved results comparable to or even exceeding those of doctors taking the orthopaedic specialisation exam. These findings suggest that artificial intelligence has the potential to become a valuable tool for supporting education, clinical decision-making, and the standardisation of knowledge in orthopaedics. However, further research is needed to evaluate the safety, limitations, and practical integration of these systems into everyday medical practice.

References

1. Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need. Neural Information Processing Systems [Internet]. Curran Associates Inc.; 2017 Jun 12;30:5998–6008. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
2. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. Sci Rep. 2023 Nov 22;13(1):20512. doi: 10.1038/s41598-023-46995-z. PMID: 37993519; PMCID: PMC10665355.
3. Jaworski A, Jasiński D, Sławińska B, et al. GPT-4o vs. Human Candidates: Performance Analysis in the Polish Final Dentistry Examination. Cureus. 2024 Sep 6;16(9):e68813. doi: 10.7759/cureus.68813. PMID: 39371744; PMCID: PMC11456324.
4. Lewandowski M, Łukowicz P, Świetlik D, Barańska-Rybak W. ChatGPT-3.5 and ChatGPT-4 dermatological knowledge level based on the Specialty Certificate Examination in Dermatology. Clin Exp Dermatol. 2024 Jun 25;49(7):686–691. doi: 10.1093/ced/llad255. PMID: 37540015.
5. Simon FM, Nielsen RK, Fletcher R. Generative AI and news report 2025: How people think about AI's role in journalism and society. https://reutersinstitute.politics.ox.ac.uk/generative-ai-and-news-report-2025-how-people-think-about-ais-role-journalism-and-society doi:10.60625/RISJ-5BJV-YT69
6. Łomzik M. ChatGPT jako skuteczne narzędzie do upraszczania dokumentacji medycznej? Próba oceny wykorzystania sztucznej inteligencji do tłumaczenia zaleceń lekarskich na prosty język. (ChatGPT as an effective tool for simplifying medical documentation? An attempt to evaluate the use of artificial intelligence to translate medical recommendations into simple language) Poradnik Językowy. 2025;821(2):31–51.
7. Bielówka M, Kufel J, Rojek M, et al. An investigative analysis – ChatGPT's capability to excel in the Polish speciality exam in pathology. Pol J Pathol. 2024;75(3):236–240. doi: 10.5114/pjp.2024.143091. PMID: 39451178.
8. Loson-Kawalec M, Kowalczyk A, Tabor A, et al. Comparison of GPT-4o With Human Performance in the Polish Vascular Surgery Specialty Examination. Cureus. 2025 Sep 23;17(9):e93022. doi: 10.7759/cureus.93022. PMID: 41141174; PMCID: PMC12548761.
9. Błecha Z, Jasiński D, Jaworski A, et al. Performance of GPT-4o and DeepSeek-R1 in the Polish Infectious Diseases Specialty Exam. Cureus. 2025 Apr 23;17(4):e82870. doi: 10.7759/cureus.82870. PMID: 40416223; PMCID: PMC12102586.
10. Sławińska B, Jasiński D, Jaworski A, et al. Performance of the ChatGPT-4o Language Model in Solving the Ophthalmology Specialization Exam. Cureus. 2025 Jul 28;17(7):e88908. doi: 10.7759/cureus.88908. PMID: 40895978; PMCID: PMC12392048.
11. Latkowska A, Sawina P, Dolata T, et al. Assessment of the Ability of the ChatGPT-5 Model to Pass the Endocrinology Specialization Exam. Cureus. 2025 Sep 29;17(9):e93479. doi: 10.7759/cureus.93479. PMID: 41170234; PMCID: PMC12570280.
12. Kowalczyk A, Loson-Kawalec M, Tabor A, et al. Comparison of GPT-5 Responses With the Official Results of the Polish Specialized Psychiatric Examination in Child and Adolescent Psychiatry. Cureus. 2025 Sep 22;17(9):e92982. doi: 10.7759/cureus.92982. PMID: 41141004; PMCID: PMC12548555.
13. Tangsrivimol JA, Darzidehkalani E, Virk HUH, et al. Benefits, limits, and risks of ChatGPT in medicine. Front Artif Intell. 2025 Jan 30;8:1518049. doi: 10.3389/frai.2025.1518049. PMID: 39949509; PMCID: PMC11821943.
14. Pamuk Ç, Uyanık AF, Kuyucu E, Uğurlar M. Can ChatGPT pass the Turkish Orthopedics and Traumatology Board Examination? Turkish orthopedic surgeons versus artificial intelligence. Ulus Travma Acil Cerrahi Derg. 2025 Mar;31(3):310-315. doi: 10.14744/tjtes.2025.07724. PMID: 40052322; PMCID: PMC11894241.
15. Kung JE, Marshall C, Gauthier C, Gonzalez TA, Jackson JB 3rd. Evaluating ChatGPT Performance on the Orthopaedic In-Training Examination. JB JS Open Access. 2023 Sep 8;8(3):e23.00056. doi: 10.2106/JBJS.OA.23.00056. PMID: 37693092; PMCID: PMC10484364.
16. Zong H, Wu R, Cha J, et al. Large Language Models in Worldwide Medical Exams: Platform Development and Comprehensive Analysis. J Med Internet Res. 2024 Dec 27;26:e66114. doi: 10.2196/66114. PMID: 39729356; PMCID: PMC11724220.

17. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. Nat Med. 2025 Mar;31(3):943-950. doi: 10.1038/s41591-024-03423-7. Epub 2025 Jan 8. PMID: 39779926; PMCID: PMC11922739.
18. Pokrywka J, Kaczmarek J, Gorzelańczyk E. GPT-4 passes most of the 297 written Polish Board Certification Examinations. http://arxiv.org/abs/2405.01589 doi:10.48550/arXiv.2405.01589
19. Bielówka M, Kufel J, Rojek M, et al. An investigative analysis – ChatGPT's capability to excel in the Polish speciality exam in pathology. Pol J Pathol. 2024;75(3):236-240. doi: 10.5114/pjp.2024.143091. PMID: 39451178.
20. Kipp M. From GPT-3.5 to GPT-4.o: A Leap in AI's Medical Exam Performance. Information. 2024; 15(9):543. https://doi.org/10.3390/info15090543
21. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT-3.5, ChatGPT-4, and Orthopaedic Resident Performance on Orthopaedic Assessment Examinations. J Am Acad Orthop Surg. 2023 Dec 1;31(23):1173–1179. doi: 10.5435/JAAOS-D-23-00396. Epub 2023 Sep 4. PMID: 37671415; PMCID: PMC10627532.
22. Takagi S, Koda M, Watari T. The Performance of ChatGPT-4V in Interpreting Images and Tables in the Japanese Medical Licensing Exam. JMIR Med Educ. 2024 May 23;10:e54283. doi: 10.2196/54283. PMID: 38787024; PMCID: PMC11148840.
23. Liu Y, Li Y, Wang Z, et al. A systematic evaluation of GPT-4V's multimodal capability for chest X-ray image analysis. Meta-Radiol. 2024 Dec 1;2(4):100099. doi:10.1016/j.metrad.2024.100099
24. Urooj B, Fayaz M, Ali S, et al. Large Language Models in Medical Image Analysis: A Systematic Survey and Future Directions. Bioengineering (Basel). 2025 Jul 29;12(8):818. doi: 10.3390/bioengineering12080818. PMID: 40868331; PMCID: PMC12383295.
25. Pal A, Sankarasubbu M. Gemini Goes to Med School: Exploring the Capabilities of Multimodal Large Language Models on Medical Challenge Problems & Hallucinations. https://aclanthology.org/2024.clinicalnlp-1.3
26. Kaneko M, Baldwin T. A Little Leak Will Sink a Great Ship: Survey of Transparency for Large Language Models from Start to Finish. http://arxiv.org/abs/2403.16139
27. Luo D, Liu M, Yu R, et al. Evaluating the performance of GPT-3.5, GPT-4, and GPT-4o in the Chinese National Medical Licensing Examination. Sci Rep. 2025 Apr 23;15(1):14119. doi: 10.1038/s41598-025-98949-2. PMID: 40269046; PMCID: PMC12018924.
28. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, Aziz S, Damseh R, Alabed Alrazak S, Sheikh J. Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions. JMIR Med Educ. 2023 Jun 1;9:e48291. doi: 10.2196/48291. PMID: 37261894; PMCID: PMC10273039.
29. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. EBioMedicine. 2023 Apr;90:104512. doi: 10.1016/j.ebiom.2023.104512. Epub 2023 Mar 15. PMID: 36924620; PMCID: PMC10025985.

Acknowledgements
We wish to express our gratitude to Piotr Koschel for his valuable support and assistance during the preparation of this article.

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) received no financial support for the research, authorship, and/or publication of this article.

Declaration of AI and AI-assisted technologies
No part of the manuscript was written by artificial intelligence.
AI was used in part of the research.