Polish-English medical knowledge transfer: A new benchmark and results

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have demonstrated significant potential in handling specialized tasks, including medical problem-solving. However, most studies predominantly focus on English-language contexts. This study introduces a novel benchmark dataset based on Polish medical licensing and specialization exams (LEK, LDEK, PES) taken by medical doctor candidates and practicing doctors pursuing specialization. The dataset was web-scraped from publicly available resources provided by the Medical Examination Center and the Chief Medical Chamber. It comprises over 24,000 exam questions, including a subset of parallel Polish-English corpora, where the English portion was professionally translated by the examination center for foreign candidates. By creating a structured benchmark from these existing exam questions, we systematically evaluate state-of-the-art LLMs, including general-purpose, domain-specific, and Polish-specific models, and compare their performance against human medical students. Our analysis reveals that while models like GPT-4o achieve near-human performance, significant challenges persist in cross-lingual translation and domain-specific understanding. These findings underscore disparities in model performance across languages and medical specialties, highlighting the limitations and ethical considerations of deploying LLMs in clinical practice.

1 Introduction

The potential of Artificial Intelligence, especially Large Language Models (LLMs), is vast, but they come with considerable risks, particularly the issue of “hallucinations”, where LLMs produce incorrect or misleading responses. This is especially concerning in fields like medicine, where errors can have serious consequences. Therefore, rigorous evaluation of LLM performance is essential before their clinical integration (Minaee et al., 2024).

LLM performance varies significantly due to differences in training methods, datasets, and objectives, which affect their ability to perform specific tasks. The quality and diversity of training datasets are particularly important for specialized domains like medicine (Minaee et al., 2024). While models trained on comprehensive, domain-specific datasets are expected to outperform those trained on general-purpose data, this assumption has been challenged (Feng et al., 2024).

Language also significantly impacts LLM performance. Most widely studied models are trained on multilingual datasets, predominantly in English, leading to better performance with English-language inputs and challenges with non-English content (Minaee et al., 2024). Additionally, LLMs trained exclusively on non-English texts may lack important knowledge available only in English.

Modern medicine is evidence-based, and one might assume that the correct management of medical issues should be nearly universal. However, in practice, clinical practices are shaped by various factors, leading to significant variations in medical guidelines across countries. For instance, Zhou et al. (2024) analyzed 22 clinical practice guidelines from 15 countries, highlighting notable differences in recommendations for managing lower back pain.

LLMs trained primarily on English-language data are likely to align with disease prevalence and clinical guidelines typical of English-speaking countries. Consequently, their diagnostic and therapeutic recommendations may be biased towards practices common in these regions. When presented with the same clinical scenario in different languages, an LLM may produce varying responses, reflecting the diversity of healthcare practices across countries represented in the training data. Such discrepancies could be revealed by evaluating LLMs on non-English medical tests, like those conducted in Poland, where disease prevalence and medical guidelines may differ from those in English-speaking countries.

To evaluate LLM performance in the medical field, we propose a new benchmark based on publicly available exam questions from medical and dental licensing exams, as well as specialist-level exams conducted in Poland[1]. This dataset includes over 24,000 questions, primarily in Polish, with a subset of licensing exam questions also available in English, enabling comparative analysis.

The proposed benchmark allowed us to study the behavior of LLMs by answering the following research questions:

- How does the performance of LLMs on Polish medical examinations differ across various models and various exam types?
- How does the performance of evaluated LLMs compare to that of human doctors and medical students?
- To what extent do LLMs provide divergent responses to general medical examination questions in Polish versus English, considering high-quality human experts' translations?
- What are the differences in the performance of LLMs on general versus specialized Polish medical examinations?
- How effectively do LLMs handle examination questions across various medical specialties (e.g., cardiology, neurology)?

2 Related work

LLMs are poised to transform various aspects of medicine by supporting medical professionals and enhancing research and education (Abd-Alrazaq et al., 2023; Clusmann et al., 2023). They can assist with literature summarization, data extraction, manuscript drafting, patient-clinical trial matching, and the creation of educational content (Clusmann et al., 2023; Harrer, 2023; Yang et al., 2023). By facilitating the conversion of unstructured to structured data, they streamline communication, translating complex medical information and summarizing patient records to simplify documentation (Clusmann et al., 2023). Applications in medical report generation and preauthorization letters can notably reduce the administrative burden on clinicians, allowing more focus on patient care (Harrer, 2023). This enhanced efficiency benefits healthcare systems not only economically but also supports personalized, patient-centered care through improved clinician-patient interactions (Clusmann et al., 2023; Nazi and Peng, 2024). Additionally, LLMs may aid diagnostics and management planning by analyzing large volumes of medical data and monitoring patient parameters (Nazi and Peng, 2024).

The integration of LLMs in healthcare necessitates rigorous evaluation to ensure reliability and safety, given the complexity and high stakes of medical decisions. Robust evaluation is crucial for assessing the performance of LLMs, identifying weaknesses, and mitigating biases to maintain patient safety and promote equity (Karabacak and Margetis, 2023; Li et al., 2023). Further, evaluation should ensure that LLMs genuinely enhance clinical care and effectively support healthcare professionals (Karabacak and Margetis, 2023).

Evaluating LLMs for medical use requires a tailored approach that goes beyond standard metrics, incorporating the specific demands of healthcare. Evaluation efforts should consider both technical performance, such as accuracy, reasoning abilities, and factual reliability using benchmark datasets (e.g., medical licensing exams), and real-world utility, including clinical impact studies and workflow integration (Karabacak and Margetis, 2023; Chang et al., 2024; Nazi and Peng, 2024).

Medical datasets are created using various approaches. The MedQA dataset comprises United States Medical Licensing Examination (USMLE)-style questions (Jin et al., 2020), while JAMA Clinical Challenge is based on the JAMA Network Clinical Challenge archive (Chen et al., 2024). Medbullets uses simulated clinical questions sourced from social media posts (Chen et al., 2024), and PubMedQA utilizes questions and contexts derived from PubMed articles (Jin et al., 2019).

Most of the current datasets focus on English, which reflects both the dominance of English in medical research and the initial English-centric development of LLMs. However, there is growing recognition of the need for multilingual and non-English datasets to ensure the broader applicability of medical LLMs. MedQA is notable for its multilingual approach, incorporating questions from medical board exams in English, Simplified Chinese, and Traditional Chinese (Jin et al., 2020). Additionally, there are datasets built around medical examinations in specific languages, including Swedish MedQA-SWE (Hertzberg and Lokrantz, 2024), Chinese CMExam (Liu et al.,

[1]The dataset is available at (anonymized)

2024), Japanese IGAKU QA (Kasai et al., 2023), and Polish.

For Polish, Lekarski Egzamin Końcowy (LEK, Eng. Medical Final Examination) was used as a benchmark (Rosoł et al., 2023; Bean et al., 2024; Suwała et al., 2023). LEK is available in both Polish and English, allowing researchers to evaluate the influence of language on LLM performance. To date, analyses have primarily focused on GPT models, though several other LLMs, including LLaMa and Med42, have also been evaluated (Bean et al., 2024).

Regarding the Państwowy Egzamin Specjalizacyjny (PES, Eng. State Specialisation Examination), a few studies have assessed GPT's performance in specialized field exams (Suwała et al., 2023; Kufel et al., 2023; Wojcik et al., 2023). Pokrywka et al. (2024) provided a comprehensive evaluation of GPT-3.5 and GPT-4 on the PES, utilizing 297 exams across 57 specialties.

Jin et al. (2024) proposed a benchmark for the cross-lingual evaluation of LLMs. However, the questions they used were translated by a machine translation system, while the questions in our benchmark were translated by human medical experts. Furthermore, we evaluated new models that demonstrate much better performance (Kipp, 2024).

3 Polish medical exams dataset overview

The LEK (Lekarski Egzamin Końcowy, Medical Final Examination) is a standardized exam for medical graduates and final-year students in Poland. Passing this exam, along with completing a postgraduate internship, is mandatory to obtain a medical license. Starting from 2022, 70% of the questions come from a publicly available database, which includes 2,870 questions for LEK. The exam is conducted twice a year and lasts four hours, consisting of 200 multiple-choice questions. Candidates are allowed to retake the exam multiple times, even after passing, to improve their scores.

The LDEK (Lekarsko-Dentystyczny Egzamin Końcowy, Dental Final Examination) is the equivalent exam for dentistry graduates and final-year students, following the same format and requirements as the LEK.

The PES (Państwowy Egzamin Specjalizacyjny, National Specialization Examination) is a mandatory exam for physicians and dentists who have completed specialization training, including required internships and courses. It consists of a written test and an oral examination. The written test, held twice a year for each specialty, typically includes 120 multiple-choice questions, with one correct answer per question, and a passing score of 60%. Candidates achieving at least 70% on the written part are exempt from the oral examination, a rule introduced in late 2022. PES is considered the most challenging exam in the professional career of a medical doctor in Poland, and unlike LEK and LDEK, its questions are not made public before the exam.

In Poland, five types of exams for physicians and dentists are conducted: LEK (Lekarski Egzamin Końcowy, Eng. Medical Final Examination), LDEK (Lekarsko-Dentystyczny Egzamin Końcowy, Eng. Dental Final Examination), LEW (Lekarski Egzamin Weryfikacyjny, Eng. Medical Verification Examination), LDEW (Lekarsko-Dentystyczny Egzamin Weryfikacyjny, Eng. Dental Verification Examination), and PES (Państwowy Egzamin Specjalizacyjny, Eng. National Specialization Examination, Board Certification Exam). LEW and LDEW are for graduates of medical or dental studies carried outside of the European Union. Passing these exams is necessary for them to legally practice in Poland.[2] However, these LEW and LDEW are taken by a relatively small number of candidates, and access to previous exam questions is limited. Therefore, they are.

The extensive descriptions of medical exams are included in Appendix A.

The dataset comprises medical exams from the Medical Examination Center (CEM) and the Supreme Medical Chamber (NIL), covering LEK, LDEK, and PES exams from 2008–2024. The exams were sourced as HTML quizzes and PDF files, with missing data from 2016–2020 (LEK/LDEK) and 2018–2022 (PES) partially filled using archives published on the NIL website. The exams were categorized by specialization, with questions and answers stored separately. Automated tools were used to scrap and process data, balancing parallelization with server constraints. Preprocessing ensured the dataset's suitability for text-only AI benchmarks by removing irrelevant files, questions containing images, and content misaligned with current medical knowledge. We refer to these as "invalidated questions" throughout the text. Detailed descriptions

[2] https://www.cem.edu.pl/lew_info.php https://www.cem.edu.pl/ldew_info.php

of data sources, acquisition methods, and quality considerations are provided in Appendix D.

Finally, we created five sub-datasets: LEK, LDEK, PES, LEK en (LEK translated into English), and LDEK en (LDEK translated into English). Not all of them were released in the same edition, particularly the Polish and English counterparts. Therefore, the results presented in Section 4 should not be used to directly compare LLM performance on Polish exams with their English translations. To address this, we focused on the overlapping years and reported these results in Section 5. For PES dataset, we collected a total of 180,712 questions. For our analysis, we selected only the latest exam from each specialty and based our analysis on these. Detailed dataset statistics are provided in Table 1. In summary, our analysis encompasses over 24,000 questions. For LLM inference, we utilized the Huggingface Transformers library (Wolf, 2019) and the OpenAI API.

4 Performance of LLMs on exams

We categorized the models under study into the following groups: medical LLMs (models fine-tuned on English medical data), general-purpose multilingual LLMs, Polish-specific models, and models with restricted APIs.

Medical Models: `BioMistral-7B` (Labrak et al., 2024), `Meditron-3` (8B and 70B versions) (OpenMeditron, 2024), `JSL-MedLlama-3-8B-v2.0` (johnsnowlabs, 2024).

General-Purpose Multilingual Models: `Qwen2.5 Instruct` (7B and 72B versions) (Team, 2024), `Llama-3.1 Instruct` (8B and 70B versions), `Llama-3.2-3B Instruct` (Dubey et al., 2024), `mistralai/Mistral-Small-Instruct-2409`, and `Mistral-Large-Instruct-2407` (Jiang et al., 2023).

Polish-Specific Model: `Bielik-11B-v2.2 Instruct` (Ociepa et al., 2024).

Restricted API Models: `GPT-4o-mini` and `GPT-4-o` (Achiam et al., 2023).

LLMs were evaluated by directly prompting them to answer exam questions. Each prompt included a brief introduction stating that the task was an exam for medical professionals consisting of single-choice questions. No additional examples or explanations were provided in the prompt; specifically, few-shot prompting was not employed. We believe this approach is appropriate for evaluating the models in a setting closely resembling the actual human exam environment.

We report the models' results as the percentage of correct answers in Table 2 and the number of exams passed in Table 3. Our findings are as follows: `GPT-4o` is the best performing model overall. Particularly in the PES category, it outperforms the second-best model. `GPT-4o` is capable of passing all evaluated exams except for six PES exams. However, `GPT-4o-mini` performs significantly worse than `GPT-4o` and is also inferior to general-purpose open models. Among the non-restricted API models, `Meta-Llama-3.1-70B-Instruct` is the best performer. Generally, general-purpose models outperform medical-specific models, possibly because the latter were fine-tuned on English medical data. The Polish-specific general-purpose model, `Bielik-11B-v2.2-Instruct`, performs worse than the top multilingual general-purpose models such as `Meta-Llama-3.1-70B-Instruct`, `Qwen2.5-72B-Instruct`, and `Mistral-Large-Instruct-2407`. However, for scenarios where deployment costs are more critical than performance, `Bielik-11B-v2.2-Instruct` may be preferable, as it still outperforms `Meta-Llama-3.1-8B-Instruct` of similar size in Polish-only exams. Our final recommendation is to use `GPT-4o` for Polish medical data tasks. If using a restricted API is not feasible (e.g., due to patient anonymity requirements), `Meta-Llama-3.1-70B-Instruct` is suggested as an alternative.

The performance of LLMs varies significantly based on specialization in PES exams, which was noted by (Pokrywka et al., 2024) before. We provide a detailed analysis across specialties in Appendix C, expanding upon the previous authors' findings with LLM other than the GPT family.

5 Cross-lingual knowledge transfer

To compare the performance of various LLMs on Polish and English versions of the same datasets, we restricted the LEK and LDEK datasets to identical subsets (LEK exams in English are exact translations of the Polish exams). The analysis results, similar to the previous one, are presented in Tables 4 and 5. As shown, all medical models, except for `OpenMeditron/Meditron3-70B`, perform better on the English versions of the datasets. This may be due to these models

Name	First	Last	Exams	Valid Questions	Invalidated Questions
LEK	2008A	2024S	22	4312	88
LDEK	2008A	2024S	22	4309	91
PES	2008A	2024S	72	8532	108
LEK (en)	2013A	2024S	14	2725	75
LDEK (en)	2013A	2024S	14	2726	74
total	2008A	2024S	156	24037	443

Table 1: Dataset statistics. S for Spring, A for Autumn.

Model Name	LEK	LDEK	PES	LEK (en)	LDEK (en)
BioMistral/BioMistral-7B	25.86	24.58	23.32	32.92	26.71
OpenMeditron/Meditron3-8B	45.57	38.32	36.99	60.51	43.21
OpenMeditron/Meditron3-70B	66.93	47.20	47.42	67.05	45.71
ProbeMedicalYonseiMAILab/medllama3-v20	40.61	34.05	31.79	52.40	38.15
aaditya/Llama3-OpenBioLLM-70B	55.15	39.78	40.06	66.09	45.27
johnsnowlabs/JSL-MedLlama-3-8B-v2.0	36.46	31.17	28.89	54.13	39.40
Qwen/Qwen2.5-7B-Instruct	51.41	42.93	41.32	67.78	48.42
Qwen/Qwen2.5-72B-Instruct	76.39	59.50	59.14	82.24	62.95
meta-llama/Meta-Llama-3.1-8B-Instruct	51.02	42.38	39.91	65.03	47.40
meta-llama/Meta-Llama-3.1-70B-Instruct	80.47	63.40	61.71	83.01	62.73
meta-llama/Meta-Llama-3.2-3B-Instruct	39.31	33.77	32.69	52.59	37.09
mistralai/Mistral-Small-Instruct-2409	51.37	40.98	38.35	64.04	43.03
mistralai/Mistral-Large-Instruct-2407	76.32	58.71	59.52	82.61	61.85
speakleash/Bielik-11B-v2.2-Instruct	61.87	45.51	42.02	57.25	42.85
gpt-4o-mini-2024-07-18	75.44	56.81	54.96	75.93	56.46
gpt-4o-2024-08-06	89.40	75.63	75.35	88.77	72.49

Table 2: The LLM results are represented as a percentage of correct answers of all datasets. The English versions of the LEK and LDEK exams are translated from the Polish versions; however, they represent only a subset of all the Polish exams.

Model Name	LEK	LDEK	PES	LEK (en)	LDEK (en)
BioMistral-BioMistral-7B	0/22	0/22	0/72	0/14	0/14
OpenMeditron-Meditron3-8B	0/22	0/22	0/72	14/14	0/14
OpenMeditron-Meditron3-70B	22/22	0/22	7/72	14/14	0/14
ProbeMedicalYonseiMAILab-medllama3-v20	0/22	0/22	0/72	3/14	0/14
aaditya-Llama3-OpenBioLLM-70B	16/22	0/22	0/72	14/14	0/14
johnsnowlabs-JSL-MedLlama-3-8B-v2.0	0/22	0/22	0/72	4/14	0/14
Qwen-Qwen2.5-7B-Instruct	3/22	0/22	2/72	14/14	0/14
Qwen-Qwen2.5-72B-Instruct	22/22	19/22	32/72	14/14	14/14
meta-llama-Llama-3.2-3B-Instruct	0/22	0/22	0/72	3/14	0/14
meta-llama-Meta-Llama-3.1-8B-Instruct	2/22	0/22	1/72	14/14	0/14
meta-llama-Meta-Llama-3.1-70B-Instruct	22/22	21/22	46/72	14/14	14/14
mistralai-Mistral-Small-Instruct-2409	2/22	0/22	0/72	14/14	0/14
mistralai-Mistral-Large-Instruct-2407	22/22	16/22	30/72	14/14	14/14
speakleash-Bielik-11B-v2.2-Instruct	22/22	1/22	1/72	9/14	0/14
gpt-4o-mini-2024-07-18	22/22	11/22	20/72	14/14	9/14
gpt-4o-2024-08-06	22/22	22/22	68/72	14/14	14/14

Table 3: The LLM results are represented as a percentage of correct answers of all datasets. The LEK and LDEK exams are considered passed with a minimum score of 56%, while the PES exam is considered passed with a minimum score of 60%.

being fine-tuned on English medical corpora. General-purpose multilingual models perform better on the English versions of the exams as well. This result is anticipated since these models are trained on corpora containing significantly more English than Polish. While these models are proficient in Polish, their performance on the tests remains lower in Polish than in English. The difference can be considerable; for example, `meta-llama-Meta-Llama-3.1-8B-Instruct` passed only one LEK exam in Polish but passed all 13 when translated into English. However, as model quality improves, the performance gap between languages narrows. For instance, with `meta-llama-Meta-Llama-3.1-8B-Instruct`, the accuracy difference between Polish LEK (51.25%) and English LEK (64.69%) is 13.44 percentage points (or a 26% relative change). In contrast, with meta-llama-Meta-Llama-3.1-70B-Instruct, the difference is only 1.66 percentage points (80.94% for Polish LEK vs. 82.60% for English LEK, or a 2% relative change).

For GPT-4o-mini, which generally performs well, the results in English are only slightly better than in Polish. Interestingly, for GPT-4o, performance is actually higher on the Polish version. The only Polish LLM, Bielik, performs better on Polish LEK and slightly better on Polish LDEK, likely due to its fine-tuning from the multilingual model Mistral-7B-v0.2 specifically for Polish. Overall, our observations suggest that language transfer is more effective as the model's general performance improves.

6 Comparison against human results

Meditron3-70B, Meta-Llama-3.1-70B-Instruct, Bielik-11B-v2.2-Instruct, and GPT-4o-2024-08-06 were selected as the top-performing models for the groups mentioned in Section 4, and compared against human students' results from the last four LEK and LDEK sessions (Spring 2024, Autumn 2023, Spring 2023, Autumn 2022), covering 977 LEK and 984 LDEK questions. While all selected models passed the chosen LEK exams, only Meta-Llama-3.1-70B-Instruct and gpt-4o-2024-08-06 scored within the range defined by an average number of points $\pm$ standard deviation achieved by humans. Assuming a normal distribution of exam results, it could be concluded that these models performed as a typical medical student. Notably, for the spring 2024 LEK exam, Meditron3-70B also achieved an average-level result, while gpt-4o-2024-08-06 exceeded the average student score. These findings are presented in Table 6(a). For the LDEK exams, all models performed noticeably worse. Assuming a normal distribution of exam results, only gpt-4o-2024-08-06 maintained a performance level comparable to that of an average medical student, consistent with its LEK exam results. In contrast, Meditron3-70B and Bielik-11B-v2.2-Instruct performed poorly, failing all exams, while Meta-Llama-3.1-70B-Instruct scored below the average but managed to pass each exam. These outcomes are summarized in Table 6(a) and Table 6(b).

The same models were used to compare their performance with human students on the PES exams. This analysis was conducted on a dataset created from the intersection of human results and LLM test results, encompassing 9,965 medical questions across 68 specializations from 12 exam sessions: Springs 2024, 2023, 2018, 2017, 2016, 2012, and Autumns 2023, 2020, 2019, 2016, 2015, 2008. The number of specializations is smaller than in the previous analysis due to inconsistencies in the specialization names between shared exams and published human results. The best-performing model was gpt-4o-2024-08-06, which achieved results in 60% of cases better than half of the student population or within the top 25% of scores. Notably, this model outperformed all students in a thoracic surgery exam. However, it is important to note that the student population for this particular exam was relatively small, consisting of only six participants. However, it is worth noting that even the best model achieved results worse than half of the student population in over 30% of specializations. For the *Audiology & phoniatrics* specialization, the model underperformed compared to all students. However, the student population for that particular case was relatively small, consisting of only nine participants. The second-best model, Meta-Llama-3.1-70B-Instruct, delivered significantly worse performance compared to the best model. Only 10% of its results across specializations were above the population median, while in over 30% of medical specializations, its performance was above the 25th percentile. The remaining models, Meditron3-70B and Bielik-11B-v2.2-Instruct, performed extremely poorly, with most of their results falling below the 25th percentile or even below the lowest scores of the entire student population. The students' results are presented in Appendix E, use whiskers to indicate the minimum and maximum student scores rather than the inter-quartile range. An aggregated exam results are provided in Table 6, X represents the distribution of human results, and the score of each model Y is categorized into the following ranges:

- $Y < min(X)$: Indicates model Y underperforms all students.
- $Y \in [min(X), p_{25})$: Model Y scores in the lowest 25% of students.
- $Y \in [p_{25}, p_{50})$: Model Y scores between the 25th and 50th percentiles, below the median but above the first quartile.

Model Name	LEK	LEK (en)	LDEK	LDEK (en)
BioMistral/BioMistral-7B	26.26	**32.74**	24.96	**26.78**
OpenMeditron/Meditron3-70B	**68.37**	66.75	**47.43**	45.97
OpenMeditron/Meditron3-8B	45.99	**60.34**	37.97	**43.35**
ProbeMedicalYonseiMAILab/medllama3-v20	40.93	**52.27**	35.09	**38.45**
aaditya/Llama3-OpenBioLLM-70B	61.33	**65.92**	41.77	**45.89**
johnsnowlabs/JSL-MedLlama-3-8B-v2.0	35.98	**54.09**	31.33	**39.44**
Qwen/Qwen2.5-72B-Instruct	76.87	**81.93**	58.35	**63.33**
Qwen/Qwen2.5-7B-Instruct	51.92	**67.73**	43.71	**48.38**
meta/llama-Llama-3.2-3B-Instruct	39.22	**52.08**	32.16	**36.87**
meta/llama-Meta-Llama-3.1-70B-Instruct	80.94	**82.60**	61.75	**63.17**
meta/llama-Meta-Llama-3.1-8B-Instruct	51.25	**64.69**	41.06	**47.71**
mistralai/Mistral-Large-Instruct-2407	76.75	**82.40**	56.29	**62.14**
mistralai/Mistral-Small-Instruct-2409	51.72	**63.70**	40.90	**43.47**
speakleash/Bielik-11B-v2.2-Instruct	**62.36**	56.98	**43.20**	42.88
gpt-4o-mini-2024-07-18	75.88	**75.92**	54.94	**56.88**
gpt-4o-2024-08-06	**89.96**	88.69	**73.89**	72.51

Table 4: The comparison of LLMs on Polish and English datasets, using the same LEK and LDEK exams, is represented as a percentage of correct answers.

Model Name	LEK	LEK (en)	LDEK	LDEK (en)
BioMistral-BioMistral-7B	0/13	0/13	0/13	0/13
OpenMeditron-Meditron3-70B	13/13	13/13	0/13	0/13
OpenMeditron-Meditron3-8B	0/13	13/13	0/13	0/13
ProbeMedicalYonseiMAILab-medllama3-v20	0/13	**3/13**	0/13	0/13
aaditya-Llama3-OpenBioLLM-70B	13/13	13/13	0/13	0/13
johnsnowlabs-JSL-MedLlama-3-8B-v2.0	0/13	**4/13**	0/13	0/13
Qwen-Qwen2.5-72B-Instruct	13/13	13/13	11/13	13/13
Qwen-Qwen2.5-7B-Instruct	2/13	**13/13**	0/13	0/13
meta-llama-Llama-3.2-3B-Instruct	0/13	**2/13**	0/13	0/13
meta-llama-Meta-Llama-3.1-70B-Instruct	13/13	13/13	12/13	**13/13**
meta-llama-Meta-Llama-3.1-8B-Instruct	1/13	**13/13**	0/13	0/13
mistralai-Mistral-Large-Instruct-2407	13/13	13/13	8/13	**13/13**
mistralai-Mistral-Small-Instruct-2409	1/13	**13/13**	0/13	0/13
speakleash-Bielik-11B-v2.2-Instruct	**13/13**	8/13	0/13	0/13
gpt-4o-mini-2024-07-18	13/13	13/13	6/13	**9/13**
gpt-4o-2024-08-06	13/13	13/13	13/13	13/13

Table 5: The comparison of LLMs on Polish and English datasets using the same LEK and LDEK exams is represented as a passed exams.

- $Y \in [p_{50}, p_{75})$: Model Y scores between the median and the top 25%.
- $Y \in [p_{75}, max(X)])$: Model Y scores in the top 25% of students.
- $Y \geq max(X)$: Model Y matches or surpasses the top human score.

7 Conclusion

In this paper we proposed a new benchmark for analyzing the performance of large language models in answering questions pertaining to the domain of medical knowledge. In contrast to the majority of previous medical datasets that collect examination questions in English, our dataset is derived from data of Polish origin. We showed that general-purpose LLMs outperform medical-specific models and that using a general-purpose model fine-tuned specifically for the Polish language is justified only if the models of the similar size are considered. We also confronted the scores obtained by LLMs with the resuts achieved by students showing that although all the analyzed models passed the standardized exam for medical graduate (LEK), even the top-performing model delivered results that were surpassed by at least half of the students in more than 30% of specializations.

The parallel sub-corpus composed of examination questions in Polish aligned with their English counterparts is a distinguished feature of the presented benchmark which allowed us to investigate the cross-lingual transfer of medical knowledge in LLMs. Our study showed that the models performed significantly on the English questions and that with increasing performance of the model, the gap between languages narrows. This is an expected result, but one that is difficult to observe without an appropriate benchmark.

Model Name	Criteria	Number of cases	Percentage share
OpenMeditron/Meditron3-70B	$Y < min(X)$	16	23.53%
	$Y \in [min(X), p_{25})$	**48**	**70.59%**
	$Y \in [p_{25}, p_{50})$	2	2.94%
	$Y \in [p_{50}, p_{75})$	2	2.94%
	$Y \in [p_{75}, max(X))$	0	0%
	$Y \geq max(X)$	0	0%
meta-llama/Meta-Llama-3.1-70B-Instruct	$Y < min(X)$	5	7.35%
	$Y \in [min(X), p_{25})$	**33**	**48.53%**
	$Y \in [p_{25}, p_{50})$	23	33.82%
	$Y \in [p_{50}, p_{75})$	5	7.35%
	$Y \in [p_{75}, max(X))$	2	2.94%
	$Y \geq max(X)$	0	0%
speakleash/Bielik-11B-v2.2-Instruct	$Y < min(X)$	24	35.29%
	$Y \in [min(X), p_{25})$	**43**	**63.24%**
	$Y \in [p_{25}, p_{50})$	1	1.47%
	$Y \in [p_{50}, p_{75})$	0	0%
	$Y \in [p_{75}, max(X)])$	0	0%
	$Y \geq max(X)$	0	0%
gpt-4o-2024-08-06	$Y < min(X)$	1	1.47%
	$Y \in [min(X), p_{25})$	11	16.18%
	$Y \in [p_{25}, p_{50})$	13	19.12%
	$Y \in [p_{50}, p_{75})$	**21**	**30.88%**
	$Y \in [p_{75}, max(X))$	**21**	**30.88%**
	$Y \geq max(X)$	1	1.47%

Table 6: Aggregated exam results categorizing model Y performance relative to the student population X across various percentiles, from scores below all students ($Y < min(X)$) to scores compared to or exceeding the best human results ($Y \geq max(X)$).

(a) LEK

Model / Human	2024S	2023A	2023S	2022A
OpenMeditron/Meditron3-70B	153	133	130	125
meta-llama/Meta-Llama-3.1-70B-Instruct	170	162	153	161
speakleash/Bielik-11B-v2.2-Instruct	129	122	123	133
gpt-4o-2024-08-06	184	177	176	179
Average human result with standard deviation	163.47 ±19.79	163.36 ±18.38	161.11 ±18.66	165.64 ±16.95

(b) LDEK

Model / Human	2024S	2023A	2023S	2022A
OpenMeditron/Meditron3-70B	103	83	94	95
meta-llama/Meta-Llama-3.1-70B-Instruct	121	119	124	123
speakleash/Bielik-11B-v2.2-Instruct	100	74	83	85
gpt-4o-2024-08-06	139	136	144	136
Average human result with standard deviation	147.62 ±26.08	148.57 ±19.08	149.42 ±21.13	156.22 ±23.52

Table 7: Comparison of top-performing LLMs and average human results, including standard deviation, across selected LEK and LDEK exams. Red represents values below the passing threshold of 112 points, orange highlights scores below average minus one standard deviation, green indicates scores above average plus one standard deviation, and black represents scores within one standard deviation of the average.

Limitations

While LLMs have demonstrated impressive performance on Polish medical multiple-choice exams, this achievement represents only a narrow facet of medical expertise. Becoming a licensed physician in Poland requires extensive training, rigorous coursework, and hands-on experience with practical medical procedures—far beyond what written exams can assess. Clinical practice necessitates analyzing diverse information and solving complex problems with multiple possible solutions. Physicians must determine what data is needed, obtain it through patient interviews, physical examinations, diagnostic tests, and consultations—all heavily reliant on direct human interaction that AI models cannot replicate. Moreover, the exams are multiple-choice, and real-world work is not narrowed to a few possible options. Therefore, despite strong exam results, LLMs cannot currently substitute the comprehensive qualifications and essential human interactions integral to effective medical care. However, this work shows that LLMs may be useful tools for medical practitioners. (Ullah et al., 2024; Park et al., 2024; Clark and Bailey, 2024; Liu et al., 2023; Lee et al., 2023).

Due to regional access restrictions, we were unable to evaluate PaLM 2 (Anil et al., 2023) and certain Llama 3.2 models. Additionally, highly resource-intensive models such as `Meta-Llama-3.1-405B-Instruct` or some other restricted access LLMs, such as Gemini (Gemini et al., 2023) were not evaluated.

References

Alaa Abd-Alrazaq, Rawan AlSaad, Dari Alhuwail, Arfan Ahmed, Padraig Mark Healy, Syed Latifi, Sarah Aziz, Rafat Damseh, Sadam Alabed Alrazak, Javaid Sheikh, et al. 2023. Large language models in medical education: opportunities, challenges, and future directions. *JMIR Medical Education*, 9(1):e48291.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Andrew M. Bean, Karolina Korgul, Felix Krones, Robert McCraith, and Adam Mahdi. 2024. Do large language models have shared weaknesses in medical question answering?

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2024. Benchmarking large language models on answering and explaining challenging medical questions. *arXiv preprint arXiv:2402.18060*.

Michelle Clark and Sharon Bailey. 2024. Chatbots in health care: Connecting patients to information. *Canadian Journal of Health Technologies*, 4(1).

Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, et al. 2023. The future landscape of large language models in medicine. *Communications medicine*, 3(1):141.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hui Feng, Francesco Ronzano, Jude LaFleur, Matthew Garber, Rodrigo de Oliveira, Kathryn Rough, Katharine Roth, Jay Nanavati, Khaldoun Zine El Abidine, and Christina Mack. 2024. Evaluation of large language model performance on the biomedical language understanding and reasoning benchmark: Comparative study. *medRxiv*.

Gemini, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Stefan Harrer. 2023. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine*, 90.

Niclas Hertzberg and Anna Lokrantz. 2024. Medqa-swe-a clinical question & answer dataset for swedish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11178–11186.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2024. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 2627–2638, New York, NY, USA. Association for Computing Machinery.

johnsnowlabs. 2024. Jsl-medllama-3-8b-v2.0. `https://huggingface.co/johnsnowlabs/JSL-MedLlama-3-8B-v2.0`. Accessed: 2024-11-02.

Mert Karabacak and Konstantinos Margetis. 2023. Embracing large language models for medical applications: opportunities and challenges. *Cureus*, 15(5).

Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. 2023. Evaluating gpt-4 and chatgpt on japanese medical licensing examinations.

Markus Kipp. 2024. From gpt-3.5 to gpt-4.o: A leap in ai's medical exam performance. *Information*, 15(9).

Jakub Kufel, Iga Paszkiewicz, Michał Bielówka, Wiktoria Bartnikowska, Michał Janik, Magdalena Stencel, Łukasz Czogalik, Katarzyna Gruszczyńska, and Sylwia Mielcarska. 2023. Will chatgpt pass the polish specialty exam in radiology and diagnostic imaging? insights into strengths and limitations. *Polish Journal of Radiology*, 88:e430.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains.

Peter Lee, Sebastien Bubeck, and Joseph Petro. 2023. Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *New England Journal of Medicine*, 388(13):1233–1239.

Hanzhou Li, John T Moon, Saptarshi Purkayastha, Leo Anthony Celi, Hari Trivedi, and Judy W Gichoya. 2023. Ethics of large language models in medicine and medical research. *The Lancet Digital Health*, 5(6):e333–e335.

Jialin Liu, Changyu Wang, and Siru Liu. 2023. Utility of chatgpt in clinical practice. *J Med Internet Res*, 25:e48568.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, et al. 2024. Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. *Advances in Neural Information Processing Systems*, 36.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Zabir Al Nazi and Wei Peng. 2024. Large language models in healthcare and medical domain: A review. In *Informatics*, volume 11, page 57. MDPI.

Krzysztof Ociepa, Łukasz Flis, Krzysztof Wróbel, Adrian Gwoździej, and SpeakLeash Teamand Cyfronet Team. 2024. Introducing bielik-7b-v0.1: Polish language model. Accessed: 2024-11-02.

OpenMeditron. 2024. Meditron3-70b. `https://huggingface.co/OpenMeditron/Meditron3-70B`. Accessed: 2024-11-02.

Ye-Jean Park, Abhinav Pillai, Jiawen Deng, Eddie Guo, Mehul Gupta, Mike Paget, and Christopher Naugler. 2024. Assessing the research landscape and clinical utility of large language models: A scoping review. *BMC Medical Informatics and Decision Making*, 24(1):72.

Jakub Pokrywka, Jeremi Kaczmarek, and Edward Gorzelańczyk. 2024. Gpt-4 passes most of the 297 written polish board certification examinations.

Maciej Rosoł, Jakub S Gąsior, Jonasz Łaba, Kacper Korzeniewski, and Marcel Młyńczak. 2023. Evaluation of the performance of gpt-3.5 and gpt-4 on the polish medical final examination. *Scientific Reports*, 13(1):20512.

S Suwała, P Szulc, A Dudek, A Białczyk, K Koperska, and R Junik. 2023. Chatgpt fails the internal medicine state specialization exam in poland: artificial intelligence still has much to learn. *Pol Arch Intern Med*, 133(11):16608.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Ehsan Ullah, Anil Parwani, Mirza Mansoor Baig, and Rajendra Singh. 2024. Challenges and barriers of using large language models (llm) such as chatgpt for diagnostic medicine with a focus on digital pathology–a recent scoping review. *Diagnostic pathology*, 19(1):43.

Simona Wojcik, Anna Rulkiewicz, Piotr Pruszczyk, Wojciech Lisik, Marcin Poboży, Iwona Pilchowska, and Justyna Domienik-Karłowicz. 2023. Beyond human understanding: Benchmarking language models for polish cariology expertise.

T Wolf. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

R. Yang, T. F. Tan, W. Lu, A. J. Thirunavukarasu, D. S. W. Ting, and N. Liu. 2023. Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4):255–263.

T. Zhou, D. Salman, and A. H. McGregor. 2024. Recent clinical practice guidelines for the management of low back pain: a global comparison. *BMC Musculoskeletal Disorders*, 25(1):344.

A Polish medical exams detailed description

Medical studies in Poland last 6 years, while dentistry takes 5 years. Final-year students and graduates can take their respective final exams — LEK for medicine and LDEK for dentistry. Passing the final examination and completing a postgraduate internship are required to obtain a medical license.[3] Both LEK and LDEK are four-hour exams conducted twice a year. Each exam consists of 200 multiple-choice questions with five possible answers, of which only one is correct. The questions cover a wide range of medical or dental disciplines. The distribution of questions from various fields is presented in Tables 8 and 9. To pass, a candidate must correctly answer at least 56% of the questions. Physicians and dentists can retake these exams multiple times, even after passing, if they are dissatisfied with their score.[4] A controversial rule (https://pulsmedycyny.pl/kadry/lekarze/samorzad-lekarski-postuluje-pilna-zmiane-bazy-pytan-w-lek-i-ldek/) has been introduced in 2022, stipulating that 70% of the exam questions come from a publicly available database, which includes 2,870 questions for LEK and 3,198 for LDEK. After these changes, the average exam scores and the percentage of passing candidates increased significantly.[5]

The PES exam is available to physicians and dentists who have completed the required internships and courses as part of their specialization training. Passing PES is mandatory to obtain the title of a specialist in a medical field. The exam consists of two parts: a written test and an oral examination. It is typically held twice a year for each medical specialty. The duration of the written test varies depending on the specialty, but it generally consists of 120 multiple-choice questions with five possible answers, of which one is correct. A minimum of 60% correct answers are required to pass. Unlike LEK and LDEK, none of the PES questions are public before the exam. Candidates who score at least 70% on the written test are exempt from taking the oral part of the exam, a rule implemented at the end of 2022. The format of the oral (practical) exam varies by specialty[6]. PES is generally considered to be the most challenging knowledge verification in the whole career of a medical doctor in Poland.

[3] https://www.cem.edu.pl/lek_info.php https://www.cem.edu.pl/ldek_info.php

[4] https://www.cem.edu.pl/lek_info.php https://www.cem.edu.pl/ldek_info.php

[5] https://www.cem.edu.pl/lep_s_h.php https://www.cem.edu.pl/ldep_s_h.php

Discipline	Questions
Internal medicine*	39
Pediatry*	29
Surgery*	27
Obstetrics and gynecology*	26
Psychiatry	14
Family medicine*	20
Emergency medicine and intensive care	20
Bioethics and medical law	10
Medical certification	7
Public health	8

Table 8: Distribution of test questions in LEK. The disciplines marked with an asterisk contribute to a minimum of 30 oncology-related questions. Internal medicine includes cardiovascular diseases. Pediatry includes neonatology. Surgery includes trauma surgery.

Discipline	Questions
Conservative dentistry*	46
Pediatric dentistry*	29
Oral surgery*	25
Prosthetic dentistry	25
Periodontology*	20
Orthodontics*	20
Emergency medicine	10
Bioethics and medical law	10
Medical certification	7
Public health	8

Table 9: Distribution of test questions in LDEK. The disciplines marked with an asterisk contribute to a minimum of 25 oncology-related questions.

B Example exam questions

B.1 LEK

Exam: 2022 Spring
Question id: 77

```
Przepuklina u starszego mężczyzny
z chorobą obturacyjną płuc uwypuklająca
się na zewnątrz jamy brzusznej przez
powięź poprzeczną stanowiącą tylną
ścianę kanału pachwinowego w miejscu
```

[6] https://www.cem.edu.pl/spec.php

ograniczonym od góry przez ścięgno
łączące, od dołu przez więzadło
pachwinowe, a bocznie przez naczynia
nabrzuszne dolne - jest rozpoznawana
jako:
A. przepuklina pachwinowa skośna.
B. przepuklina mosznowa.
C. przepuklina pachwinowa prosta.
D. przepuklina udowa.
E. przepuklina Spigela.

Correct answer: C.

B.2 LEK (en)

Exam: 2022 Spring
Question id: 77

An elderly male patient with obturative
lung disease was diagnosed with hernia.
It was protruding from the abdominal
cavity through the transverse fascia
which forms the posterior wall of the
inguinal canal, at the site bordering
the conjoint tendon at the top, the
inguinal ligament at the bottom, and
laterally, through inferior epigastric
vessels. The hernia in such location is
known as:
A. oblique inguinal hernia.
B. scrotal hernia.
C. direct inguinal hernia.
D. femoral hernia.
E. spigelian hernia.

Correct answer: C.

B.3 LDEK

Exam: 2022 Spring
Question id: 77

Jednostronny wyciek z nosa posokowatej
treści z domieszką krwi, rozchwianie
zębów górnych, łzawienie, wytrzeszcz
gałki ocznej, a niekiedy bóle i
mrowienie policzka mogą być wczesnym
objawem:
A. pseudotorbieli zatoki szczękowej.
B. raka zatoki szczękowej.
C. raka policzka.
D. przewlekłego zapalenia zatoki szczękowej.
E. ostrego zapalenia zatoki szczękowej.

Correct answer: B.

B.4 LDEK (en)

Exam: 2022 Spring
Question id: 77

Unilateral ichorous discharge from the
nose with a blend of blood, gomphiasis
of the upper teeth, lacrimation,
exopathalmos, and sometimes pain and
tingling sensation in the cheek, might
be an early symptom of:
A. pseudocyst of the maxillary sinus.
B. cancer of the maxillary sinus.
C. buccal cancer.
D. chronic maxillary sinusitis.
E. acute maxillary sinusitis.

Correct answer: B.

B.5 PES

Exam: 2019 Autumn
Question id: 68
Specialty: Family medicine

Kliniczne cechy sepsy u dzieci to:
1) gorączka;
2) leukocytoza;
3) leukopenia;
4) tachykardia bez innej przyczyny;
5) tachypnoë bez innej przyczyny.
Prawidłowa odpowiedź to:

Correct answer: E.

C Specialty performance on PES

Among the 72 unique PES specialties, certain areas of medicine consistently challenge the majority of tested models, while others frequently rank among the highest-scoring categories based on model accuracy. By identifying the top five highest and lowest-scored categories, we gain insights into specific domains where models excel or struggle, highlighting their potential limitations in these fields.

The general field of medicine where LLMs struggle the most is dentistry, specifically in orthodontics, which appeared ten times in the top five lowest scores across 17 models, followed by conservative dentistry with endodontics and pediatric dentistry. These results suggest that certain nuances in dental specialties may not yet be fully captured by modern LLMs, leading to difficulties in understanding this broad field.

The most frequently occurring specialty among the highest-scoring categories was laboratory diagnostics, which appeared twelve times. This observation may indicate that diagnostics tasks align well with the pattern recognition and data interpretation capabilities of LLMs. Additionally, other specialties with high scores, such as public health and pulmonary diseases reflect the vast quantity and accessibility of data in those fields. The COVID-19 pandemic could have largely increased the resource pool regarding pulmonary and respiratory conditions.

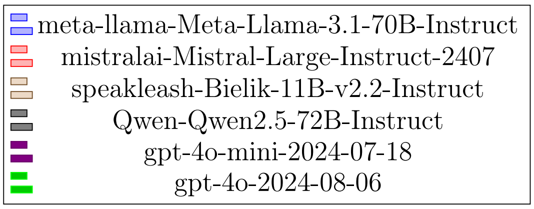


Figure 1: Models performance on different specialties on PES exams (part 1/2). Dotted lines indicate the passing threshold for the exam (60%) and exemption from the oral part (75%).

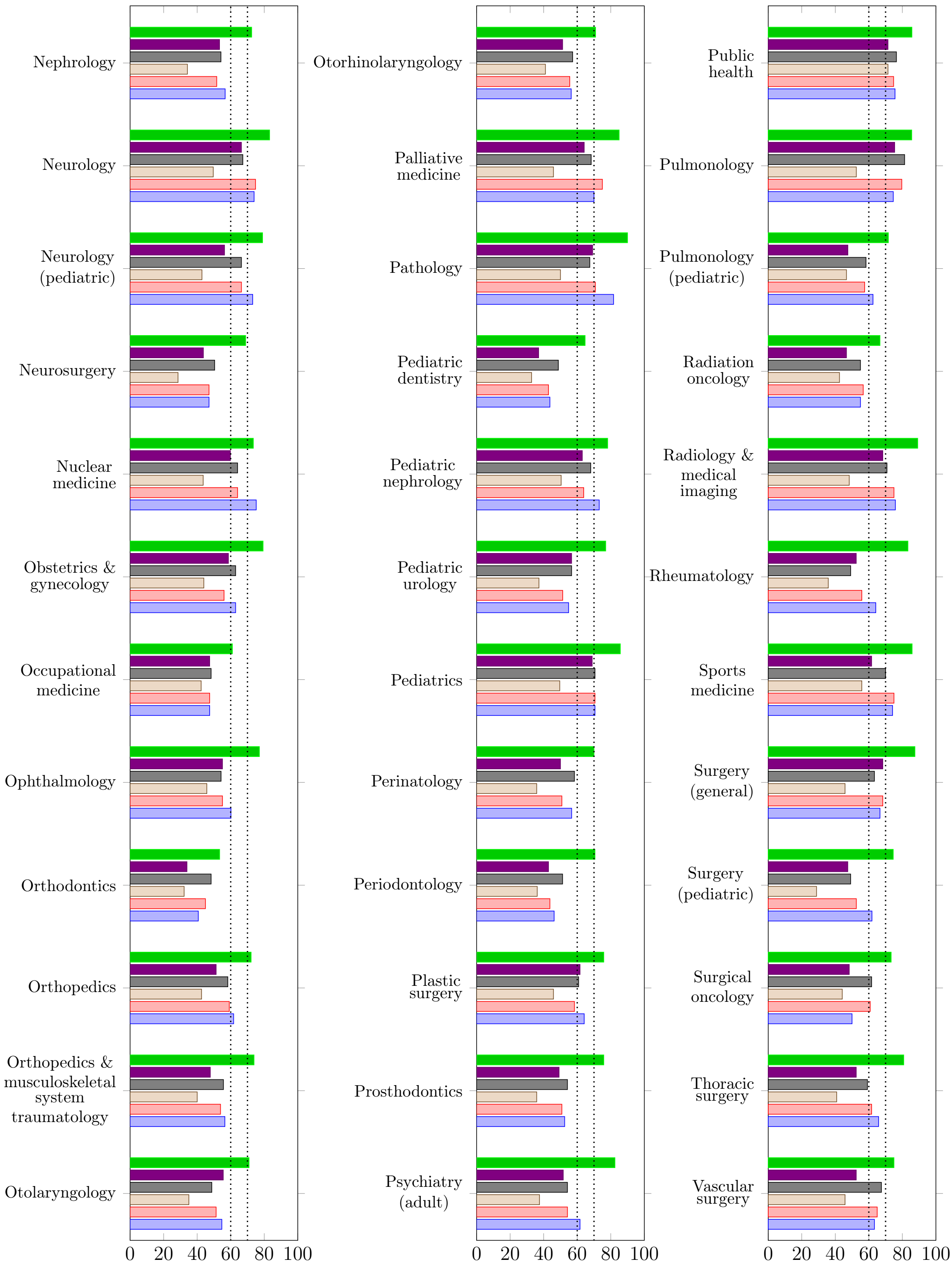


Figure 2: Models performance on different specialties on PES exams (part 2/2). Dotted lines indicate the passing threshold for the exam (60%) and exemption from the oral part (75%).

D Data preparation

D.1 Data sources

Medical exams in Poland are conducted biannually, in spring and autumn. Past exam content and corresponding answers are available on the Medical Examination Center (Centrum Egzaminów Medycznych, CEM) website, either as quizzes or PDF files. The site archives the following exams in the Polish language:

- LEK exams from autumn 2008 to autumn 2012 are provided as PDF files,
- LEK exams from spring 2013 to autumn 2015, and from spring 2021 to autumn 2024 are available as quizzes,
- LDEK exams from autumn 2008 to autumn 2012 are available as PDF files,
- LDEK exams from spring 2013 to autumn 2015, and from spring 2021 to autumn 2024 are provided as quizzes,
- PES exams from spring 2003 to autumn 2017, and from spring 2023 to spring 2024 are available as quizzes.

LEK and LDEK exams published as quizzes are also available in English. The missing LEK and LDEK exams from spring 2016 to autumn 2020 have not been found. The missing PES exams from spring 2018 to autumn 2022 have been published as PDF files on the Supreme Medical Chamber (Naczelna Izba Lekarska, NIL) website.

Figure 3: Quiz interface on the Medical Examination Center website.

The Medical Examination Center also provides detailed information about human answers for the PES exams. The initial view displays a list of examinees, represented by code numbers, along with their total achieved points and final grades. For all exams conducted since autumn 2006, detailed answers for each examinee are available by clicking on the examinee's code number. This detailed view includes the question number, the answer provided, and the correct answer.

D.2 Data acquisition and processing

The missing PES exams were published on the Supreme Medical Chamber platform across two distinct pages, with separate archives for the periods 2018–2020 and 2021–2022. Each medical specialization's exams were compressed into a zip file and provided as individual download links. To streamline the downloading process, a JavaScript script was executed via Chrome's Developer Tools, iterating through the links and simulating clicks for automatic downloads. The exams were then categorized by specialization, with each folder containing two types of PDF files: questions and the corresponding correct answers.

Custom Python scraping scripts were developed to automate the downloading of quizzes from the Medical Examination Center platform. Separate scripts were created for LEK/LDEK exams, PES exams, and exam statistics. Due to the server's slow response time, the entire process took several days, even with parallelized data download. When too many concurrent threads were used, the server became overwhelmed, resulting in timeouts.

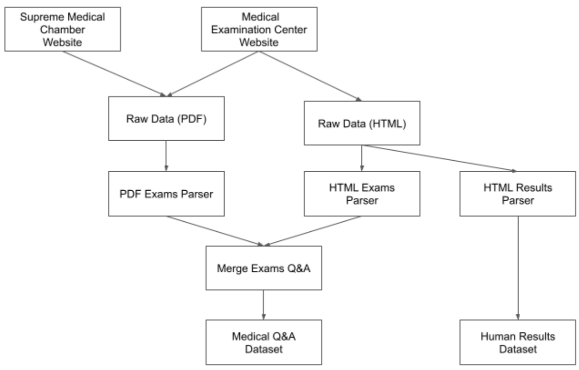


Figure 4: Data acquisition and processing workflow

D.3 Data quality

Data is stored in two formats: PDF and HTML, both of which are inconsistent and present several challenges. Since the goal of creating this dataset is to establish a Polish medical benchmark for Large Language Models, questions containing images were excluded. Additionally, some questions were disqualified by their authors due to errors or inconsistencies with current medical knowledge.

D.3.1 HTML format

HTML format is relatively straightforward to process, as specific HTML tags can be used to extract information such as questions and correct answers. However, some questions contain images that are essential for context, which poses a challenge for AI models designed to process text. Since the final dataset is intended for text-based AI models, questions containing images were excluded using specific tags. Additionally, the quiz interface allows anonymous users to leave comments on individual questions. These comments could potentially highlight areas where the content's alignment with current knowledge has been questioned. However, many of the comments appeared unprofessional and seemed not to be moderated by the platform administrators. As a result, the presence of comments was not considered a valid indicator for filtering questions, and all of them were kept in the final dataset.

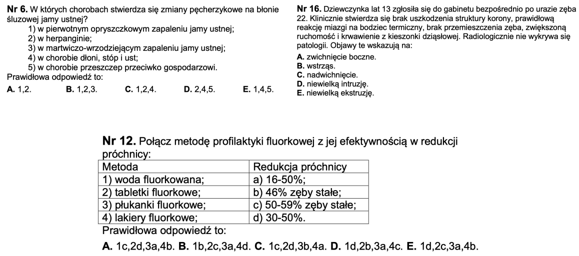
Moreover, the raw dataset contains empty questions. The platform uses two static drop-down lists to browse questions based on exam date and medical specialization, even when no corresponding exam or question is available in the database. According to the platform's messages, missing data occurs either due to the absence of questions in the database or because exams were not conducted during a specific time. This design leads to a collection of HTML files with no meaningful content. Since the user interface does not manage these cases, it was necessary to filter out and remove such files from the dataset after downloading.

Figure 5: Example of missing data caused by an absent question.

Figure 6: Example of missing data due to an exam not being conducted.

D.3.2 PDF files

Processing PDF files is more challenging compared to HTML due to the need to handle content sequentially, line by line, while applying multiple conditions to accurately extract medical exam questions. Additionally, the structure of questions is inconsistent across points, pages, and files. The question content or answer options may be presented in various formats, such as horizontal lists, vertical lists, two separate lists of options, or a table where points must be matched across columns. This inconsistency complicates the extraction process and poses difficulties for data processing.



Nr 6. W których chorobach stwierdza się zmiany pęcherzykowe na błonie śluzowej jamy ustnej?
1) w pierwotnym opryszczkowym zapaleniu jamy ustnej;
2) w herpanginie;
3) w martwiczo-wrzodziejącym zapaleniu jamy ustnej;
4) w chorobie dłoni, stóp i ust;
5) w chorobie przeszczep przeciwko gospodarzowi.
Prawidłowa odpowiedź to:
A. 1,2. **B.** 1,2,3. **C.** 1,2,4. **D.** 2,4,5. **E.** 1,4,5.

Nr 16. Dziewczynka lat 13 zgłosiła się do gabinetu bezpośrednio po urazie zęba 22. Klinicznie stwierdza się brak uszkodzenia struktury korony, prawidłową reakcję miazgi na bodziec termiczny, brak przemieszczenia zęba, zwiększoną ruchomość i krwawienie z kieszonki dziąsłowej. Radiologicznie nie wykrywa się patologii. Objawy te wskazują na:
A. zwichnięcie boczne.
B. wstrząs.
C. nadwichnięcie.
D. niewielką intruzję.
E. niewielką ekstruzję.

Nr 12. Połącz metodę profilaktyki fluorkowej z jej efektywnością w redukcji próchnicy:

Metoda	Redukcja próchnicy
1) woda fluorkowana;	a) 16-50%;
2) tabletki fluorkowe;	b) 46% zęby stałe;
3) płukanki fluorkowe;	c) 50-59% zęby stałe;
4) lakiery fluorkowe;	d) 30-50%.

Prawidłowa odpowiedź to:
A. 1c,2d,3a,4b. **B.** 1b,2c,3a,4d. **C.** 1c,2d,3b,4a. **D.** 1d,2b,3a,4c. **E.** 1d,2c,3a,4b.

Figure 7: Answer options presented horizontally, vertically, or in a table within the same PDF file.

The quality of the PDF files varies significantly. While some are digitally generated with perfect clarity, others resemble scanned printed documents of noticeably lower quality. Fortunately, this variation does not impact the data extraction process. However, certain PDF files lack text layers, making them significantly harder to process, as Optical Character Recognition (OCR) must be applied to extract the text. This challenge arose for 212 exams from 2021 and 2022 year. Due to the complexity, even with OCR, it was decided to omit these documents from the analysis.

Correct answers are stored in separate PDF files. To obtain comprehensive results, content must be extracted from both the question and answer files, and the corresponding points matched. Typically, the correct answer is indicated by a letter between A and E. However, in some cases, an 'X' appears in the answer file, indicating that the question is no longer aligned with current knowledge and has been annulled.

1059

E Comparison of human results and best-performing LLMs.

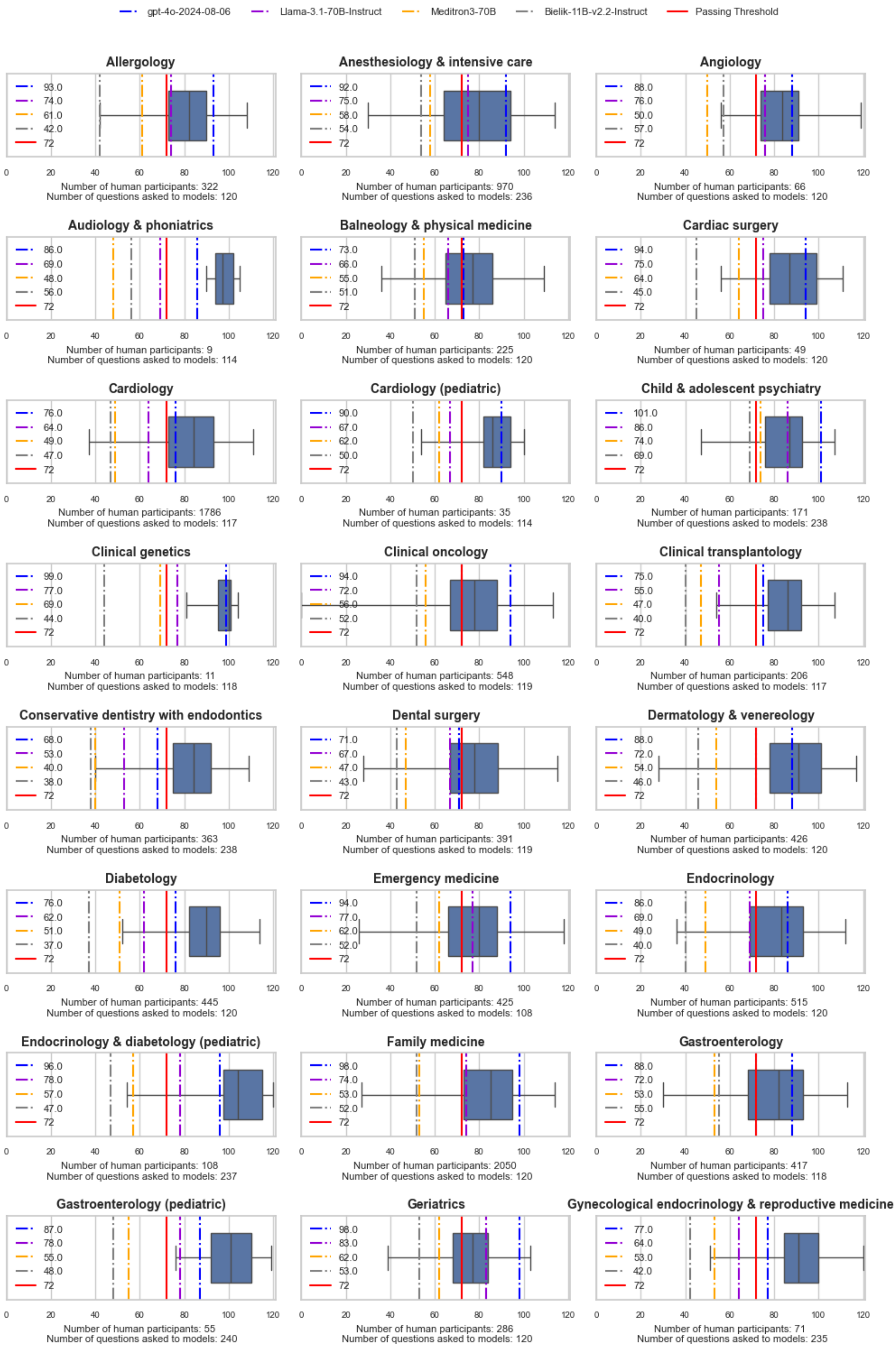


Figure 8: Students performance compared to top-performing LLMs on different specialties on PES exam (part 1/3).

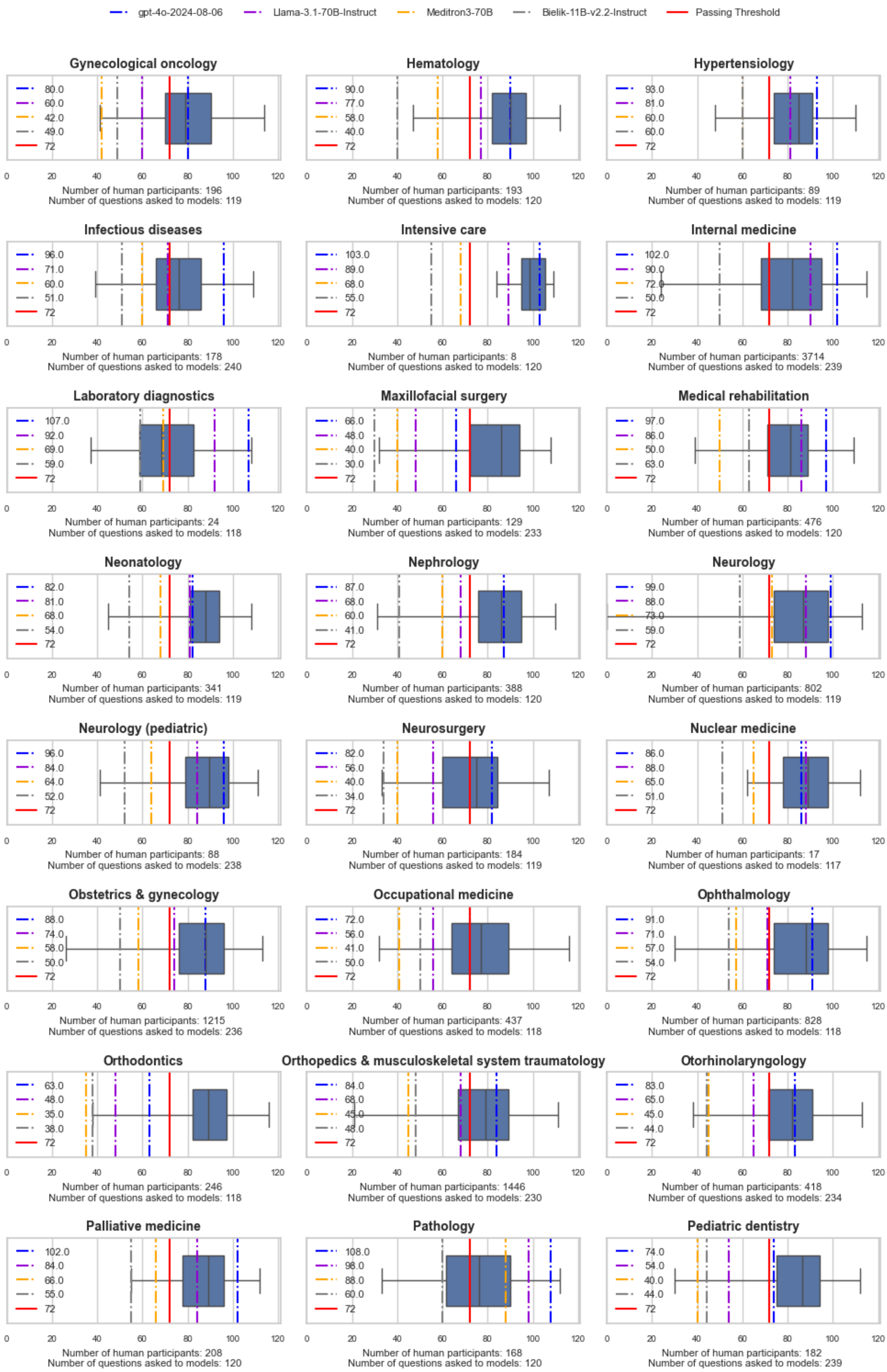


Figure 9: Students performance compared to top-performing LLMs on different specialties on PES exam (part 2/3).

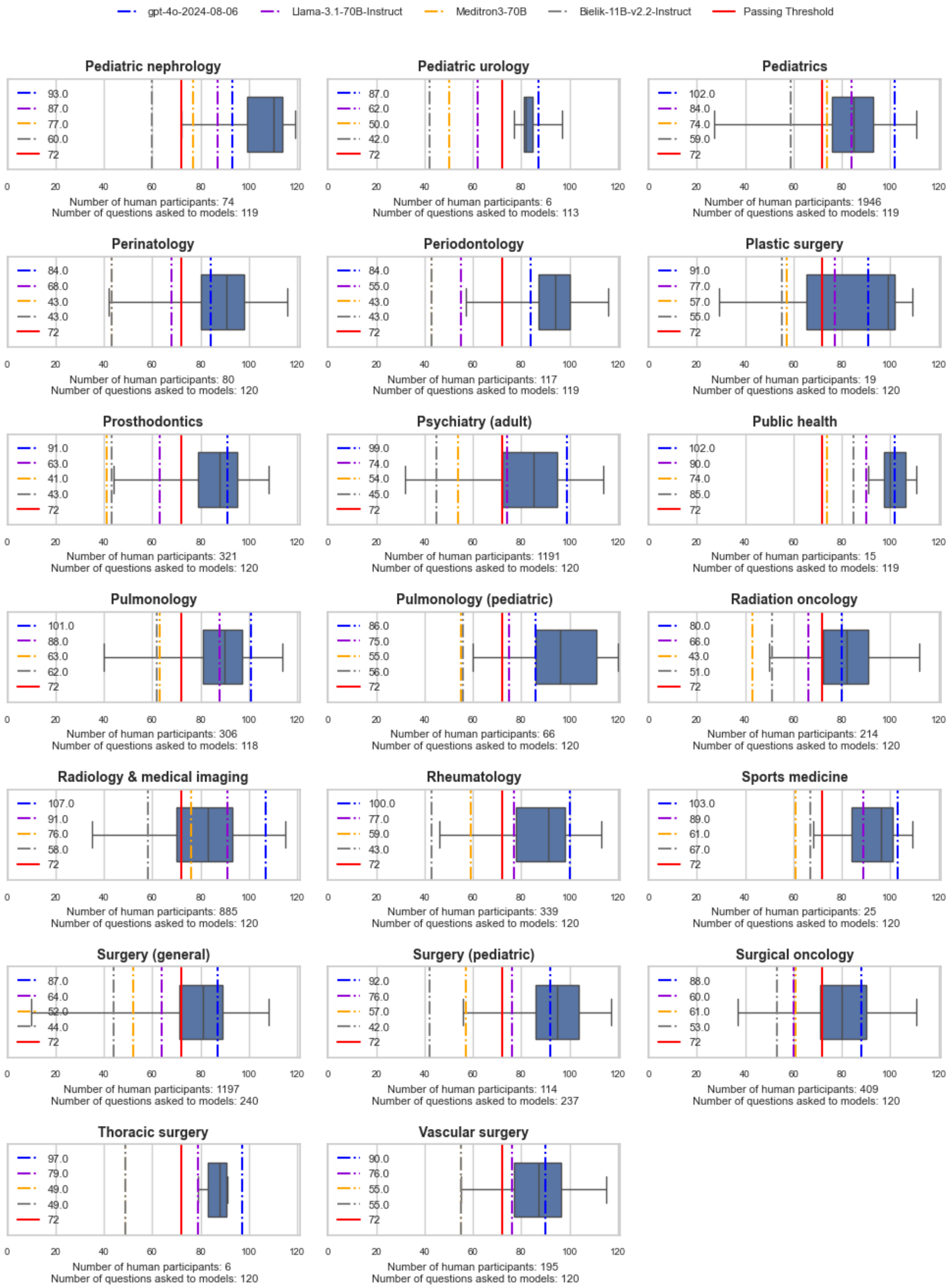


Figure 10: Students performance compared to top-performing LLMs on different specialties on PES exam (part 3/3).