Reassessing High-Performing LLMs on Polish Medical Exams: True Competence or Bias-Driven Performance?

Anonymous ACL submission

Abstract

Large language models (LLMs) have shown strong performance on medical question-answering benchmarks, yet such evaluations often rely on single-choice formats that may overestimate true clinical competence. Moreover, medical LLM performance varies substantially across languages, highlighting the importance of evaluations grounded in local medical practice and linguistic context. In this work, we reassess LLM performance on Polish medical examinations by extending and refining an existing benchmark based on Polish Medical Exams. We broaden the evaluation scope by incorporating questions from additional professional medical exams, and by modifying question structures to create a more challenging and informative evaluation setting. Through these extensions, we examine how evaluation design and question formulation influence model performance across diverse medical domains in Polish. Our results provide deeper insights into the robustness and limitations of LLMs in non-English medical contexts and highlight the need for more expanded evaluation frameworks in medical NLP. To facilitate further research, we make our benchmark publicly available.[1]

1 Introduction

Large language models (LLMs) are increasingly explored for medical applications (Yang et al., 2023), motivating the need for reliable and informative evaluation frameworks. In recent years, a wide range of benchmarks has been introduced to track progress in medical LLM capabilities, with most evaluations relying on question-answering tasks. However, most existing evaluation setups provide only a limited view of a model's true clinical reasoning abilities. Predominantly single-choice and multiple-choice questions can overestimate performance and fail to capture reasoning depth, uncertainty handling, or realistic clinical decision-making (Gu et al., 2025). Consequently, high benchmark scores do not necessarily translate into robust medical competence.

In addition, LLM performance in the medical domain varies appreciably across languages (Alonso et al., 2024; Jin et al., 2024). These findings suggest that medical questions require not only general biomedical knowledge but also local contextual understanding, including region-specific disease patterns, health-system constraints, and socio-cultural factors (Nimo et al., 2025). This highlights the need for evaluation benchmarks grounded in local medical practice and linguistic context.

Considering all limitations of the question-answer benchmarks, we decided to extend and refine the previously introduced medical benchmark for Polish medical examinations (Grzybowski et al., 2025). Our work addresses the following research questions:

- How do large language models perform across a broader and more diverse set of medical domains in the Polish language?
- How does modifying the structure of medical questions influence LLM performance?
- How to address limitations imposed by the evaluation, consisting of the exam-based single-choice questions?

To address the research questions outlined above, we introduce a set of extensions that meaningfully broaden the evaluation landscape of LLMs in the medical domain for the Polish language. Figure 1 presents the overview of our methodology. Our contributions include:

- Extension of the previously introduced Polish medical knowledge evaluation dataset[2] by incorporating over 12,600 new questions from

[1] https://anonymous.4open.science/r/MedicalExams-6341

[2] https://huggingface.co/spaces/amu-cai/Polish_Medical_Exams

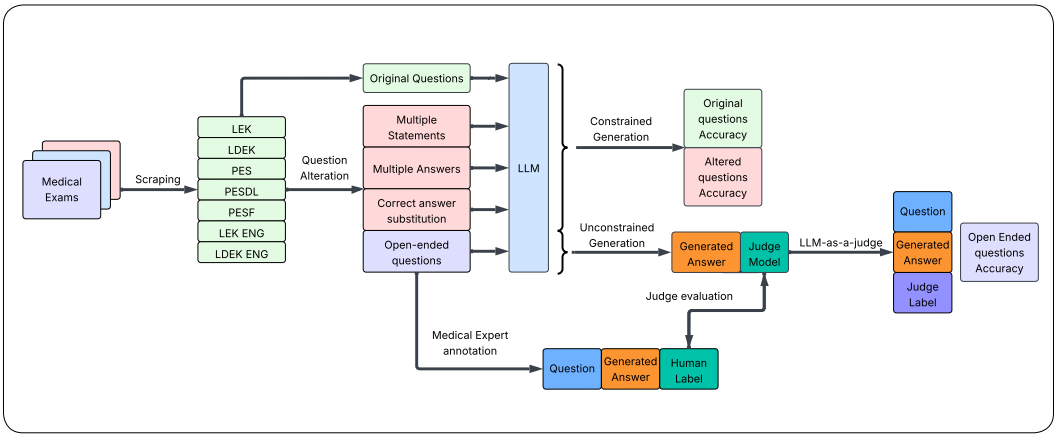

Figure 1: Overview of our methodology.

the Pharmaceutical Specialist Examination and Examination for Laboratory Diagnosticians Specialization.

- Alteration of question structure, resulting in a new evaluation method backed by a recomposed, specialized, and more challenging benchmark.

- Comprehensive evaluation of 20 LLMs, covering a broad range of model families, spanning general-purpose models, medical-domain specialized models, and Polish-language models, enabling a systematic comparison of their performance and robustness to question modifications.

2 Related work

A substantial portion of effort on medical benchmark construction has focused on question–answering (QA) datasets, resulting in widely adopted benchmarks such as MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), and more. While these resources have played an essential role in assessing progress, recent studies have identified several limitations of QA-based evaluation. In particular, many benchmarks fail to capture the complexity of realistic clinical scenarios, lack systematic validation against expert medical judgment, and are increasingly saturated, leaving limited headroom for meaningful performance differentiation. Moreover, rather than demonstrating robust medical reasoning, models often exploit innate pattern recognition abilities, leading to an overestimation of LLMs' capabilities in clinical reasoning and decision support. This behavior has been observed in settings where models are able to correctly quess answers even when key elements of the input question are masked or removed (Gu et al., 2025; Singh et al., 2025).

Another limitation of the QA-based evaluation approach concerns biases exhibited by LLMs. Selection bias refers to the tendency of a model to preferentially select specific answer option identifiers (e.g., A/B/C/D) independently of their semantic content. This behavior persists even when the option or question text is altered, provided that the option identifiers remain unchanged (Zheng et al.). Positional bias, in contrast, captures the model's sensitivity to the structural placement of answer options within a multiple-choice question. This bias arises from factors such as the order in which options are presented, their relative positions (e.g., first or last), and the overall ordering of the option list, irrespective of the option identifiers (Pezeshkpour and Hruschka, 2023; Zheng et al.). More generally, label bias refers to a model's tendency to systematically prefer certain answer labels over others, regardless of the content of the input task. In prompt-based classification settings, the model selects the answer with the highest probability, but this probability is often biased by the prompt rather than the task semantics. Such biases may be influenced by the choice of label verbalization, as well as the selection and ordering of in-context examples (Reif and Schwartz, 2024).

Reforming evaluation paradigms toward free-text generation or multi-turn conversational settings has therefore been proposed as a more informative and robust approach to assessing LLM competence in medical domains (Arora et al., 2025; Singh et al.,

2025). However, evaluating open-ended text generation remains substantially more challenging, as it typically relies on expert human annotation, which is costly, time-consuming, and difficult to scale. In response, LLM-as-a-Judge approaches have been explored. Nevertheless, research work indicates that expert involvement remains necessary when evaluating knowledge-intensive domains such as healthcare, particularly with respect to medical correctness and potential harm (Szymanski et al., 2025; Diekmann et al., 2025).

LLMs are increasingly being explored in Polish-language medical settings, ranging from medical benchmarks to systems designed to support emergency care (Chojnicki et al., 2025). Selected specializations, question sets, and exam subsets from the Medical Final Exam (LEK, *Lekarski Egzamin Końcowy*), the Medical–Dental Final Exam (LDEK, *Lekarsko-Dentystyczny Egzamin Końcowy*), and the Polish Board Certification Examination (PES, *Państwowy Egzamin Specjalizacyjny*) have been used in multiple studies to evaluate the LLMs' medical capacity in the Polish language (Rosoł et al., 2023; Wójcik et al., 2024; Suwala et al., 2023; Nicikowski et al., 2024; Pokrywka et al., 2024; Siebielec et al., 2024; Jassem et al., 2025).

The most comprehensive analysis to date of multiple LLMs on official Polish medical examinations is provided by Grzybowski et al. (2025). This study aggregates and publicly releases the largest collection of available exam materials. However, existing evaluations have not yet incorporated the Pharmaceutical Specialist Examination (PESF, *Państwowy Egzamin Specjalizacyjny Farmaceutów*) or the Examination for Laboratory Diagnosticians' Specialization (PESDL, *Państwowy Egzamin Specjalizacyjny Diagnostów Laboratoryjnych*).

3 Exams

The LEK Medical Final Exam in Poland is required for medical faculty graduates to obtain the right to practice and be admitted into a specialization. The exam consists of 200 closed questions with four distractors and one correct answer. Although a score of 56% is sufficient to pass, specialization placement depends on ranking, as the number of available training positions is limited. As a result, most candidates aim for a score of around 90%. Similarly, LDEK is an equivalent of LEK but for dentistry graduates. The PES Polish Board Certification Exam is a mandatory exam for physicians and dentists who have completed their specialization. It consists of two parts: a written exam and an oral. The written part consists of 120 closed questions with four distractors and one correct answer, and the participant passes if they obtain at least 60% of the possible points. Laboratory diagnosticians and pharmacists have their own equivalents of the PES, the Pharmaceutical Specialist Examination (PESF), and the Examination for Laboratory Diagnosticians Specialization (PESDL). Both of them have the same structure and characteristics as PES. Previous studies (Grzybowski et al., 2025) included from LEK, LDEK and PES together with English counterparts of LEK and LDEK, from the Centre of Medical Examination (Centrum Egzaminów Medycznych - CEM)[3] and Supreme Medical Chamber (Naczelna Izba Lekarska - NIL).[4] Recently, these institutions have released the question sets from the PESF and PESDL exams, which have not been included in any of the prior studies. This has opened an avenue for more diverse evaluation of medical knowledge

4 Dataset creation

In this section we describe the process behind the curation of additional exams and our modifications to the question format.

4.1 Additional datasets sourcing

As described in the previous Section, we included the PESF and PESDL datasets using scraping scripts, which removed unnecessary information, extracting only the contents of the questions and answer key. All of the exam questions and keys were available in the PDF format. Two of the PESF exam editions were in the form of PDF scans, resulting in multiple OCR errors and rendering them unsuitable for future processing. They were discarded because they constituted only a minority of the exam's editions. As a result, over 12,000 questions were added to the benchmark.

4.2 Question modification

The multiple-choice format, due to its inherent limitations, is not well-suited for evaluating the true depth of knowledge possessed by LLMs. Looking at the formulation of a multiple-choice QA, we can observe that given a question Q and a set of choices A including a correct answer $a_i \in A$ the

[3] https://cem.edu.pl/index.php
[4] https://nil.org.pl/

task in fact is not revolved around the generation of the correct answer, but around the separation between incorrect and correct ones. This is backed by the experiments performed by (Chandak et al., 2025) that highlight that models fine-tuned for QA given only the set of answers *A*, are significantly outperforming the random accuracy baseline. This suggests that the question content may be entirely redundant, and the model can discriminatively select the correct answer based solely on the context of the choices.

To limit this phenomenon and propose modifications to the original multiple-choice formulation, three main groups of question structures that enable alteration have been identified and are described below. The examples of modified questions in the English version are provided in Section E.

1) Hidden multiple-response questions In the Multiple Statements (**MS**) setting, questions originally formulated as multiple-choice with composite answer options (each option encoding a subset of statements) are reformulated to require the model to output the set of correct statements directly. An answer is considered correct only if the predicted set exactly matches the ground truth.

In the Multiple Answers (**MA**) setting, questions containing meta-options (e.g., "answers C and D are correct") are transformed into standard multiple-response questions by removing such options and marking the underlying answers as correct. Models must identify all of the answers to be correct.

2) Questions where the correct answer can be substituted with "None of the answers is correct." Correct answer substitution (**AS**) tests a model's ability to abstain when no provided option is correct. In this setting, the original correct answer is replaced with an option indicating that none of the remaining answers is correct, requiring the model to recognize the absence of a valid choice.

3) Questions that can be modified to be open-ended. Open-ended questions (**OE**) are derived from exam items that can be answered solely based on their textual context, without reliance on predefined options. This format reduces multiple-choice bias and better reflects realistic model usage. However, evaluation is more challenging, as multiple valid phrasings may exist and domain expertise is required for judgment. In this setting, we employ LLM-as-a-judge.

The structure of the exam questions is consistent, enabling the filtering of questions into modification groups using regular expressions. The regular expression analysis and filtering were done in Polish. Since the English question sets are parallel to Polish, the same filters have been applied to them, as the question structures were the same.

Exam	#Original	#Covered	Coverage
LEK	4,312	2,610	61%
LDEK	4,309	3,059	71%
PES	9,965	7,070	71%
LEK ENG	2,725	1,627	60%
LDEK ENG	2,726	1,878	69%
PESDL	10,908	7,627	70%
PESF	1,710	982	57%
All	36,655	24,853	68%

Table 1: Coverage statistics by source.

The obtained dataset spans over 35,000 questions covers multiple medical domains Table 1, and proposed alterations cover 68% of the original questions.The distributions of altered question types vary across exam sets Figure 2, because the questions suited for modifications are not distributed equally in each exam set. Additionally, open-ended alteration had priority over others, since it had the most rigorous filtering, so if the question was suited for both open-ended and *None of the answers is correct* alterations, the open-ended was chosen.

5 Methodology

5.1 Obtaining answers

Since LLMs are extremely vulnerable to prompt formulations, we have decided to inspect whether the prompt formulation strategy taken in (Grzybowski et al., 2025), does not influence the distribution of the produced answers on the Polish medical exams benchmark. To achieve this, analysis of the produced answers by various LLMs, including Polish Bielik (Ociepa et al., 2025b,a,d,c) and PLLuM (Kocoń et al., 2025) variants, but also English-centered models like Qwen (Team, 2025b) and GPT-5-mini (OpenAI, 2025a) has been conducted using the following instruction:

> **Your task is to provide answers to a medical test for doctors. From all the provided answers A, B, C, D, E select only one. If you are not sure, choose the most probable one. Answer in a manner: Correct answer is B.**

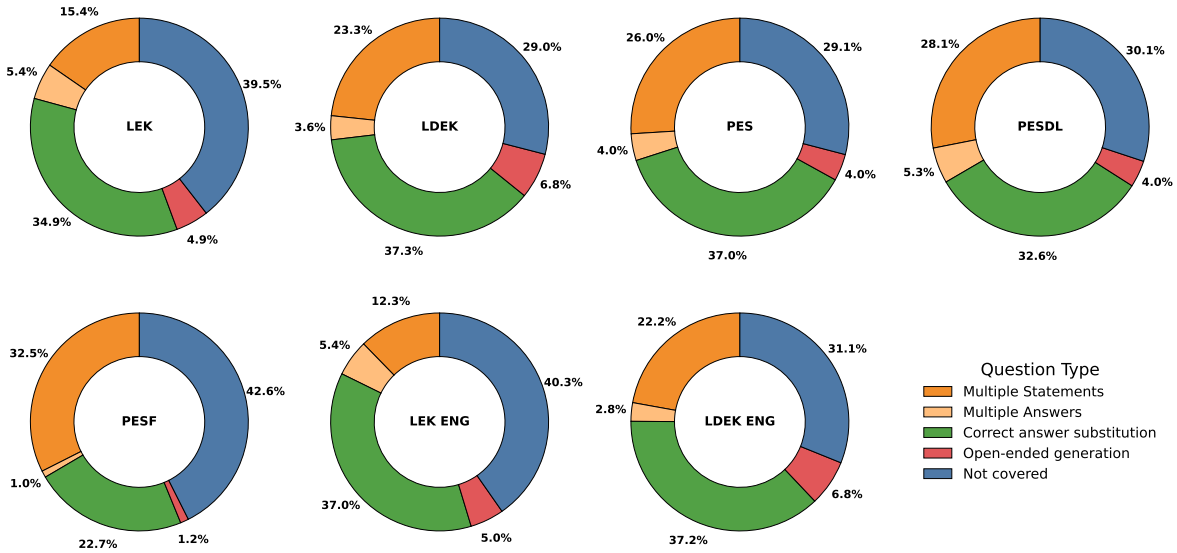

Figure 2: Question type composition across exam question sets.

The initial analysis suggests that the considered models are biased towards the answer that is included as a sample answer format in the prompt, and that even the order of listing the answers in the instruction might influence the model's prediction. These initial takeaways are based on the comparison between the ground-truth answer distribution (correct answers for the medical exams) and the distribution of answers generated by the aforementioned models. Since the formulation of the questions that comprise the benchmark already includes the possible answers in a classical order (A, B, C, D, E), providing the possible answers in the instruction is not necessary.

> **Your task is to provide answers to a medical test for doctors. From the provided answers, select the correct one and respond only with that answer. End your answer with a period.**

When applicable[5], constrained generation based on a set of allowed tokens was used paired with the instruction suited for the modified question types.

For example, in one of the multiple-choice scenarios, the model had to return the set of correct statements; the generation was constrained to return only the indices of the statements considered true, and a suitable instruction was modified to meet these requirements. All of the instructions for generating answers are provided in the Appendix C.

[5]Some of the tokenizers have unique way of encoding the numbers into multiple tokens, when constrained to generate only tokens that are used to form numbers from 1-20 (indices of correct statements) the model generated complex multi-digit numbers instead, which rendered a constrained generation unapplicable.

5.2 Answer evaluation

Answer-key During the dataset alteration, the answer key for a given modification was modified on the fly to resemble the modifications made. When considering multi-answer or multi-statement questions, the answer was deemed only correct when the provided output included an identical set of correct statement indices or answers.

LLM-as-a-judge For evaluating open-ended questions, the LLM-as-a-judge approach was adopted using GPT-oss-120B (OpenAI, 2025b). However, since the original answer to the modified question may not be the only correct one, and judgment based solely on the original answer key would be insufficient, the model requires extensive medical knowledge to make a correct assessment. Additionally, binary judgments can overstate the accuracy of LLMs as evaluators, since the two-class setup allows even chance-level predictions to score relatively high. Instruction for the LLM judgement is provided in the Appendix C.

Annotations To address the issues of inflated judgment scores and resolve whether out-of-domain LLM can be sufficient for the evaluation, a meta-evaluation dataset has been introduced using human annotation to measure the alignment be-

All Annotation Samples	1,200
Invalid Samples – Missing Labels	16
Invalid Samples – Invalid Question	167
Positive Labels	280
Negative Labels	647
Fleiss Kappa	0.839

Table 2: Annotation process summary

Samples judged	927
Positive Labels	305
Negative Labels	622
Cohen Kappa (Human, LLM)	0.598

Table 3: LLM-as-a-judge labels on the meta evaluation dataset with measured agreement with human annotation.

tween LLM-as-a-judge and domain experts - doctors. A sample from LEK and PES open-ended questions was used to generate answers from four models: PLLuM-12B-instruct (Kocoń et al., 2025), gpt-oss-20b (OpenAI, 2025b), Bielik-11B-v2.6-Instruct (Ociepa et al., 2025a,d) and Qwen/Qwen3-30B-A3B-Instruct-2507 (Team, 2025b). Each annotator, a medical professional, was given a set of 200 (question, generated-answer, original answer) triplets, together with annotation guidelines. Each set of samples included 15 control triplets to measure the agreement between the annotators. The annotation task was defined as a binary assessment of whether the generated answer is, in fact, a correct answer to the question, with the original answer provided as an additional context. The annotators could additionally decide if the question is poorly constructed, outdated, or unsuitable for a clear answer and discard it from the annotation pool.

The annotation process resulted in 927 valid annotation samples and 90 valid control samples. Over 150 triplets were invalid due to the question construction or being outdated. We have calculated the Fleiss Kappa to measure the agreement between the annotators on the control group Table 2. To facilitate our LLM-as-judge evaluation, we have compared the labels produced by Experts and LLM using the Cohen Kappa Table 3. We have concluded that the obtained kappa of 0.598 is sufficient to extrapolate the judge method to unsupervised examples. Both the annotation guidelines and the instructions for the judge were similarly constructed.

6 Experiment setup

The experiments consist of two parts. The first part focuses on the outcomes of the efforts to limit the bias and create a fair evaluation. In this part, the LLMs are evaluated on the unmodified exam sets, comparing two evaluation techniques: the technique reproduced from the Polish Medical Exams benchmark (Grzybowski et al., 2025) and the introduced one that changes the instruction to minimize bias and utilizes constrained generation. Additionally, including the PESF and PESDL exam questions as an extension to the benchmark. The second part of the experiments presents a comparison of the results obtained from the evaluation using the original and altered question structures. Each exam set is divided according to the alteration method and then evaluated on the same questions, both unmodified and altered.

6.1 Models

Models under study are categorized in the following way: medical-domain fine-tuned models, Polish LLMs, and general-purpose LLMs.

Medical models: BioMistral-7B and BioMistral-7B-DARE (Labrak et al., 2024) , Meditron3-70B (OpenMeditron), Llama3-OpenBioLLM-70B (Pal and Sankarasubbu, 2024), MedGemma (4B and 27B versions) (Sellergren et al., 2025).

General-purpose models: Qwen2.5-72B (Yang et al., 2024; Team, 2024), Qwen3-30B (Team, 2025b), Llama3.3-70B (AI@Meta, 2025), Llama3-8B (AI@Meta, 2024), Gemma3 (12B and 27B versions) (Team, 2025a).

Polish LLMs: PLLuM-12B-instruct, Llama-PLLuM-70B-instruct (Kocoń et al., 2025), Bielik-11B-v2.6-Instruct (Ociepa et al., 2025a,d) and Bielik-4.5B-v3.0-Instruct (Ociepa et al., 2025b,c)

6.2 Evaluation

We use accuracy expressed in % as a metric for evaluation. Additionally, we aggregate the accuracies from respective modification groups, taking an average of **MS MA** and **AS** and an accuracy on the unmodified equivalents to measure the average change of the performance of a given model on a given exam between the original and modified questions. The evaluation of the **OE** questions is also expressed as an accuracy measure where the correct answer is determined based on the judg-

Model	Exam Performance (%)													
	LEK		LDEK		PES		PESDL		PESF		LEK-ENG		LDEK-ENG	
	OM	PM	OM	PM	OM	PM	OM	PM	OM	PM	OM	PM	OM	PM
Bielik-4.5B	47.91	43.39	37.25	33.37	32.15	30.58	31.21	34.20	23.92	27.31	52.70	50.13	39.62	38.19
Bielik-11B	64.05	58.93	47.04	41.94	44.38	39.04	50.14	43.27	46.73	37.89	59.71	52.26	44.24	36.90
PLLuM-12B-NC-Inst	48.84	48.77	36.04	35.76	32.91	32.49	37.30	35.84	31.23	32.22	42.83	39.96	32.17	31.07
PLLuM-12B-NC-Chat	49.56	48.59	37.34	35.48	32.94	32.75	37.59	36.29	32.92	31.40	43.89	44.55	33.16	33.24
PLLuM-12B-Inst	46.24	36.46	35.16	27.52	32.65	25.14	35.20	27.83	30.58	22.92	40.88	28.84	30.45	24.17
PLLuM-12B-Chat	38.50	20.80	29.47	17.96	27.49	14.96	30.66	14.88	23.92	12.92	30.94	25.69	24.94	18.97
PLLuM-70B-Inst	70.64	45.99	51.84	32.93	50.43	31.12	57.67	35.78	47.84	27.72	71.27	61.10	50.48	41.82
LLaMA3-8B	44.53	45.73	36.88	39.13	35.93	37.41	40.82	42.58	31.99	34.09	60.81	61.28	42.74	43.21
Gemma-3-12B	64.29	53.83	48.99	38.25	45.90	36.16	53.08	42.30	46.20	37.66	68.73	61.80	47.36	41.64
Gemma-3-27B	72.19	68.02	54.44	49.18	52.23	48.00	59.75	56.81	50.47	50.64	74.39	70.13	52.49	49.16
Qwen3-30B-A3B	74.81	72.52	57.67	54.17	57.42	54.60	66.14	61.72	54.39	50.23	80.84	78.97	60.49	57.15
LLaMA3.3-70B	78.39	79.04	60.57	60.64	59.98	59.70	66.35	67.29	56.43	55.56	81.21	82.28	59.79	62.40
Qwen2.5-72B	77.95	75.83	61.87	58.53	61.04	57.69	69.69	66.93	57.89	54.56	82.75	77.83	64.64	56.53
MedGemma-4B	50.12	43.95	39.61	34.56	37.86	32.78	42.45	36.41	36.08	29.53	58.50	53.72	39.51	36.43
MediPhi-Instruct	33.28	26.95	32.12	26.43	29.87	24.61	34.27	26.62	26.32	20.64	60.77	55.38	42.37	37.89
BioMistral-7B	32.03	19.71	29.06	19.47	27.35	18.52	29.83	19.16	21.64	12.57	42.86	33.06	32.87	25.35
BioMistral-7B-DARE	31.86	24.05	28.92	23.04	26.33	21.47	30.05	22.78	20.00	17.08	44.77	35.16	33.79	27.48
MedGemma-27B	73.86	70.64	53.84	50.17	52.47	49.77	58.92	55.57	51.35	50.88	76.48	73.43	54.37	48.13
OpeOMioLLM-70B	67.67	59.53	50.17	42.17	50.62	40.84	57.43	47.56	48.83	35.03	73.65	74.83	53.37	53.08
Meditron3-70B	65.47	26.00	50.52	32.84	48.44	21.55	54.44	19.42	47.08	15.15	71.74	2.50	53.85	1.10

Table 4: Model evaluation results expressed as percentages. *OM* = Our evaluation Methodology, *PM* = Previous evaluation Methodology.

ment from the described LLM-as-a-judge approach using GPT-oss-120B (OpenAI, 2025b).

7 Results and discussion

Limiting the bias The results of the replicated evaluation methodology used in (Grzybowski et al., 2025) and the proposed one are presented in Table 4. The vast majority of the models outperform the previous methodology. This is consistent with the hypothesis that the bias inferred by the instruction and generation method negatively affects the benchmark results.

Question alteration effect The evaluation of models' performance on the altered question groups (**MS, MA, AS**) presented in the Table 5 showcases a significant drop in performance, suggesting that the high scores on the original questions might be attributed not to the competence of the models in medical domains but to the formulation of the evaluation. Detailed results for each alteration group are presented in the Appendix D.

Open-ended questions Answering open-ended questions is the most natural setting for evaluating LLMs, as it resembles both day-to-day and in-domain use. The results of the evaluation on the **OE** group, presented in the Table 6, show that the models are not competent enough to provide the correct answers in this setting. None of the models has surpassed the average accuracy of 45%, with the best model being Qwen3-30B. Moreover, there is no significant improvement between the specialised medical models and the general-purpose ones. We observe a consistent trend where larger models achieve higher accuracy, suggesting that memorisation capacity may play an important role in this setting.

Model groups Comparing general-purpose and specialised (Polish-language oriented, medical) models highlights several differences. The general-purpose group not only consistently performs well on the original questions from multiple medical domains but also performs well both in English and Polish. There is no significant gap between the general-purpose models and Polish LLMs on Polish exams, and the discrepancy between medical models and general ones on various exams is in favour of the multi-purpose ones.

8 Conclusion

We extended and refined a Polish medical exam benchmark to provide a more informative evaluation of LLMs beyond standard single-choice questions. Our results show that both question structure and domain coverage substantially influence model performance, suggesting that high scores may not reliably reflect true medical competence. These findings underscore the need for more robust, linguistically and contextually grounded evaluation frameworks as well as better pre-training protocols for medical LLMs.

	Original vs modified avg difference (%)						
Model	LEK	LDEK	PES	PESDL	PESF	LEK-ENG	LDEK-ENG
Bielik-4.5B	-32.22	-23.50	-21.73	-17.79	-12.07	-30.74	-23.93
Bielik-11B	-44.56	-30.80	-29.35	-35.60	-34.86	-31.81	-24.23
PLLuM-12B-NC-Inst	-33.97	-24.62	-22.20	-26.22	-17.28	-31.47	-24.11
PLLuM-12B-NC-Chat	-35.93	-27.98	-23.18	-25.95	-20.71	-35.10	-26.00
PLLuM-12B-Inst	-31.97	-25.27	-22.64	-26.49	-24.04	-32.62	-22.58
PLLuM-12B-Chat	-27.64	-21.29	-20.26	-24.37	-23.00	-22.28	-17.44
PLLuM-70B-Inst	-51.60	-38.83	-37.37	-42.89	-35.90	-51.61	-37.76
LLaMA3-8B	-21.22	-17.46	-16.23	-23.24	-22.69	-32.82	-21.04
Gemma-3-12B	-54.44	-41.80	-38.53	-46.99	-53.49	-47.85	-34.22
Gemma-3-27B	-50.68	-38.40	-35.15	-41.84	-31.33	-49.53	-34.22
Qwen3-30B-A3B	-36.64	-28.85	-28.52	-34.58	-28.52	-39.13	-32.68
LLaMA3.3-70B	-49.10	-42.98	-40.19	-43.17	-38.91	-53.63	-41.90
Qwen2.5-72B	-36.43	-31.42	-29.76	-33.43	-26.02	-38.39	-32.62
MedGemma-4B	-37.55	-30.23	-28.64	-31.89	-33.82	-41.88	-29.37
MediPhi-Instruct	9.46	9.40	6.40	-0.95	-4.58	-34.84	-24.88
BioMistral-7B	-25.72	-21.12	-21.31	-25.10	-19.87	-34.30	-25.00
BioMistral-7B-DARE	-22.05	-17.28	-17.40	-22.93	-13.53	-34.43	-25.83
MedGemma-27B	-55.90	-42.52	-39.78	-44.56	-40.59	-58.45	-45.39
OpenBioLLM-70B	-36.10	-26.40	-27.38	-33.40	-27.68	-37.65	-25.95
Meditron3-70B	-29.10	-27.48	-21.25	-26.95	-22.79	-47.51	-37.47

Table 5: Weighted (by the number of questions in each question group) average differences between the original question scores and the modified (Multiple Statement – MS, Multiple Answer – MA, Correct answer substitution – AS).

	Open-ended question accuracies (%)							
Model	LEK	LDEK	PES	PESDL	PESF	LEK-ENG	LDEK-ENG	Avg
Bielik-4.5B	32.70	20.14	15.15	18.98	0.00	27.74	16.67	18.77
Bielik-11B	38.39	25.60	17.93	26.62	23.81	36.50	21.51	27.19
PLLuM-12B-NC-Inst	44.55	25.94	20.20	32.41	14.29	38.69	26.34	28.92
PLLuM-12B-NC-Chat	49.76	24.57	21.21	29.40	9.52	43.07	24.19	28.82
PLLuM-12B-Inst	36.97	22.18	16.41	23.84	14.29	43.80	26.34	26.26
PLLuM-12B-Chat	40.28	19.45	16.16	21.53	4.76	37.23	22.58	23.14
PLLuM-70B-Inst	46.45	30.38	21.97	37.27	38.10	62.77	44.09	40.15
LLaMA3-8B	28.44	18.77	15.66	22.45	9.52	37.96	32.80	23.66
Gemma-3-12B	39.34	26.96	19.70	23.84	9.52	54.74	31.18	29.33
Gemma-3-27B	46.45	30.72	25.51	34.72	28.57	50.36	40.32	36.66
Qwen3-30B-A3B	51.18	31.06	28.54	39.58	38.10	66.42	46.24	43.02
LLaMA3.3-70B	51.18	29.01	26.26	40.05	23.81	60.58	44.62	39.36
Qwen2.5-72B	43.60	28.33	26.52	37.50	47.62	53.28	44.09	40.13
MedGemma-4B	22.27	20.82	11.87	15.05	14.29	40.88	22.58	21.11
MediPhi-Instruct	9.48	6.14	5.56	9.49	0.00	38.69	33.33	14.67
BioMistral-7B	10.43	9.22	5.30	8.10	0.00	39.42	27.42	14.27
BioMistral-7B-DARE	20.38	13.99	8.08	11.11	14.29	31.39	23.12	17.48
MedGemma-27B	48.82	29.35	24.24	36.34	23.81	52.55	50.00	37.87
OpenBioLLM-70B	48.34	30.38	25.00	38.19	28.57	54.74	45.16	38.63
Meditron3-70B	41.71	27.30	25.51	34.95	19.05	57.66	47.85	36.29

Table 6: Performance of models on open-ended questions across exams. Values denote the percentage of correctly generated answers judged by the GPT-oss-120B. "*Avg*" column is the mean across accuracies on all exams.

Limitations

While we present comprehensive evaluation of various models, due to computational and cost limitations, we did not exhaust all of the experimental possibilities, which could include more models in the same size range, another family of models like Mistral, or flagship commercial models like Claude or GPT5.

The GPT-oss-120B model used in the LLM-as-a-judge approach, although shows high agreement with expert annotations, is not specialised in the medical domain. This may limit its ability to consistently identify subtle clinical inaccuracies or domain-specific reasoning errors.

The regex approach employed to filter the questions into modification groups does not capture all edge cases, leaving the possibility of errors in generated answers due to mismatches between questions and instructions. Such errors are expected to be infrequent but could introduce noise into the analysis for certain modification types.

Ethical Considerations

The exam questions used in this work were originally created by the Centre of Medical Examination (Centrum Egzaminów Medycznych - CEM) and made publicly available. Our contribution is limited to restructuring and modifying them. The part of our dataset was also developed based on an existing benchmark dataset (Grzybowski et al., 2025), with permission from its authors, and we preserve attribution to the original sources.

To validate the quality of the LLM-as-a-judge setup, we used Polish-speaking human annotators with medical expertise. Annotators were recruited via direct personal invitations and collaborated outside any commercial crowd-annotation platform. This helped ensure domain competence and reduced risks related to low-quality or uninformed annotations.

Importantly, performance on written medical exams reflects only a narrow slice of medical competence. Clinical practice additionally requires, among other things, taking a patient history, performing a physical examination, interpreting laboratory and imaging results, considering contraindications and comorbidities, communicating uncertainty, and making decisions under incomplete information. Therefore, high benchmark scores should not be interpreted as evidence that LLMs "outperform doctors" or are ready to replace clinical professionals.

Finally, while LLMs can be useful for medical education and as decision-support tools, they may still produce hallucinated or incorrect outputs. This creates a safety risk if such outputs are treated as authoritative. We stress that clinicians should be explicitly informed about these limitations and encouraged to verify model outputs, particularly because medical errors can have serious consequences for patient health and life.

References

AI@Meta. 2024. Llama 3 model card.

AI@Meta. 2025. meta-llama/llama-3.3-70b-instruct. https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct. Instruction-tuned Llama 3.3 70B large language model.

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Artificial intelligence in medicine*, 155:102938.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, and 1 others. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

Nikhil Chandak, Shashwat Goel, Ameya Prabhu, Moritz Hardt, and Jonas Geiping. 2025. Answer matching outperforms multiple choice for language model evaluation. *Preprint*, arXiv:2507.02856.

Michał Chojnicki, Katarzyna Kaczmarek-Majer, Paweł Burchardt, Yanwu Ren, and Marek Z Reformat. 2025. Pilot assessment of transparency of llm-based systems to support emergency rooms. In *Proceedings of the Second Workshop on Explainable Artificial Intelligence for the medical domain-25-30 October*.

Yella Diekmann, Chase Fensore, Rodrigo Carrillo-Larco, Eduard Castejon Rosales, Sakshi Shiromani, Rima Pai, Megha Shah, and Joyce Ho. 2025. Llms as medical safety judges: Evaluating alignment with human annotation in patient-facing qa. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 217–224.

Łukasz Grzybowski, Jakub Pokrywka, Michał Ciesiółka, Jeremi Ignacy Kaczmarek, and Marek Kubis. 2025. Polish-english medical knowledge transfer: A new benchmark and results. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9042–9063.

Yu Gu, Jingjing Fu, Xiaodong Liu, Jeya Maria Jose Valanarasu, Noel CF Codella, Reuben Tan, Qianchu Liu, Ying Jin, Sheng Zhang, Jinyu Wang, and 1 others. 2025. The illusion of readiness: Stress testing large frontier models on multimodal medical benchmarks. *arXiv preprint arXiv:2509.18234*.

Krzysztof Jassem, Michał Ciesiółka, Filip Graliński, Piotr Jabłoński, Jakub Pokrywka, Marek Kubis, Monika Jabłońska, and Ryszard Staruch. 2025. Llmzsz {\L}: a comprehensive llm benchmark for polish. *arXiv preprint arXiv:2501.02266*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2567–2577.

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*, pages 2627–2638.

Jan Kocoń, Maciej Piasecki, Arkadiusz Janz, Teddy Ferdinan, Łukasz Radliński, Bartłomiej Koptyra, Marcin Oleksy, Stanisław Woźniak, Paweł Walkowiak, Konrad Wojtasik, Julia Moska, Tomasz Naskręt, Bartosz Walkowiak, Mateusz Gniewkowski, Kamil Szyc, Dawid Motyka, Dawid Banach, Jonatan Dalasiński, Ewa Rudnicka, and 80 others. 2025. Pllum: A family of polish large language models. *arXiv preprint arXiv:2511.03823*.

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. *Preprint*, arXiv:2402.10373.

Jan Nicikowski, Mikołaj Szczepański, Miłosz Miedziaszczyk, and Bartosz Kudliński. 2024. The potential of chatgpt in medicine: an example analysis of nephrology specialty exams in poland. *Clinical kidney journal*, 17(8):sfae193.

Charles Nimo, Tobi Olatunji, Abraham Toluwase Owodunni, Tassallah Abdullahi, Emmanuel Ayodele, Mardhiyah Sanni, Ezinwanne C Aka, Folafunmi Omofoye, Foutse Yuehgoh, Timothy Faniran, and 1 others. 2025. Afrimed-qa: a pan-african, multi-specialty, medical question-answering benchmark dataset. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1948–1973.

Krzysztof Ociepa, Łukasz Flis, Remigiusz Kinas, Adrian Gwoździej, Krzysztof Wróbel, SpeakLeash Team, and Cyfronet Team. 2025a. Bielik-11b-v2.6-instruct model card.

Krzysztof Ociepa, Łukasz Flis, Remigiusz Kinas, Adrian Gwoździej, Krzysztof Wróbel, SpeakLeash Team, and Cyfronet Team. 2025b. Bielik-4.5b-v3-instruct model card.

Krzysztof Ociepa, Łukasz Flis, Remigiusz Kinas, Krzysztof Wróbel, and Adrian Gwoździej. 2025c. Bielik v3 small: Technical report. *Preprint*, arXiv:2505.02550.

Krzysztof Ociepa, Łukasz Flis, Krzysztof Wróbel, Adrian Gwoździej, and Remigiusz Kinas. 2025d. Bielik 11b v2 technical report. *Preprint*, arXiv:2505.02410.

OpenAI. 2025a. GPT-5 mini. `https://chat.openai.com/`.

OpenAI. 2025b. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

OpenMeditron. Meditron3-70b. `https://huggingface.co/OpenMeditron/Meditron3-70B`. Large language model specialized in clinical medicine, based on Llama-3.1.

Ankit Pal and Malaikannan Sankarasubbu. 2024. aaditya/llama3-openbiollm-70b. `https://huggingface.co/aaditya/Llama3-OpenBioLLM-70B`. Open source biomedical LLM fine-tuned from LLaMA-3 with 70B parameters.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions, 2023. *URL https://arxiv. org/abs/2308.11483*.

Jakub Pokrywka, Jeremi Kaczmarek, and Edward Gorzelańczyk. 2024. Gpt-4 passes most of the 297 written polish board certification examinations. *arXiv preprint arXiv:2405.01589*.

Yuval Reif and Roy Schwartz. 2024. Beyond performance: Quantifying and mitigating label bias in llms. *arXiv preprint arXiv:2405.02743*.

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.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2025. Medgemma technical report. *Preprint*, arXiv:2507.05201.

Julia Siebielec, Michal Ordak, Agata Oskroba, Anna Dworakowska, and Magdalena Bujalska-Zadrozny. 2024. Assessment study of chatgpt-3.5's performance on the final polish medical examination: Accuracy in answering 980 questions. In *Healthcare*, volume 12, page 1637. MDPI.

Shrutika Singh, Anton Alyakin, Daniel Alexander Alber, Jaden Stryker, Ai Phuong S Tong, Karl Sangwon, Nicolas Goff, Mathew De La Paz, Miguel Hernandez-Rovira, Ki Yun Park, and 1 others. 2025. The pitfalls of multiple-choice questions in generative ai and medical education. *Scientific Reports*, 15(1):42096.

Szymon Suwala, Paulina Szulc, Aleksandra Dudek, Aleksandra Bialczyk, Kinga Koperska, and Roman Junik. 2023. Chatgpt fails the polish board certification examination in internal medicine: artificial intelligence still has much to learn. *Pol Arch Int Med Pol Arch Med Wewnet*, 133(11).

Annalisa Szymanski, Noah Ziems, Heather A Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A Metoyer. 2025. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 952–966.

Gemma Team. 2025a. Gemma 3.

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

Qwen Team. 2025b. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Dorota Wójcik, Ola Adamiak, Gabriela Czerepak, Oskar Tokarczuk, and Leszek Szalewski. 2024. A comparative analysis of the performance of chatgpt4, gemini and claude for the polish medical final diploma exam and medical-dental verification exam. *MedRxiv*, pages 2024–07.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Rui Yang, Ting Fang Tan, Wei Lu, Arun James Thirunavukarasu, Daniel Shu Wei Ting, and Nan Liu. 2023. Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4):255–263.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors, 2024. *URL https://arxiv.org/abs/2309.03882*.

A Usage of GenAI in research

We used Grammarly models for improving the grammar of this manuscripts as well as ChatGPT for coding. All GenAI outputs were manually verified and accepted by the authors.

B Experimental Setup and Hyperparameters

Experiments were conducted on a cluster consisting of NVIDIA A100 SXM4-40GB GPUs and lasted around 500 GPU/Hours. The maximum number of used GPUS for a single model evaluation was 4. To ensure reproducibility, the generation of the answers was done with a temperature equal to 0, constrained generation (when applicable) was limited to answers A, B, C, D, E, and integers from 1 to 20 with the stop token set to "." depending on the question form.

C Instructions

Your task is to provide answers to a medical test for doctors. From the provided statements, select all that are correct and return only their indexes. End your answer with a period.

Table 7: Instruction for inferring the answer for a **Multiple Statements** type of question.

Your task is to provide answers to a medical test for doctors. From the provided options, select all the correct ones and respond only with those answers. End your answer with a period.

Table 8: Instruction for inferring the answer for a **Multiple Answers** type of question.

Your task is to provide a precise and specialized answer to a medical question for doctors. Answer with one sentence, phrase, medical term, most appropriate clinical course of action, or diagnosis.

Table 9: Instruction for inferring the answer for an **Open-Ended** type of question.

Take on the role of a medical examiner in an oral exam. You ask the doctor a question (QUESTION), and the doctor provides a proposed answer (PROPOSED ANSWER). You also know the correct answer (ORIGINAL ANSWER). Based on your objective judgment, medical knowledge, and known guidelines, is the PROPOSED ANSWER correct? Do you determine that the PROPOSED ANSWER is the correct answer to the QUESTION? If the PROPOSED ANSWER is correct return True, if the PROPOSED ANSWER is incorrect return False. Do not return anything more than a Boolean.

Table 10: Instruction used in **LLM-as-a-judge**.

D Detailed results

Tables 11, 12 and 13 present the detailed results over the different question alteration groups. Each group of modified questions from a given exam is compared with the same set of questions, but unmodified, and an evaluation method with constrained generations is used.

The results of the evaluation on the **MS, and AS** questions (Table 11 and 12) present a significant drop in the accuracies between the original questions and the multiple response ones, showcasing the vulnerabilities in models' reasoning and their medical knowledge.

When it comes to the **AS** results (Table 13), which present the models' ability to abstain when there is no correct answer, the observed drop in accuracy is substantial. The only model that presents improvement over the original evaluations is MedPhi, which on Polish exams obtains an accuracy of over 50%. However, in English, it does not showcase the same trend. Since it is an English-centered model, there might be some training bias that has interfered with the evaluation method in Polish.

Model	Exam Performance (%)													
	LEK		LDEK		PES		PESDL		PESF		LEK-EN		LDEK-EN	
	OM	MS	OM	MS	OM	MS	OM	MS	OM	MS	OM	MS	OM	MS
Bielik-4.5B	22.78	18.25	20.7	10.05	21.98	10.16	3.79	8.98	8.11	19.64	43.88	16.12	35.6	9.93
Bielik-11B	56.26	24.59	37.91	16.12	38.82	12.86	44.32	13.84	60.54	32.61	51.34	19.1	37.42	10.93
PLLuM-12B-NC-Inst	33.94	22.62	23.78	11.04	25.26	11.9	26.89	12.6	29.55	31.71	21.19	10.15	18.05	5.3
PLLuM-12B-NC-Chat	31.37	15.54	25.47	8.26	24.45	9.66	25.82	12.63	31.89	29.19	25.67	7.16	21.52	4.97
PLLuM-12B-Inst	26.4	8.75	22.49	4.98	23.29	7.07	25.26	5.71	28.29	14.23	21.79	10.45	22.68	7.12
PLLuM-12B-Chat	24.13	3.17	21.89	1.59	21.78	2.55	24.54	2.19	23.6	6.13	17.31	5.97	20.36	5.13
PLLuM-70B-Inst	58.67	40.27	39.1	19.3	38.43	20.39	48.63	24.67	57.12	40.0	55.22	42.99	34.77	20.03
LLaMA3-8B	33.79	15.69	28.26	10.35	28.81	10.97	31.07	11.68	38.92	20.36	43.28	25.37	32.12	13.41
Gemma-3-12B	59.13	2.71	40.7	1.59	40.29	2.24	46.41	1.66	61.26	5.59	63.58	41.19	42.72	18.05
Gemma-3-27B	66.06	42.99	46.97	22.29	44.26	23.95	52.32	28.33	66.67	50.99	69.55	46.27	45.36	23.84
Qwen3-30B-A3B	71.34	44.34	51.24	22.59	51.87	24.87	60.97	28.07	70.09	45.59	78.81	50.45	53.64	27.81
LLaMA3.3-70B	73.45	50.53	53.93	25.97	54.73	28.43	60.02	34.3	74.95	50.63	78.21	49.55	51.99	24.83
Qwen2.5-72B	74.81	48.72	56.82	27.06	54.46	29.39	65.63	36.36	76.4	56.04	83.28	56.42	59.11	30.3
MedGemma-4B	39.97	15.38	31.14	11.14	30.67	9.66	31.89	10.35	44.14	17.84	43.88	20.6	29.64	11.92
MediPhi-Instruct	31.83	10.41	27.86	10.55	27.39	8.3	27.64	7.87	33.33	12.61	47.16	25.07	38.58	16.56
BioMistral-7B	25.04	9.8	20.4	6.77	21.67	8.27	20.14	4.24	20.72	11.53	19.7	12.54	19.54	7.62
BioMistral-7B-DARE	18.85	12.82	18.91	8.86	18.66	9.73	20.82	6.63	18.56	15.32	19.4	15.52	21.19	7.78
MedGemma-27B	66.82	30.17	43.98	10.25	44.03	14.33	49.9	18.15	68.47	38.02	72.24	14.33	46.52	3.81
OpenBioLLM-70B	54.9	38.61	36.52	18.41	41.14	20.97	47.26	24.74	62.88	44.68	62.69	42.39	38.25	21.36
Meditron3-70B	52.94	45.7	32.84	21.79	37.2	26.19	43.86	31.59	57.66	50.09	51.64	45.07	34.93	21.69

Table 11: Model evaluation results. OM = Our evaluation methodology used on original questions, MS = *Multiple Statements* results on the same set of questions as OM

Model	Exam Performance (%)													
	LEK		LDEK		PES		PESDL		PESF		LEK-EN		LDEK-EN	
	OM	MA	OM	MA	OM	MA	OM	MA	OM	MA	OM	MA	OM	MA
Bielik-4.5B	46.12	32.33	43.79	22.88	43.07	22.67	46.21	23.79	41.18	35.29	47.97	31.76	52.0	21.33
Bielik-11B	58.62	35.34	51.63	26.8	53.4	21.91	59.31	23.79	58.82	23.53	60.14	36.49	53.33	22.67
PLLuM-12B-NC-Inst	40.52	25.43	32.03	22.88	35.26	18.64	36.72	17.41	47.06	29.41	39.19	16.22	38.67	13.33
PLLuM-12B-NC-Chat	44.83	26.72	35.29	21.57	39.04	18.14	38.97	19.83	41.18	29.41	39.86	19.59	32.0	14.67
PLLuM-12B-Inst	32.76	20.26	29.41	13.73	35.77	13.6	31.38	14.83	17.65	11.76	30.41	13.51	26.67	12.0
PLLuM-12B-Chat	38.79	25.0	24.18	11.76	33.25	15.37	27.76	16.72	35.29	17.65	17.57	14.19	13.33	9.33
PLLuM-70B-Inst	59.05	26.29	50.98	18.95	52.9	19.4	49.83	18.28	76.47	17.65	54.05	30.41	49.33	20.0
LLaMA3-8B	49.57	19.4	47.06	14.38	49.87	16.12	64.66	16.03	52.94	23.53	56.08	15.54	48.0	14.67
Gemma-3-12B	51.29	18.53	44.44	15.69	44.84	17.63	46.03	16.21	64.71	11.76	58.78	25.68	52.0	21.33
Gemma-3-27B	61.21	21.55	50.33	13.73	54.41	14.61	60.69	16.03	52.94	17.65	63.51	23.65	54.67	16.0
Qwen3-30B-A3B	66.38	23.28	57.52	19.61	57.18	17.88	68.1	16.72	76.47	17.65	68.24	29.05	57.33	17.33
LLaMA3.3-70B	71.98	56.47	62.09	38.56	62.47	43.32	67.76	48.1	76.47	70.59	68.24	57.43	60.0	48.0
Qwen2.5-72B	74.57	21.98	64.71	16.34	61.96	16.88	75.52	16.72	82.35	17.65	79.73	25.0	62.67	12.0
MedGemma-4B	47.41	25.86	50.33	20.26	45.59	16.62	51.03	21.55	29.41	11.76	47.97	38.51	49.33	14.67
MediPhi-Instruct	43.1	19.4	48.37	14.38	49.37	17.88	51.21	13.28	47.06	29.41	46.62	26.35	46.67	16.0
BioMistral-7B	40.52	13.36	45.1	13.07	44.33	9.32	50.86	5.52	35.29	11.76	34.46	18.24	41.33	12.0
BioMistral-7B-DARE	43.53	12.5	43.79	13.07	45.09	10.08	51.55	6.55	29.41	5.88	38.51	14.86	48.0	9.33
MedGemma-27B	68.53	23.71	50.98	16.34	55.42	14.61	61.38	17.93	52.94	17.65	66.22	26.35	58.67	20.0
OpeOMioLLM-70B	61.21	21.55	56.86	16.99	54.16	20.91	65.0	17.24	76.47	17.65	62.16	23.65	62.67	12.0
Meditron3-70B	56.47	23.71	54.25	26.14	55.92	20.65	59.14	19.31	76.47	35.29	58.78	34.46	60.0	30.67

Table 12: Model evaluation results. OM = Our evaluation methodology used on original questions, MA = *Multiple Answers* results on the same set of questions as OM

Model	Exam Performance (%)													
	LEK		LDEK		PES		PESDL		PESF		LEK-EN		LDEK-EN	
	OM	AS	OM	AS	OM	AS	OM	AS	OM	AS	OM	AS	OM	AS
Bielik-4.5B	56.05	8.78	40.98	9.2	35.47	6.64	44.47	7.6	50.64	4.63	54.52	20.66	39.59	17.18
Bielik-11B	65.76	12.23	47.39	10.39	44.52	13.02	51.79	11.74	60.67	15.94	58.39	25.52	44.72	22.31
PLLuM-12B-NC-Inst	52.93	6.05	39.49	5.97	35.06	6.05	42.69	5.04	51.16	6.17	45.18	5.66	36.62	5.82
PLLuM-12B-NC-Chat	53.79	6.25	40.67	4.6	35.22	5.91	43.09	5.01	53.98	7.2	46.18	3.38	36.33	4.05
PLLuM-12B-Inst	50.13	8.84	38.56	7.52	35.76	8.57	41.9	7.8	46.79	7.71	46.18	4.17	32.77	5.43
PLLuM-12B-Chat	41.36	8.64	30.1	7.34	29.42	8.19	35.0	6.73	37.53	6.43	35.95	7.25	26.55	6.81
PLLuM-70B-Inst	74.4	5.25	54.35	2.99	55.23	3.88	65.7	4.62	67.1	5.4	73.19	4.37	54.2	2.07
LLaMA3-8B	44.55	23.34	36.26	20.52	37.07	23.86	42.72	20.3	46.02	17.74	61.27	24.63	44.03	22.51
Gemma-3-12B	66.42	9.51	50.25	5.53	47.56	7.48	58.8	7.07	59.9	9.51	69.71	11.22	48.08	7.9
Gemma-3-27B	74.87	10.31	55.66	8.52	54.69	9.63	65.0	8.22	63.24	9.77	74.68	15.0	52.52	11.06
Qwen3-30B-A3B	75.6	35.7	55.78	27.67	58.22	29.77	70.43	37.14	72.49	39.59	80.14	37.44	61.6	25.37
LLaMA3.3-70B	78.59	12.77	60.88	6.65	60.9	8.7	71.56	9.49	73.01	11.83	82.13	13.9	59.62	6.71
Qwen2.5-72B	78.59	40.09	60.88	30.04	62.93	31.53	72.88	40.02	75.32	42.93	82.32	42.5	65.65	32.08
MedGemma-4B	51.0	5.25	40.61	3.98	38.12	4.18	45.03	3.83	49.61	4.37	58.39	5.56	39.88	3.95
MediPhi-Instruct	32.91	61.1	31.59	61.82	30.21	58.57	36.24	57.59	41.65	60.67	62.26	21.05	42.74	16.58
BioMistral-7B	33.38	3.26	27.8	3.05	28.63	3.25	33.6	3.86	38.05	3.08	47.27	1.29	34.06	1.58
BioMistral-7B-DARE	34.11	6.38	27.24	6.72	27.68	6.24	32.7	5.83	31.88	4.11	49.16	2.98	34.85	2.57
MedGemma-27B	76.0	9.91	55.04	6.28	54.04	7.29	63.93	8.14	67.61	12.34	75.57	14.2	56.66	9.18
OpenBioLLM-70B	70.08	25.8	51.87	21.58	54.07	22.26	62.29	21.85	65.55	25.71	75.57	32.27	55.28	25.77
Meditron3-70B	70.94	32.78	56.34	18.66	52.96	26.03	63.48	25.96	67.87	24.16	74.88	10.33	57.95	5.43

Table 13: Model evaluation results. OM = Our evaluation methodology used on original questions, AS = *Correct answer substitution* results on the same set of questions as OM

E Question modification examples

1) Hidden multiple-response questions

Multiple Statements (MS)

Original question:

> Indicate true statements regarding complications associated with using chemotherapy in cancer treatment:1) the most frequent haematological complication is neutropenia (found in 60-88% of the patients treated);2) neutropenic fever is found in ca. 10-50% of patients treated for solid tumours and in over 80% of patients treated for haematological malignancies;3) the G-CSF prophylaxis is recommended only in radical and palliative treatment.
> The correct answer is:
> A. 1,2.
> B. all of the above.
> C. 1,3.
> D. 2 only.
> E. 3 only.

Correct answer: **A**

Modified question:

> Indicate true statements regarding complications associated with using chemotherapy in cancer treatment:1) the most frequent haematological complication is neutropenia (found in 60-88% of the patients treated);2) neutropenic fever is found in ca. 10-50 of patients treated for solid tumours and in over 80% of patients treated for haematological malignancies;3) the G-CSF prophylaxis is recommended only in radical and palliative treatment.

Correct answer: **1, 2**

Multiple Answers (MA)

Original question:

> Which of the following activities are characteristic of six-month-old infants?
> A. supporting their body on the extended arms with partly or fully opened hands.
> B. bringing a toy from one hand to the other.
> C. dropping things on purpose.
> D. answers A,B are correct.
> E. answers A,B,C are correct.

Correct answer: **D**

Modified question:

> Which of the following activities are characteristic of six-month-old infants:
> A. supporting their body on the extended arms with partly or fully opened hands.
> B. bringing a toy from one hand to the other.
> C. dropping things on purpose.

Correct answer: **A, B**

2) Correct Answer substitution (AS)

Original question:

> 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**

Modified question:

> 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. none of the answers is correct
> D. femoral hernia.
> E. spigelian hernia.

Correct answer: **C**

3) Open-Ended (OE)

Original question:

> A 36-year-old multiparous woman went to see a gynecologist because of regular but excessive and painful periods which has lasted for the last few years. This is accompanied by increasing fatigue, general weakness, and more frequent urination. The pathological findings included pale mucosa. Gynecological exam revealed a tumour the size of a 4-month pregnancy. Blood test and transvaginal ultrasound were made. What lab test results and diagnosis can you expect? A. aneamia, uterine myomas.
> B. anaemia, pregnancy.
> C. normal blood count, pregnancy.
> D. normal blood count, simple ovarian cyst.
> E. anaemia, simple ovarian cyst.

Correct answer: **A**

Modified question:

> A 36-year-old multiparous woman went to see a gynecologist because of regular but excessive and painful periods which has lasted for the last few years. This is accompanied by increasing fatigue, general weakness, and more frequent urination. The pathological findings included pale mucosa. Gynecological exam revealed a tumour the size of a 4-month pregnancy. Blood test and transvaginal ultrasound were made. What lab test results and diagnosis can you expect?

Correct answer: **aneamia, uterine myomas**