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

**Antoni Lasik**[1,†] **Jakub Pokrywka**[2] **Łukasz Grzybowski**[2,3]

**Jeremi Ignacy Kaczmarek**[2,4,6] **Gabriela Korzańska**[4] **Janusz Świeczkowski-Feiz**[5,7]

**Oskar Pastuszek**[6] **Paulina Hoffman**[7] **Jakub Tomasz Dąbrowski**[5] **Wojciech Kusa**[1,†]

[1]NASK National Research Institute [2]Adam Mickiewicz University [3]ARAAI Poland
[4]Poznań University of Medical Sciences [5]Centre of Postgraduate Medical Education, Poland
[6]T. Marciniak Lower Silesian Specialist Hospital [7]Medical University of Warsaw

[†] **Correspondence:** {firstname.lastname}@nask.pl

## Abstract

Large language models (LLMs) in medicine are mainly evaluated using multiple-choice question answering (MCQA), which can overestimate real clinical ability due to guessing strategies and answer biases. To address these limitations, we introduce an expanded and more challenging benchmark based on Polish medical exams, adding over 15,000 questions, two new domains, and four structural modifications that reduce MCQA-specific artifacts and better test reasoning. We evaluate 21 LLMs and show that evaluation design strongly affects results. Under our harder setup, the best model (Qwen3.5-122B) drops by 28.4 and 31 pp on English and Polish exams, respectively. Despite low evidence of data contamination, standard MCQA scores do not reliably reflect true medical competence. 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 multiple-choice questions answering (MCQA).

However, MCQA can overestimate performance and fail to capture reasoning depth, uncertainty handling due to the constrained answer space, where models may exploit statistical regularities, answer priors, and inter-choice relationships rather than perform genuine problem-solving. The presence of explicit options enables elimination strategies, pattern matching, and reconstruction of plausible questions from the choices themselves, which can inflate accuracy without corresponding improvements in underlying reasoning capabilities (Balepur et al., 2024, 2025). This is particularly important in the medical domain, where the goal is not only to compare LLMs against one another, but also to accurately assess the limitations of their knowledge, as inadequately evaluated models may lead to harmful consequences in healthcare.

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 the limitations inherent to MCQA benchmarks, we extend and refine the previously introduced medical benchmark for Polish medical examinations (Grzybowski et al., 2025). The objective of this work is to develop a more rigorous evaluation framework that more accurately reflects the underlying knowledge of LLMs by mitigating their ability to exploit structural properties of the MCQA format. Furthermore, we systematically assess the discrepancies between conventional MCQA-based performance and results obtained under the proposed evaluation protocol.

[1] https://github.com/NASK-NLP/Polish-Medical-Exams



*Findings of the Association for Computational Linguistics: EMNLP 2026*, pages 51441–51467
October 24-29, 2026 ©2026 Association for Computational Linguistics

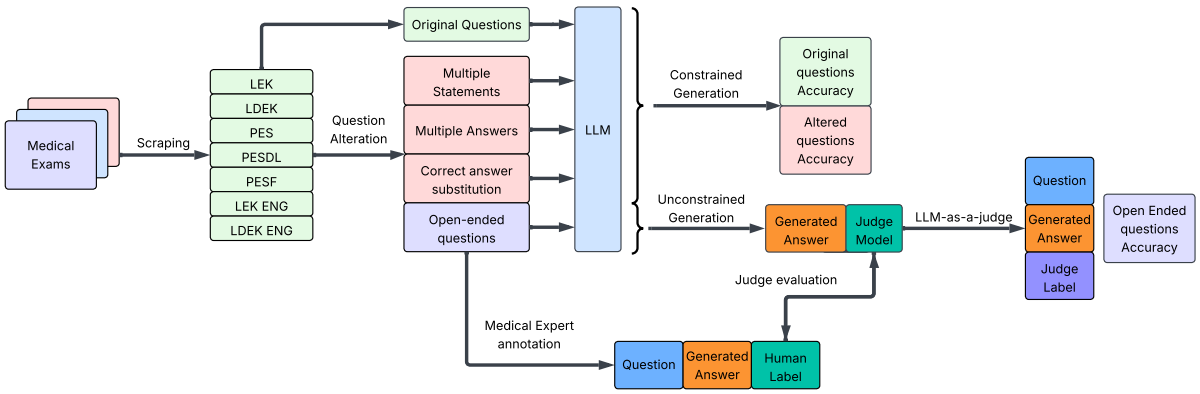



Figure 1: Overview of the evaluation pipeline, from exam question scraping and structural alteration (Multiple Statements, Multiple Answers, Answer Substitution, Open-ended) through constrained/unconstrained answer generation to accuracy scoring and LLM-as-a-judge evaluation.

Our contributions are as follows:

- We extend the previously introduced Polish medical knowledge evaluation dataset[2] by incorporating over 15,000 additional questions from the Pharmaceutical Specialist Examination and the Examination for Laboratory Diagnosticians Specialization, and providing updates for other exams.
- We introduce structural alterations to the question format and an open-ended setup, resulting in a refined evaluation protocol supported by a recomposed and more challenging benchmark, designed to better assess true LLM knowledge while limiting the exploitation of MCQA-specific artifacts.
- We conduct a comprehensive evaluation of 21 LLMs spanning diverse model families, including general-purpose models, domain-specific medical models, and Polish-language models, enabling a systematic comparison of their performance on MCQA tasks and their underlying medical knowledge.

## 2 Related Work

### 2.1 MCQA Evaluation

Balepur et al. (2024) showed that LLMs can achieve above-baseline accuracy in multiple-choice settings even when only answer options are provided, with gains observed in 11 out of 12 evaluated cases and improvements of up to 0.33 accuracy. Their analysis indicates that this behavior cannot be explained solely by memorization or simple priors over individual answers; instead, models exploit group dynamics among choices and are sometimes able to infer plausible underlying questions from the options alone. In a follow-up study, Balepur et al. (2025) identify fundamental limitations of MCQA as an evaluation paradigm, showing that it poorly captures generative reasoning, is susceptible to shortcut strategies, and is affected by issues such as dataset leakage, unanswerability, and benchmark saturation. They further demonstrate that model performance is sensitive to evaluation design choices and advocate for incorporating generative answering formats, improved scoring schemes, and principled test construction methods (e.g., Item Response Theory) to better assess model knowledge and robustness. Chandak et al. (2025) proves that models fine-tuned for MCQA given only the set of answers without the question, are significantly outperforming the random accuracy baseline.

Another limitation of MCQA-based evaluation is the presence of systematic biases in LLM predictions. Selection bias refers to a model's tendency to favor specific answer identifiers (e.g., A/B/C/D) regardless of their semantic content, even when the underlying text is modified (Zheng et al., 2024). Positional bias captures sensitivity to the ordering and placement of options within a question, independent of their meaning (Pezeshkpour and Hruschka, 2023; Zheng et al., 2024). More broadly, label bias arises when models prefer certain labels due to prompt-specific factors, such as label verbalization or the structure of in-context examples, rather than task-relevant information (Reif and Schwartz, 2024).

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

### 2.2 Medical MCQA Evaluation

A substantial portion of research on medical benchmark construction has focused on MCQA datasets, including MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), and PubMedQA (Jin et al., 2019), while comparatively fewer works examine the limitations of MCQA specifically from an evaluation perspective. Although these benchmarks have been instrumental in measuring progress, they often fail to reflect real clinical complexity, lack rigorous expert validation, and are increasingly saturated. Consequently, models tend to rely on pattern recognition rather than robust reasoning, sometimes correctly predicting answers even when key parts of the question are missing (Gu et al., 2025; Singh et al., 2025). Other recent benchmarks, such as MedXpertQA (Zuo et al., 2025), attempt to counter this saturation through expert-level sourcing, difficulty-based filtering, and LLM-driven question rewriting to reduce leakage risk.

To address these issues, alternative evaluation paradigms based on free-text generation or multi-turn interactions have been proposed as more realistic and informative (Arora et al., 2025; Singh et al., 2025). However, such approaches remain difficult to evaluate at scale due to their reliance on expert annotation, and while LLM-as-a-Judge methods offer partial automation, expert oversight is still required in safety-critical domains like healthcare (Szymanski et al., 2025; Diekmann et al., 2025).

### 2.3 Polish MCQA Benchmarks

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*). To our knowledge, there are no studies evaluating Polish LLMs' professional medical knowledge using open-ended responses.

## 3 Polish Medical Exams

LEK and LDEK are Polish final examinations for medical and dental graduates. PES is a board certification examination for physicians and dentists after specialization training. PESDL and PESF are analogous specialist examinations for laboratory diagnosticians and pharmacists. All of them are multiple-choice exams with A–E answer options. They are described in more detail in Appendix C.

Most exam questions are presented in the form of clinical vignettes reflecting realistic diagnostic and therapeutic scenarios. Such tasks often require interpretation of symptoms, laboratory findings, imaging results, and treatment options, thereby assessing applied clinical reasoning in addition to factual medical knowledge (example questions are provided in Appendix F).

## 4 Dataset Creation

In this section we describe how we collected additional exam data and modified the question format.

### 4.1 Additional Datasets Sourcing

We included PESF and PESDL exams using scraping scripts that extract questions and answer keys from PDFs. Two PESF editions were discarded due to OCR errors from scanned PDFs. We also added the most recent available exam editions. In total, this adds over 15,000 questions to the benchmark.

### 4.2 Question Modification

The MCQA format has inherent limitations when used to evaluate the actual knowledge of LLMs. Given a question $Q$ and a set of answer choices $A$, including the correct answer $a_i \in A$, the task is often reduced to distinguishing the correct option from incorrect ones, rather than generating the correct answer independently.

To address this, we define three groups of question structures that can be modified to better re-

flect clinical reasoning. The examples of modified questions in the English version are provided in Appendix F. These changes make questions more clinically oriented by reducing reliance on fixed answer choices.

**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.

The obtained dataset spans over 38,000 questions covers multiple medical domains (see Table 1 for details), and proposed alterations cover 68% of the original questions. The distributions of altered question types vary across exam sets (as shown in 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.

| Exam | #Exam editions (*Years*) | #Original | #Covered | Coverage |
|---|---|---|---|---|
| LEK | 26 (*'08–'15, '20–'26*) | 4,728 | 2,875 | 61% |
| LDEK | 26 (*'08-'15, '20–'26*) | 4,773 | 3,401 | 71% |
| PES | 12 (*'08,'12,'15–'20,'23,'24*) | 9,965 | 7,070 | 71% |
| PESDL | 25 (*'08–'20*) | 10,908 | 7,627 | 70% |
| PESF | 22 (*'08–'20*) | 1,710 | 982 | 57% |
| LEK EN | 18 (*'13–'15, '20–'26*) | 3,151 | 1,899 | 60% |
| LDEK EN | 18 (*'13–'15, '20–'26*) | 3,196 | 2,223 | 70% |
| All | | 38,431 | 26,077 | 68% |

Table 1: Coverage statistics by source.

## 5 Methodology

In this section we describe in detail our evaluation protocol, comparing it to previous effort using Polish medical exams. Moreover, we discuss the LLM-as-a-judge approach taken to evaluate the open-ended questions, the annotation procedure used to construct the meta-evaluation set, the resulting alignment between the judge and human experts, and our analysis of the recurring patterns behind their disagreements.

### 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,c) and PLLuM (Kocoń et al., 2025) variants, but also English-centered models like Qwen (Qwen, 2025) and GPT-5-mini (OpenAI, 2025a) has been conducted using the instruction with example output format, which contained a sample answer (later denoted as **Previous Methodology (PM)**.

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)

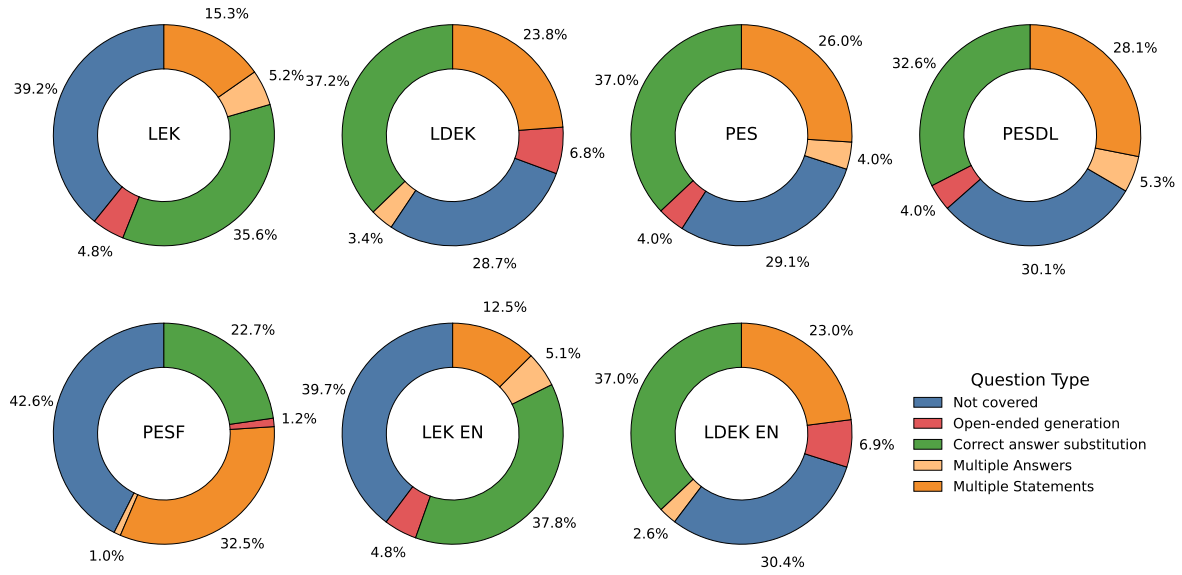



Figure 2: Question type composition across exam question sets.

and the distribution of answers generated by the aforementioned models. Since the formulation of the original 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. Simple instruction paired with constrained generation limits this answer bias. This approach is later denoted as **Our evaluation Methodology (OM)**.

When applicable[3], constrained generation based on a set of allowed tokens was used paired with the instruction suited for the modified and original question types. For example, in **MS** scenario, 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. The refined instruction and all of the instructions for generating answers for modified questions are provided in the Appendix E.

### 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 Deepseek-V4-Pro (DeepSeek-AI, 2026). 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 E.

**Annotations** We build a human meta-evaluation set to test whether LLM-as-a-judge aligns with medical experts and whether out-of-domain LLMs are reliable for evaluation. We sample LEK and PES open-ended questions and collect 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), and Qwen3-30B-A3B-Instruct-2507 (Qwen, 2025). Six medical experts annotated 200 (question, answer, reference answer) triplets each, with 15 control items per set. The task was binary: decide whether the generated answer is correct, with the reference answer as context; annotators were also able to discard unclear or invalid questions.

In total, we obtained 927 valid annotations and

[3]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.

| | Human-Human | Human-Model |
|---|---|---|
| All Annotation Samples | 1,200 | — |
| Unique Annotation Samples | 1,110 | — |
| *Invalid Samples:* | 367 | — |
| → Missing Labels | 16 | — |
| → Invalid Question | 167 | — |
| → Annotator - Outlier | 184 | — |
| *Valid samples:* | — 743 — | |
| → Positive Labels | 220 | 270 |
| → Negative Labels | 523 | 473 |
| Kappa | 0.921 (*Fleiss - w/o outlier*) | 0.649 (*Cohen*) |

Table 2: Annotation process summary.

| | Discrepancy pattern | Definition |
|---|---|---|
| A | Errors in redundant content | The answer included unnecessary additional content which introduced at least one defect. |
| B | Valid alternative interpretation of the question | The answer addressed a valid interpretation of the open-ended question, but not the issue implied by the original closed-ended question and answer key. |
| C | Underspecified expected detail | The answer addressed the expected issue, but with less specificity than the official correct answer. |
| D | Non-existent medical terms | The answer used non-existent medical terms, including hybrid Polish–English or Polish–Latin expressions. |
| E | Misuse of existing medical terms | The answer used existing medical terms inconsistently with their accepted meaning. |
| F | Incorrect spelling or notation of medical terms | The answer used existing medical terms with incorrect spelling or notation. |

Table 3: Overview of discrepancy patterns used to label potential sources of disagreement between human annotators and AI judges.

90 control samples (see Table 2); over 150 items were removed due to poor or outdated questions. We computed inter-annotator agreement (Fleiss' $\kappa$) on controls and removed one outlier annotator. We then measured agreement between experts and the LLM judge using Cohen's $\kappa$, obtaining 0.649, which indicates substantial agreement but shows that LLM-based evaluation is not equivalent to expert judgment. Annotation and LLM judge guidelines are aligned across both setups.

**Disagreement analysis** We analysed disagreements between human annotators and LLM-as-a-judge models to identify common sources of divergence. A senior physician reviewed a subset of 74 cases with inconsistent judgments and identified six recurring **discrepancy patterns** (Table 3).

Further annotation of these cases showed that disagreements mainly arise from two factors: (1) linguistic issues (e.g., poor Polish fluency, incorrect terminology), and (2) differences in judging correctness. Human annotators tend to focus on clinically or linguistically relevant errors, while LLM judges rely more on surface overlap with the reference answer and often miss contextual or off-target errors. Additional details are provided in Appendix I.

## 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), MediPhi (Corbeil et al., 2025).

**General-purpose models:** Qwen3.5 in three sizes 9B, 35B and 122B (Team Qwen, 2026), Llama3.3-70B (AI@Meta, 2025), Llama3-8B (AI@Meta, 2024), Gemma3 (12B and 27B versions) (Team, 2025).

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

| Model | Avg PM | Avg OM | Δ (OM−PM) |
|---|---|---|---|
| *Polish LLMs* | | | |
| Bielik-4.5B | 36.8 | 37.8 | +1.1 |
| Bielik-11B | **55.0** | 55.8 | +0.8 |
| PLLuM-12B-NC-Inst | 36.5 | 37.3 | +0.8 |
| PLLuM-12B-NC-Chat | 37.3 | 38.1 | +0.7 |
| PLLuM-12B-Inst | 27.5 | 35.9 | +8.4 |
| PLLuM-12B-Chat | 17.9 | 29.5 | +11.6 |
| PLLuM-70B-Inst | 39.4 | **57.1** | +17.7 |
| *Average* | *35.8* | *41.6* | *+5.9* |
| *General LLMs* | | | |
| LLaMA3-8B | 43.3 | 41.8 | -1.4 |
| Qwen3.5-9B | 42.9 | 63.1 | +20.2 |
| Gemma-3-12B | 44.5 | 53.3 | +8.9 |
| Gemma-3-27B | 55.9 | 59.2 | +3.3 |
| Qwen3.5-35B-A3B | 71.0 | 72.2 | +1.2 |
| LLaMA3.3-70B | 66.4 | 65.8 | -0.6 |
| Qwen3.5-122B-A10B | **75.8** | **75.1** | -0.7 |
| *Average* | *57.1* | *61.5* | *+4.4* |
| *Medical LLMs* | | | |
| MedGemma-4B | 38.0 | 43.4 | +5.3 |
| MediPhi-Instruct | 31.2 | 37.0 | +5.8 |
| BioMistral-7B | 21.1 | 30.6 | +9.6 |
| BioMistral-7B-DARE | 24.3 | 30.7 | +6.4 |
| MedGemma-27B | 56.9 | **60.0** | +3.1 |
| OpenBioLLM-70B | 50.2 | 57.2 | +7.0 |
| Meditron3-70B | 16.8 | 55.8 | +38.9 |
| *Average* | *34.1* | *45.0* | *+10.9* |

Table 4: Average model performance under the previous MCQA protocol (PM) and our proposed bias-mitigated methodology (OM). Averages computed over all available exams (LEK, LDEK, PES, PESDL, PESF, LEK-EN, LDEK-EN). Best result per model family in **bold**, best overall is underlined. Full per-exam breakdown in Table 11.

| Model | Avg Mod | Avg Og | Δ EN | Δ PL |
|---|---|---|---|---|
| *Polish LLMs* | | | | |
| Bielik-4.5B | 13.01 | 36.30 | -27.64 | -21.56 |
| Bielik-11B | **17.49** | 56.27 | -40.64 | -38.04 |
| PLLuM-12B-NC-Inst | 10.65 | 36.39 | -27.94 | -24.87 |
| PLLuM-12B-NC-Chat | 9.50 | 37.22 | -30.40 | -26.66 |
| PLLuM-12B-Inst | 8.03 | 34.71 | -27.92 | -26.18 |
| PLLuM-12B-Chat | 6.72 | 29.11 | -20.24 | -23.26 |
| PLLuM-70B-Inst | 14.62 | **57.04** | -45.26 | -41.29 |
| Average | 11.43 | 41.01 | -31.43 | -28.84 |
| *General LLMs* | | | | |
| LLaMA3-8B | 19.12 | 41.27 | -26.92 | -20.25 |
| Qwen3.5-9B | 21.93 | 64.43 | -42.38 | -42.55 |
| Gemma-3-12B | 9.08 | 54.35 | -41.04 | -46.96 |
| Gemma-3-27B | 19.92 | 59.96 | -41.56 | -39.43 |
| Qwen3.5-35B | 39.98 | 73.96 | -37.59 | -32.53 |
| LLaMA3.3-70B | 22.92 | 67.14 | -47.82 | -42.78 |
| Qwen3.5-122B | **46.97** | **77.28** | -28.49 | -31.04 |
| Average | 25.70 | 62.63 | -37.97 | -36.51 |
| *Medical LLMs* | | | | |
| MedGemma-4B | 9.10 | 42.49 | -36.08 | -32.33 |
| BioMistral-7B | 5.49 | 30.02 | -29.36 | -22.60 |
| BioMistral-7B-DARE | 7.54 | 29.48 | -30.25 | -18.62 |
| MediPhi-Instruct | **31.87** | 37.77 | -30.16 | +3.80 |
| MedGemma-27B | 14.14 | **60.78** | -52.01 | -44.49 |
| OpenBioLLM-70B | 26.66 | 57.41 | -32.22 | -30.17 |
| Meditron3-70B | 25.93 | 56.41 | -42.51 | -25.45 |
| Average | 17.25 | 44.89 | -36.08 | -24.27 |

Table 5: Average model performance on modified questions (Avg Mod), original variants of the same questions (Avg Og) and deltas between the performance on modfied questions and original questions on English (Δ EN) exams (LEK-EN, LDEK-EN) and Polish (Δ PL) exams (LEK, LDEK, PES, PESDL, PESF). **Bold** values indicate highest scores within model group, while highest value overall is underlined. Full per-exam breakdown in Table 12.

### 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 judgment from the described LLM-as-a-judge approach using Deepseek-V4-Pro (DeepSeek-AI, 2026).

## 7 Results and Discussion

**Limiting the bias** Table 4 compares the replicated evaluation protocol from Grzybowski et al. (2025) with the proposed bias-mitigated methodology. Most models achieve better results under the proposed protocol, which supports the hypothesis that the original instruction and generation setup introduces bias that lowers benchmark scores. The difference is especially visible for some instruction-tuned models, such as PLLuM-70B-Inst, Qwen3.5-9B, and Meditron3-70B.

The proposed protocol also changes the ranking of the Polish LLMs. Under the bias-mitigated evaluation, PLLuM-70B-Inst achieves a higher average score than Bielik-11B, while the previous methodology suggests the opposite ranking. This suggests that the original protocol may affect some models more strongly than others. Overall, the results indicate that limiting instruction-related bias provides a more reliable comparison between models.

**Question alteration effect** Table 5 presents the differences in accuracy between the original and altered questions. The significant performance drop observed across models and examinations in both

| Model | Open-ended question accuracies (%) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | LEK | LDEK | PES | PESDL | PESF | LEK-EN | LDEK-EN | Avg |
| *Polish LLMs* | | | | | | | | |
| Bielik-4.5B | 34.67 | 21.23 | 14.14 | 22.92 | 4.76 | 29.61 | 19.46 | 20.97 |
| Bielik-11B | **58.22** | **38.77** | **31.06** | 38.43 | 38.10 | 59.87 | 33.94 | 42.63 |
| PLLuM-12B-NC-Inst | 48.44 | 31.38 | 22.73 | 36.34 | 28.57 | 36.84 | 28.05 | 33.19 |
| PLLuM-12B-NC-Chat | 54.22 | 32.31 | 25.25 | 35.19 | 14.29 | 42.76 | 25.34 | 32.77 |
| PLLuM-12B-Inst | 41.33 | 24.62 | 19.95 | 26.85 | 19.05 | 41.45 | 24.43 | 28.24 |
| PLLuM-12B-Chat | 43.11 | 24.00 | 16.41 | 24.77 | 19.05 | 37.50 | 20.81 | 26.52 |
| PLLuM-70B-Inst | 54.67 | 36.62 | 26.52 | **45.37** | **52.38** | **63.16** | **46.15** | **46.41** |
| *General LLMs* | | | | | | | | |
| LLaMA3-8B | 25.78 | 18.77 | 15.66 | 22.69 | 9.52 | 39.47 | 32.58 | 23.50 |
| Qwen3.5-9B | 45.78 | 32.92 | 25.25 | 39.81 | 42.86 | 63.16 | 48.87 | 42.66 |
| Gemma-3-12B | 45.33 | 29.54 | 26.77 | 29.63 | 23.81 | 54.61 | 33.48 | 34.74 |
| Gemma-3-27B | 56.00 | 40.92 | 33.59 | 46.76 | 52.38 | 54.61 | 43.44 | 46.81 |
| Qwen3.5-35B-A3B | 63.56 | 42.77 | 38.64 | 55.79 | 47.62 | 65.79 | 52.04 | 52.32 |
| LLaMA3.3-70B | 60.89 | 36.00 | 30.05 | 46.76 | 33.33 | 62.50 | 48.87 | 45.49 |
| Qwen3.5-122B-A10B | **68.00** | **53.54** | **52.78** | **61.11** | **76.19** | **73.03** | **57.01** | **63.09** |
| *Medical LLMs* | | | | | | | | |
| MedGemma-4B | 27.11 | 18.46 | 12.63 | 17.36 | 23.81 | 41.45 | 22.62 | 23.35 |
| MediPhi-Instruct | 6.67 | 5.85 | 4.80 | 7.41 | 9.52 | 40.13 | 33.48 | 15.41 |
| BioMistral-7B | 16.00 | 8.92 | 6.06 | 9.95 | 4.76 | 36.84 | 26.70 | 15.60 |
| BioMistral-7B-DARE | 22.67 | 13.85 | 7.32 | 13.66 | 9.52 | 36.18 | 27.15 | 18.62 |
| MedGemma-27B | **56.44** | 33.54 | **33.33** | **44.68** | 33.33 | 56.58 | 45.25 | 43.31 |
| OpenBioLLM-70B | 55.11 | **35.08** | 28.79 | 41.90 | 23.81 | 57.24 | **46.61** | 41.22 |
| Meditron3-70B | 51.11 | 34.46 | 32.07 | 42.36 | **52.38** | **66.45** | **46.61** | **46.49** |

Table 6: Performance of models on open-ended questions across exams as the percentage of correctly generated answers judged by the Deepseek-V4-Pro. "*Avg*" column is the mean across accuracies on all exams. **Bold** indicates the best performer within each model group, while underline denotes the best score overall.

Polish and English indicates that model performance is highly dependent on the evaluation format. General-purpose models achieved the strongest results under the altered conditions, with models from the Qwen family achieving the best overall performance and Gemma-3-27B outperforming its medical counterpart. Notably, Qwen3.5-35B-A3B stood out for its efficiency, achieving better performance than Meditron3-70B, Llama3.3-70B, and OpenBioLLM-70B despite its substantially smaller parameter count. At the same time, Llama3.3-70B failed to outperform medical models of comparable size.

The alterations expose the vulnerabilities of models, showing that under the changed conditions they fail to provide the correct answer, especially in lower-resource language like Polish, where the average performance on modified questions is particularly low. This only shows how it is important to create a robust and reliable evaluation for lower-resource languages in the critical domains like medicine.

The only model that experienced gains in performance over the refined evaluation was MediPhi-Instruct, becoming the best performer in the Medical LLMs group. However, under further investigation the source of the improvement was found in the questions from the **AS** question group on Polish exams (see Appendix D Table 10 and Appendix F.3). The MediPhi inconsistency between performance on English and Polish within the same alteration group can be attributed to the model's overfitting towards the "None of the answers is correct" option.

**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 presented in Table 6 show that, on average, the evaluated LLMs fail to correctly answer expert-level exam questions. Medical LLMs, which are expected to perform particularly well in this setting, achieved results comparable to those of general-purpose Polish LLMs trained holistically rather than specifically for medical tasks.

The relatively low performance on the question pool spanning multiple medical specialisations and domains, since this setting is closest to the day-to-day use of the LLMs by regular consumers.

**Model groups** Among models with comparable parameter counts across the groups, we observe that the Polish PLLuM-70B experienced a substantial performance drop on the altered questions compared to general LLaMA-3.3-70B, or medical Meditron3 or OpenBioLLM, which performed the best for their size. On open-ended questions, these models perform comparably, but Meditron3 outperforms the others by a slight margin.

These comparisons are complex since the models need to be fluent in English, Polish, and have expert-level medical knowledge. Additionally, the reformulated evaluation is more demanding since it does not follow the classic MCQA approach.

**Detailed results** The detailed breakdown of the evaluation results, per model, exam, question alteration group, and evaluation methodology, is provided in the Appendix D.

**Data contamination** Since the benchmark was constructed from publicly available data, exposure of the questions in LLM training corpora is possible. To investigate this, we conducted additional temporal contamination analysis using the Data Contamination Quiz (DCQ), which is a lightweight method introduced by Golchin and Surdeanu (2025) for evaluating data contamination in a model's training data. The method is based on creating a quiz in which the model must identify

| Model | 2015 | 2020 | 2026 |
|---|---|---|---|
| Bielik-11B | (17.0, 24.5) | (13.5, 19.0) | (12.5, 23.5) |
| PLLuM-12B-Inst | (13.5, 19.0) | (10.37, 26.5) | (18.0, 20.0) |
| LLaMA-3.3-70B | (17.5, 18.0) | (18.5, 19.0) | (13.5, 18.0) |
| Qwen3.5-35B | (8.59, 9.5) | (5.53, 6.0) | (9.6, 10.5) |
| MedGemma-27B | (6.63, 15.5) | (3.3, 12.0) | (4.74, 9.5) |
| Meditron3-70B | (9.5, 9.5) | (10.5, 12.5) | (11.0, 13.0) |

Table 7: Contamination level ranges reported using DCQ methodology across editions of LEK and LDEK, provided in format (min contamination, max contamination).

the original instances from a target dataset among several perturbed variants of those instances. We created such a quiz for LEK and LDEK exam questions from three different spring editions (2015, 2020, and 2026) 200 questions each. This setup allows us to assess whether the editions suspected of being leaked into the training data (2015 and 2020) exhibit higher contamination scores than the edition that could not have been leaked (2026). The results suggest limited contamination (see Table 7), yet the reported levels do not align with the performance drops observed when applying question alterations. Notably, the DCQ measured exposure to exam questions only, which are considerably more accessible than their corresponding answers. Furthermore, if contamination were present, the levels reported for the 2026 edition would be expected to be substantially lower than those for the 2015 and 2020 editions – a pattern that is not observed in the data.

Moreover, we have performed Web Search Audit for LEK and PES exam questions and temporal analysis of the results on LEK EN and LDEK EN. The audit revealed exposure of LEK exam questions and none for PES. However, the vast majority of the retrieved results for the LEK exam have not contained corresponding answers. The analysis of the model's performance on LEK and LDEK examinations over the years does not suggest that questions from the latest edition (2026) are an outlier compared to the potentially exposed editions.

## 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.

## 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 Deepseek-V4-Pro model used in the LLM-as-a-judge approach, although shows substantial 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 contraindi-

cations 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.

## Acknowledgments

We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2026/019405.

## 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*.

Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. 2024. Artifacts or abduction: How do LLMs answer multiple-choice questions without the question? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10308–10330, Bangkok, Thailand. Association for Computational Linguistics.

Nishant Balepur, Rachel Rudinger, and Jordan Lee Boyd-Graber. 2025. Which of these best describes multiple choice evaluation with LLMs? a) forced B) flawed C) fixable D) all of the above. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3394–3418, Vienna, Austria. Association for Computational Linguistics.

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*.

Jean-Philippe Corbeil, Amin Dada, Jean-Michel Attendu, Asma Ben Abacha, Alessandro Sordoni, Lucas Caccia, Francois Beaulieu, Thomas Lin, Jens Kleesiek, and Paul Vozila. 2025. A modular approach for clinical SLMs driven by synthetic data with pre-instruction tuning, model merging, and clinical-tasks alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19352–19374, Vienna, Austria. Association for Computational Linguistics.

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.

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.

Shahriar Golchin and Mihai Surdeanu. 2025. Data contamination quiz: A tool to detect and estimate contamination in large language models. *Transactions of the Association for Computational Linguistics*, 13:809–830.

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

Krzysztof Ociepa, Remigiusz Kinas, Krzysztof WrĂłbel, Adrian GwoŁşdziej, and 1 others. 2025b. Bielik 11b v3: Multilingual large language model for european languages. *arXiv preprint arXiv:2601.11579*.

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

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*.

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

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. 2025. Gemma 3.

Team Qwen. 2026. Qwen3.5: Towards native multimodal agents.

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.

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. 2024. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, volume 2024, pages 19426–19454.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*.

## A Usage of GenAI in Research

We used Grammarly and GPT 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.

For the data creation and formatting we have used standard packages: Pandas, Huggingface Datasets, re, pickle. For the LLMs evaluation, we were using vLLM and Huggingface Transformers with two separate environments: one for most of the models (vLLM 0.10.1.1, Transformers 4.56.0), the second one for the newest models from the Qwen3.5 series (vLLM 0.19.0, Transformers 5.5.4). For the LLM-as-a-judge testing and deployment: OpenAI, vLLM, and anthropic packages. For visualisations seaborn and matplotlib.

## C 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, placement into a specialization depends on ranking; therefore, most candidates aim for scores 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)[4] and Supreme Medical Chamber (Naczelna Izba Lekarska - NIL).[5] Recently, these

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

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.

## D Detailed results

Tables 8, 9 and 10 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 8 and 9) 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 10), 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 | 22.96 | 17.57 | 21.99 | 9.94 | 21.98 | 10.16 | 3.79 | 8.98 | 8.11 | 19.64 | 43.54 | 15.95 | 36.24 | 9.26 |
| Bielik-11B | **62.79** | **42.05** | **44.06** | **19.26** | **40.17** | 19.04 | 47.88 | 21.54 | **65.05** | **45.05** | **58.73** | 35.44 | **40.33** | 16.35 |
| PLLuM-12B-NC-Inst | 33.75 | 22.41 | 23.66 | 10.73 | 25.26 | 11.9 | 26.89 | 12.6 | 29.55 | 31.71 | 22.28 | 9.62 | 18.12 | 5.59 |
| PLLuM-12B-NC-Chat | 31.12 | 15.63 | 25.42 | 8.18 | 24.45 | 9.66 | 25.82 | 12.63 | 31.89 | 29.19 | 26.33 | 8.1 | 20.84 | 4.9 |
| PLLuM-12B-Inst | 27.25 | 8.44 | 22.25 | 4.66 | 23.29 | 7.07 | 25.26 | 5.71 | 28.29 | 14.23 | 22.53 | 10.63 | 21.93 | 5.59 |
| PLLuM-12B-Chat | 24.76 | 3.46 | 21.55 | 1.58 | 21.78 | 2.55 | 24.54 | 2.19 | 23.6 | 6.13 | 17.97 | 7.09 | 20.03 | 5.04 |
| PLLuM-70B-Inst | 57.4 | 39.97 | 39.14 | 18.73 | 38.43 | **20.39** | **48.63** | **24.67** | 57.12 | 40.0 | 54.94 | **41.77** | 35.83 | **18.94** |
| LLaMA3-8B | 34.3 | 15.35 | 29.02 | 10.11 | 28.81 | 10.97 | 31.07 | 11.68 | 38.92 | 20.36 | 43.8 | 25.06 | 31.74 | 13.62 |
| Qwen3.5-9B | 70.54 | 7.47 | 47.23 | 3.52 | 50.41 | 3.55 | 58.16 | 4.77 | 72.07 | 13.15 | 74.94 | 29.37 | 47.68 | 11.99 |
| Gemma-3-12B | 57.68 | 2.63 | 40.37 | 1.5 | 40.29 | 2.24 | 46.41 | 1.66 | 61.26 | 5.59 | 63.04 | 41.27 | 42.23 | 17.71 |
| Gemma-3-27B | 65.42 | 43.43 | 46.35 | 22.25 | 44.26 | 23.95 | 52.32 | 28.33 | 66.67 | 50.99 | 68.1 | 46.08 | 45.37 | 23.3 |
| Qwen3.5-35B-A3B | 81.05 | 59.75 | 56.64 | 29.64 | 64.43 | 36.89 | 70.1 | 38.74 | 79.1 | 58.56 | 81.77 | 56.46 | 60.22 | 29.97 |
| LLaMA3.3-70B | 73.72 | 50.48 | 52.95 | 26.47 | 54.73 | 28.43 | 60.02 | 34.3 | 74.95 | 50.63 | 77.22 | 48.61 | 51.36 | 24.25 |
| Qwen3.5-122B | **83.96** | **62.79** | **62.88** | **33.16** | **66.55** | **39.67** | **72.32** | **42.43** | **82.16** | **60.72** | **83.54** | **63.04** | **63.08** | **37.87** |
| MedGemma-4B | 39.28 | 15.49 | 30.96 | 10.82 | 30.67 | 9.66 | 31.89 | 10.35 | 44.14 | 17.84 | 42.03 | 19.49 | 31.2 | 11.17 |
| MediPhi-Instruct | 31.81 | 9.96 | 28.41 | 9.67 | 27.39 | 8.3 | 27.64 | 7.87 | 33.33 | 12.61 | 46.33 | 26.58 | 36.78 | 15.26 |
| BioMistral-7B | 24.34 | 9.27 | 20.23 | 6.6 | 21.67 | 8.27 | 20.14 | 4.24 | 20.72 | 11.53 | 19.24 | 13.42 | 17.98 | 7.77 |
| BioMistral-7B-DARE | 18.95 | 12.59 | 18.56 | 8.88 | 18.66 | 9.73 | 20.82 | 6.63 | 18.56 | 15.32 | 21.52 | 15.95 | 20.3 | 8.45 |
| MedGemma-27B | **65.56** | 29.46 | **43.18** | 10.47 | **44.03** | 14.33 | **49.9** | 18.15 | **68.47** | 38.02 | **70.13** | 13.67 | **46.32** | 4.5 |
| OpenBioLLM-70B | 54.63 | 38.73 | 36.5 | 18.65 | 41.14 | 20.97 | 47.26 | 24.74 | 62.88 | 44.68 | 61.52 | 42.78 | 38.42 | **21.66** |
| Meditron3-70B | 53.11 | **46.06** | 32.1 | **21.99** | 37.2 | **26.19** | 43.86 | **31.59** | 57.66 | **50.09** | 50.63 | **45.57** | 35.42 | 20.71 |

Table 8: 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 | 46.34 | 33.33 | 43.83 | 22.22 | 43.07 | 22.67 | 46.21 | 23.79 | 41.18 | 35.29 | 49.38 | 33.33 | **50.00** | 22.62 |
| Bielik-11B | **63.01** | **44.31** | 46.91 | **32.10** | 50.38 | **30.23** | **53.28** | **26.55** | 70.59 | **47.06** | **63.58** | **47.53** | 46.43 | **23.81** |
| PLLuM-12B-NC-Inst | 41.06 | 25.20 | 32.72 | 24.07 | 35.26 | 18.64 | 36.72 | 17.41 | 47.06 | 29.41 | 40.74 | 14.81 | 36.90 | 13.10 |
| PLLuM-12B-NC-Chat | 44.31 | 26.02 | 35.19 | 23.46 | 39.04 | 18.14 | 38.97 | 19.83 | 41.18 | 29.41 | 39.51 | 17.90 | 29.76 | 16.67 |
| PLLuM-12B-Inst | 30.89 | 19.11 | 30.25 | 15.43 | 35.77 | 13.60 | 31.38 | 14.83 | 17.65 | 11.76 | 30.25 | 12.35 | 27.38 | 11.90 |
| PLLuM-12B-Chat | 39.43 | 25.20 | 22.84 | 13.58 | 33.25 | 15.37 | 27.76 | 16.72 | 35.29 | 17.65 | 18.52 | 12.35 | 17.86 | 9.52 |
| PLLuM-70B-Inst | 59.76 | 26.83 | **51.85** | 18.52 | **52.90** | 19.40 | 49.83 | 18.28 | **76.47** | 17.65 | 55.56 | 29.63 | **50.00** | **23.81** |
| LLaMA3-8B | 50.00 | 19.51 | 46.30 | 16.05 | 49.87 | 16.12 | 64.66 | 16.03 | 52.94 | 23.53 | 54.32 | 16.05 | 46.43 | 16.67 |
| Qwen3.5-9B | 76.83 | 23.58 | 58.02 | 21.60 | 60.20 | 17.63 | 71.38 | 15.86 | 76.47 | 11.76 | 74.07 | 27.78 | **66.67** | 23.81 |
| Gemma-3-12B | 50.41 | 18.29 | 43.21 | 17.90 | 44.84 | 17.63 | 46.03 | 16.21 | 64.71 | 11.76 | 60.49 | 24.07 | 48.81 | 21.43 |
| Gemma-3-27B | 61.79 | 20.33 | 50.62 | 16.05 | 54.41 | 14.61 | 60.69 | 16.03 | 52.94 | 17.65 | 63.58 | 21.60 | 53.57 | 19.05 |
| Qwen3.5-35B-A3B | **82.52** | 55.28 | 64.81 | **44.44** | 70.28 | 41.81 | 77.24 | 46.03 | 82.35 | 58.82 | 80.86 | 37.65 | 58.33 | 33.33 |
| LLaMA3.3-70B | 71.54 | **56.50** | 61.11 | 37.65 | 62.47 | **43.32** | 67.76 | **48.10** | 76.47 | **70.59** | 69.75 | **58.02** | 59.52 | **44.05** |
| Qwen3.5-122B-A10B | **82.52** | 55.28 | **70.37** | 41.36 | **71.54** | **43.32** | **78.10** | 44.66 | **82.35** | 58.82 | **82.10** | 51.23 | **66.67** | 42.86 |
| MedGemma-4B | 47.56 | 25.61 | 48.77 | 19.14 | 45.59 | 16.62 | 51.03 | **21.55** | 29.41 | 11.76 | 49.38 | 40.12 | 46.43 | 15.48 |
| MediPhi-Instruct | 47.15 | 18.29 | 46.30 | 16.67 | 49.37 | 17.88 | 51.21 | 13.28 | 47.06 | 29.41 | 45.06 | 28.40 | 45.24 | 14.29 |
| BioMistral-7B | 41.06 | 13.01 | 44.44 | 12.96 | 44.33 | 9.32 | 50.86 | 5.52 | 35.29 | 11.76 | 37.04 | 17.90 | 40.48 | 11.90 |
| BioMistral-7B-DARE | 41.87 | 11.79 | 42.59 | 12.96 | 45.09 | 10.08 | 51.55 | 6.55 | 29.41 | 5.88 | 40.12 | 13.58 | 45.24 | 9.52 |
| MedGemma-27B | **68.29** | **22.76** | 51.85 | 19.14 | 55.42 | 14.61 | 61.38 | 17.93 | 52.94 | 17.65 | **66.67** | 24.69 | **58.33** | 22.62 |
| OpenBioLLM-70B | 60.98 | 20.73 | **54.94** | 19.14 | 54.16 | **20.91** | **65.00** | 17.24 | **76.47** | 17.65 | 62.35 | 22.22 | **58.33** | 13.10 |
| Meditron3-70B | 55.69 | **22.76** | 53.09 | **26.54** | **55.92** | 20.65 | 59.14 | 19.31 | **76.47** | **35.29** | 59.26 | **35.19** | 55.95 | **30.95** |

Table 9: 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 | 55.92 | 8.57 | 40.52 | 9.17 | 35.47 | 6.64 | 44.47 | 7.6 | 50.64 | 4.63 | 54.96 | 20.84 | 39.36 | 16.47 |
| Bielik-11B | 72.52 | **21.0** | 54.14 | **15.93** | 49.32 | **17.43** | 58.88 | **18.67** | 62.47 | **23.14** | 70.42 | **16.3** | 48.82 | **12.58** |
| PLLuM-12B-NC-Inst | 53.0 | 6.54 | 39.17 | 5.63 | 35.06 | 6.05 | 42.69 | 5.04 | 51.16 | 6.17 | 45.8 | 5.8 | 36.4 | 6.17 |
| PLLuM-12B-NC-Chat | 53.78 | 6.84 | 40.01 | 4.45 | 35.22 | 5.91 | 43.09 | 5.01 | 53.98 | 7.2 | 46.64 | 4.12 | 36.06 | 3.63 |
| PLLuM-12B-Inst | 50.74 | 8.98 | 38.44 | 7.32 | 35.76 | 8.57 | 41.9 | 7.8 | 46.79 | 7.71 | 46.39 | 4.29 | 32.85 | 5.07 |
| PLLuM-12B-Chat | 41.7 | 8.98 | 29.83 | 7.71 | 29.42 | 8.19 | 35.0 | 6.73 | 37.53 | 6.43 | 36.39 | 7.56 | 27.2 | 6.42 |
| PLLuM-70B-Inst | **74.54** | 5.59 | **53.74** | 2.87 | **55.23** | 3.88 | **65.7** | 4.62 | **67.1** | 5.4 | **74.62** | 4.96 | **53.97** | 1.77 |
| LLaMA3-8B | 44.68 | 22.84 | 35.62 | 20.37 | 37.07 | 23.86 | 42.72 | 20.3 | 46.02 | 17.74 | 62.1 | 24.79 | 43.58 | 22.04 |
| Qwen3.5-9B | 76.44 | 37.66 | 55.77 | 24.48 | 56.91 | 29.15 | 69.53 | 31.65 | 73.52 | 42.16 | 77.9 | 29.75 | 57.69 | 19.68 |
| Gemma-3-12B | 66.69 | 9.58 | 49.86 | 5.46 | 47.56 | 7.48 | 58.8 | 7.07 | 59.9 | 9.51 | 70.0 | 11.76 | 48.48 | 7.94 |
| Gemma-3-27B | 74.54 | 9.88 | 54.98 | 8.05 | 54.69 | 9.63 | 65.0 | 8.22 | 63.24 | 9.77 | 74.62 | 15.71 | 52.36 | 11.23 |
| Qwen3.5-35B-A3B | 85.78 | 48.13 | 65.84 | 25.21 | 69.96 | 34.3 | 80.12 | 39.45 | 78.92 | 40.36 | 85.55 | 44.71 | 67.31 | 24.24 |
| LLaMA3.3-70B | 78.41 | 12.61 | 60.1 | 6.53 | 60.9 | 8.7 | 71.56 | 9.49 | 73.01 | 11.83 | 82.61 | 14.12 | 59.04 | 6.84 |
| Qwen3.5-122B | **88.28** | **52.77** | **72.03** | **33.71** | **74.24** | **39.13** | **84.46** | **47.79** | **80.21** | **50.9** | **88.99** | **56.97** | **70.19** | **40.71** |
| MedGemma-4B | 50.62 | 5.41 | 39.9 | 3.83 | 38.12 | 4.18 | 45.03 | 3.83 | 49.61 | 4.37 | 58.91 | 5.21 | 40.12 | 3.97 |
| MediPhi-Instruct | 33.49 | **61.04** | 30.44 | **60.66** | 30.21 | **58.57** | 36.24 | **57.59** | 41.65 | **60.67** | 63.45 | 20.08 | 42.82 | 16.05 |
| BioMistral-7B | 33.55 | 3.81 | 27.86 | 2.98 | 28.63 | 3.25 | 33.6 | 3.86 | 38.05 | 3.08 | 46.64 | 1.09 | 34.21 | 1.35 |
| BioMistral-7B-DARE | 34.09 | 6.6 | 27.29 | 6.36 | 27.68 | 6.24 | 32.7 | 5.83 | 31.88 | 4.11 | 49.33 | 2.69 | 35.05 | 2.45 |
| MedGemma-27B | 75.67 | 10.23 | 53.97 | 6.02 | 54.04 | 7.29 | **63.93** | 8.14 | 67.61 | 12.34 | 76.39 | 14.45 | 56.84 | 8.78 |
| OpenBioLLM-70B | 70.14 | 25.52 | 51.27 | 21.05 | **54.07** | 22.26 | 62.29 | 21.85 | 65.55 | 25.71 | **76.81** | **31.43** | 55.66 | **25.42** |
| Meditron3-70B | **71.56** | 33.19 | **55.94** | 18.4 | 52.96 | 26.03 | 63.48 | 25.96 | **67.87** | 24.16 | 75.88 | 10.34 | **57.18** | 5.49 |

Table 10: Model evaluation results. OM = Our evaluation methodology used on original questions, AS = *Correct answer substitution* results on the same set of questions as OM

| 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 | 47.97 | 43.53 | 37.36 | 33.35 | 32.15 | 30.58 | 31.21 | 34.20 | 23.92 | 27.31 | 52.68 | 50.52 | 39.55 | 37.98 |
| Bielik-11B | **70.69** | **71.19** | **52.27** | **51.62** | 48.12 | **47.85** | 54.39 | **53.30** | 48.77 | **48.65** | 69.91 | **67.06** | 46.68 | **45.34** |
| PLLuM-12B-NC-Inst | 48.58 | 48.39 | 36.04 | 35.49 | 32.91 | 32.49 | 37.30 | 35.84 | 31.23 | 32.22 | 42.81 | 39.57 | 32.04 | 31.23 |
| PLLuM-12B-NC-Chat | 49.30 | 48.33 | 37.04 | 35.18 | 32.94 | 32.75 | 37.59 | 36.29 | 32.92 | 31.40 | 43.80 | 43.86 | 32.95 | 33.48 |
| PLLuM-12B-Inst | 46.83 | 36.06 | 34.99 | 27.78 | 32.65 | 25.14 | 35.20 | 27.83 | 30.58 | 22.92 | 40.75 | 28.72 | 30.16 | 24.22 |
| PLLuM-12B-Chat | 38.90 | 21.21 | 29.16 | 17.77 | 27.49 | 14.96 | 30.66 | 14.88 | 23.92 | 12.92 | 31.32 | 24.72 | 25.22 | 19.09 |
| PLLuM-70B-Inst | 70.39 | 45.64 | 51.46 | 32.60 | **50.43** | 31.12 | **57.67** | 35.78 | **47.84** | 27.72 | **71.56** | 60.96 | **50.38** | 41.83 |
| LLaMA3-8B | 44.31 | 45.90 | 36.77 | 38.55 | 35.93 | 37.41 | 40.82 | 42.58 | 31.99 | 34.09 | 60.74 | 61.22 | 42.30 | 43.21 |
| Qwen3.5-9B | 75.85 | 74.37 | 55.42 | 52.82 | 55.80 | 53.63 | 64.30 | 62.59 | 54.80 | 52.98 | 78.48 | 2.44 | 56.85 | 1.10 |
| Gemma-3-12B | 64.21 | 53.83 | 48.29 | 37.98 | 45.90 | 36.16 | 53.08 | 42.30 | 46.20 | 37.66 | 68.49 | 61.82 | 47.25 | 41.55 |
| Gemma-3-27B | 72.00 | 67.75 | 53.78 | 48.67 | 52.23 | 48.00 | 59.75 | 56.81 | 50.47 | 50.64 | 74.07 | 70.10 | 52.07 | 49.00 |
| Qwen3.5-35B-A3B | 84.67 | 84.09 | 65.45 | 63.23 | 68.73 | 66.36 | 75.06 | 72.58 | 60.23 | 58.77 | 85.12 | 85.62 | 66.21 | 66.05 |
| LLaMA3.3-70B | 78.21 | 79.02 | 59.71 | 59.71 | 59.98 | 59.70 | 66.35 | 67.29 | 56.43 | 55.56 | 81.21 | 82.23 | 58.82 | 61.23 |
| Qwen3.5-122B-A10B | **87.06** | **87.71** | **71.00** | **71.15** | **72.19** | **72.44** | **77.45** | **78.69** | **61.64** | **62.98** | **87.46** | **87.78** | **68.65** | **69.68** |
| MedGemma-4B | 49.66 | 43.76 | 39.28 | 34.19 | 37.86 | 32.78 | 42.45 | 36.41 | 36.08 | 29.53 | 58.24 | 53.09 | 40.05 | 36.51 |
| MediPhi-Instruct | 33.57 | 26.73 | 31.99 | 26.27 | 29.87 | 24.61 | 34.27 | 26.62 | 26.32 | 20.64 | 60.87 | 55.70 | 42.08 | 38.11 |
| BioMistral-7B | 32.17 | 19.63 | 28.89 | 19.51 | 27.35 | 18.52 | 29.83 | 19.16 | 21.64 | 12.57 | 42.24 | 32.53 | 32.45 | 25.59 |
| BioMistral-7B-DARE | 31.68 | 23.67 | 28.62 | 23.09 | 26.33 | 21.47 | 30.05 | 22.78 | 20.00 | 17.08 | 45.10 | 34.72 | 33.45 | 27.53 |
| MedGemma-27B | **73.54** | **70.39** | **53.07** | **49.70** | **52.47** | **49.77** | **58.92** | **55.57** | **51.35** | **50.88** | **76.39** | 73.53 | **54.29** | 48.50 |
| OpenBioLLM-70B | 67.45 | 59.24 | 49.51 | 41.76 | 50.62 | 40.84 | 57.43 | 47.56 | 48.83 | 35.03 | 73.85 | **74.67** | 52.97 | **52.41** |
| Meditron3-70B | 65.69 | 25.53 | 49.82 | 32.26 | 48.44 | 21.55 | 54.44 | 19.42 | 47.08 | 15.15 | 71.88 | 2.60 | 53.07 | 1.22 |

Table 11: Model evaluation results expressed as percentages. *OM* = Our evaluation Methodology, *PM* = Previous evaluation Methodology.

| Model | Original vs modified avg difference (%) | | | | | | |
|---|---|---|---|---|---|---|---|
| | LEK | LDEK | PES | PESDL | PESF | LEK-ENG | LDEK-ENG |
| Bielik-4.5B | -32.56 | -23.64 | -21.73 | **-17.79** | **-12.07** | -30.90 | -24.39 |
| Bielik-11B | -46.13 | -35.65 | -30.32 | -37.77 | -40.33 | -47.51 | -33.77 |
| PLLuM-12B-NC-Inst | -34.02 | -24.61 | -22.20 | -26.22 | -17.28 | -32.21 | -23.67 |
| PLLuM-12B-NC-Chat | -35.77 | -27.67 | -23.18 | -25.95 | -20.71 | -35.09 | -25.70 |
| PLLuM-12B-Inst | -32.48 | -25.26 | -22.64 | -26.49 | -24.04 | -32.91 | -22.92 |
| PLLuM-12B-Chat | **-27.81** | **-20.85** | **-20.26** | -24.37 | -23.00 | **-22.55** | **-17.92** |
| PLLuM-70B-Inst | -51.57 | -38.73 | -37.37 | -42.89 | -35.90 | -52.47 | -38.04 |
| LLaMA3-8B | **-21.65** | **-17.42** | **-16.23** | **-23.24** | **-22.69** | -33.08 | **-20.75** |
| Qwen3.5-9B | -46.76 | -36.15 | -36.05 | -45.91 | -47.87 | -47.39 | -37.37 |
| Gemma-3-12B | -54.30 | -41.49 | -38.53 | -46.99 | -53.49 | -47.94 | -34.14 |
| Gemma-3-27B | -50.81 | -38.01 | -35.15 | -41.84 | -31.33 | -49.16 | -33.97 |
| Qwen3.5-35B | -32.23 | -34.53 | -32.08 | -35.94 | -27.88 | -37.55 | -37.62 |
| LLaMA3.3-70B | -49.35 | -42.28 | -40.19 | -43.17 | -38.91 | -54.04 | -41.59 |
| Qwen3.5-122B | -30.83 | -34.65 | -31.51 | -33.53 | -24.66 | **-29.31** | -27.67 |
| MedGemma-4B | -37.32 | -29.96 | -28.64 | -31.89 | -33.82 | -42.33 | -29.82 |
| MediPhi-Instruct | **9.03** | **9.11** | **6.40** | **-0.95** | **-4.58** | -35.33 | -24.98 |
| BioMistral-7B | -25.62 | -21.08 | -21.31 | -25.10 | -19.87 | **-34.17** | **-24.56** |
| BioMistral-7B-DARE | -21.99 | -17.25 | -17.40 | -22.93 | -13.53 | -35.17 | -25.33 |
| MedGemma-27B | -55.69 | -41.83 | -39.78 | -44.56 | -40.59 | -58.73 | -45.29 |
| OpenBioLLM-70B | -36.29 | -26.09 | -27.38 | -33.40 | -27.68 | -38.50 | -25.93 |
| Meditron3-70B | -29.25 | -27.03 | -21.25 | -26.95 | -22.79 | -47.87 | -37.15 |

Table 12: 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).

## E Instructions

| |
|---|
| 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 |

Table 13: Instruction for inferring the answer used in **Previous Methodology (PM)**, with a sample answer.

| |
|---|
| 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. |

Table 14: Instruction for inferring the answer used in **Our Metodology (OM)**, with a sample answer.

| |
|---|
| 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 15: 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 16: 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 17: 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 18: Instruction used in **LLM-as-a-judge**.

## F Question Modification Examples

### F.1 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**

### F.2 Multiple Answers (MA)

**Original question**

Which of the following activities are characteristic of six-month-old infants?

A. supporting their body on 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 and B are correct
E. answers A, B, and 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 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**

### F.3 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 forming the posterior wall of the inguinal canal, bordered superiorly by the conjoint tendon, inferiorly by the inguinal ligament, and laterally by the 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 forming the posterior wall of the inguinal canal, bordered superiorly by the conjoint tendon, inferiorly by the inguinal ligament, and laterally by the 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**

### F.4 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. anaemia, 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: anaemia, uterine myomas**

## G Contamination Studies

### G.1 Web Search Audit of Question and Answer Availability

Polish examination questions are publicly available, but access is restricted by a CAPTCHA system. To check whether they may be included in LLM training corpora, we audited whether they could be found through a web search. For this experiment, we used the 2023 autumn editions of LEK and PES and queried Tavily[6] for each question without the answer. We then compared the best retrieved text fragment with the original question using normalized Levenshtein distance.

Figure 4 shows the distribution of normalized character-level Levenshtein distances for LEK and PES. The distributions differ substantially. LEK questions were found much more frequently in near-exact or highly similar form on the web, while PES matches were generally more distant. The median distance was 0.52 for LEK and 0.77 for PES.

We then performed a second, stricter analysis for LEK. Instead of relying on the short search preview, we retrieved the full content of the returned page and asked `gpt-5-mini` to judge whether the page contained the same question and whether it also contained the correct answer. We treated a case as evidence of potential answer leakage only when both conditions were satisfied.

For LEK, `gpt-5-mini` judged 169 of 194 returned pages to contain the same or a materially equivalent question. However, most of these pages did not expose the answer. This pattern was consistent with many results pointing to paid LEK learning portals, where the question text is visible but the answer is hidden behind access control. In total, 36 of 194 pages contained the correct answer. Some of these answer-containing cases came from freely available flashcard-style learning platforms.

For PES, we did not find evidence of data leakage in which both the same question and its corresponding answer were available together in the top returned web result. This is consistent with the character of the exam: PES is a highly specialized board-certification examination, less commonly targeted by public learning materials. By contrast, LEK is taken by every candidate seeking the right to practice medicine in Poland, and the larger ecosystem of preparation materials makes repeated or highly similar questions easier to find online. The observed pattern is also consistent with candidates' perception of the exams: LEK preparation materials are widely available, whereas PES materials are substantially scarcer. Some LEK leakage likely reflects the fact that questions repeat, or appear in very similar form, across consecutive years.

[6] https://www.tavily.com/

### G.2 Temporal Performance Analysis

Additionally, to check whether the LLMs perform better on older exams (which would prove data leakage), we conducted a temporal contamination analysis using the English versions of LEK and LDEK. We focused on the 2024 spring editions, whose questions were released after the training cutoff of the majority of the evaluated models and, therefore, could not have been directly memorized during pre-training.

If contamination were the primary explanation for high benchmark performance, one would expect a systematic degradation across recent editions relative to earlier years, depending on the model's knowledge cutoff date. However, such a pattern was not consistently observed. In several cases, performance on the 2025 data was comparable to the average performance on examinations from 2020–2024, whereas performance in 2026 was not a significant outlier. No visible deterioration in performance over recent years is observed in the LEK ENG and LDEK ENG exams, as presented in Figures 5 and 6.

### G.3 DCQ Visualisation

The visualization of the results of the DCQ contamination estimation is presented in Figure 3.

## H Regular Expressions for Question Filtering

We use regular expressions to (i) identify parts of questions that require removal or modification and (ii) filter questions eligible for a given modification group (e.g., MA-group).

For the MA-group, we apply the following language-specific patterns:

**Polish:**

```
pattern = r"prawdziwe sa odpowiedzi"
filtered = [p if re.search(pattern, p, re.
    IGNORECASE) else "" for p in parts]
pattern = r"wszystkie\s+.*?\s+(prawidlowe|
    prawdziwe)"
filtered2 = [p for p in parts if re.search(
    pattern, p, re.IGNORECASE)]
```

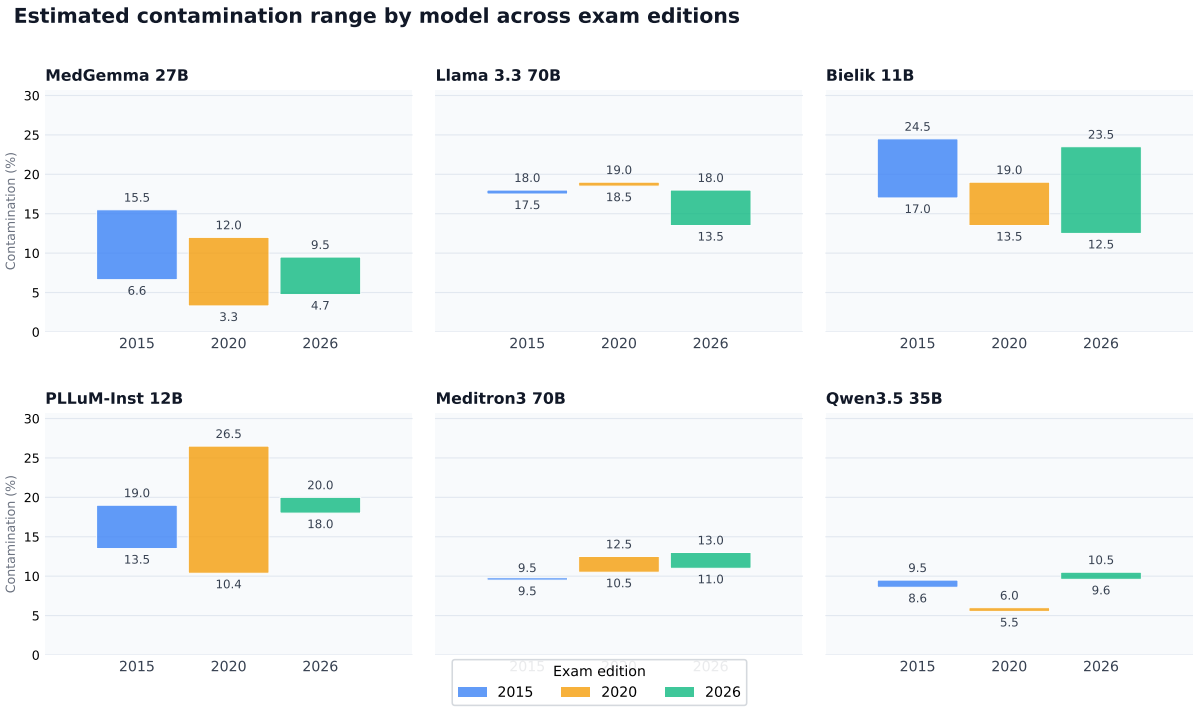



Figure 3: Estimated contamination in LLMs from different groups and model families, across three spring editions 2015, 2020, 2026 of combined LEK and LDEK exams.

**English:**

```
pattern = r"(?:are\s+(?:correct|true)|true\s+(?:
    answers|statements)|correct\s+answers?|
    answers|\&)"
filtered = [p if re.search(pattern, p, re.
    IGNORECASE) else "" for p in parts]
```

# I Detailed Summary of Disagreement Analysis

After calculating agreement metrics for human annotators and AI judges in the evaluation of open-ended answers, we performed an additional analysis of cases in which their assessments diverged. The aim was to identify recurring sources of disagreement, define them as reproducible discrepancy patterns, and examine how these patterns were treated by human annotators and LLM-as-a-judge systems.

## I.1 Analytical Process

The analysis consisted of three stages: preliminary review of evaluated cases, supplementary annotation of discrepancy patterns, and pattern-level analysis of evaluator behaviour.

### I.1.1 Preliminary case review

A senior annotator who was also a physician reviewed 15 control cases completed by all six human annotators and 74 disagreement cases in which a single human annotator's decision contradicted the unanimous assessment of seven AI judges (gpt-5.4-nano, gpt-5.4-mini, gpt-5.5, claude-haiku-4-5, claude-sonnet-4-6, claude-opus-4-7, gpt-oss-120b). For each case, the senior annotator reviewed all materials available to, and produced by, the initial annotator, including the original question, the official correct answer, the AI-generated answer, the human annotator's decision, and, when available, their written commentary.

The source of disagreement was not always identifiable, and in some cases appeared unique to a single answer. However, multiple outputs shared recurring features, and their relevance was further supported by the available initial annotators' comments.

The main outcome of this preliminary review was the definition of six **discrepancy patterns**: recurring features of evaluated answers that were repeatedly suspected of contributing to disagreement between human and AI evaluators, and in some cases also to variability among human annotators. These patterns form a practical framework for describing the main sources of disagreement in the analysed sample, rather than a complete taxonomy of answer defects.

The discrepancy patterns used in the supplementary annotation were defined as follows:

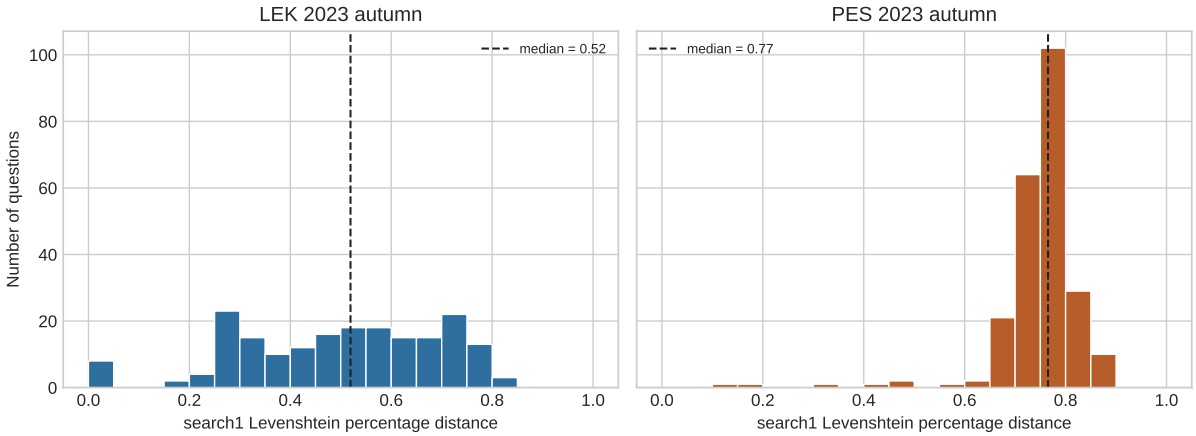



Figure 4: Distribution of normalized Levenshtein distances between the original question and the closest text fragment retrieved by Tavily for LEK and PES. Lower values indicate closer textual matches.

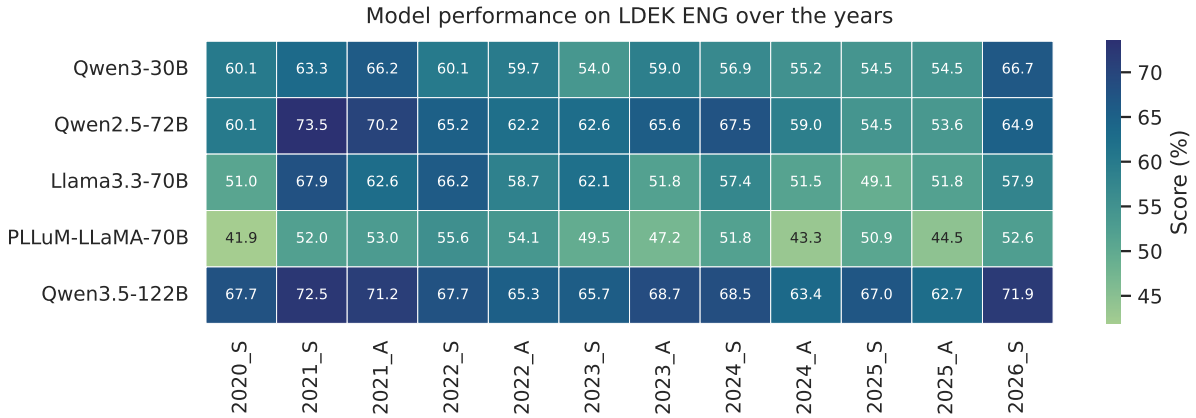



Figure 5: Model performance on LDEK in English over the editions Spring-**S** Autumn-**A** over the years

- **A – Defects in redundant content:** The output addressed the issue indicated by the official correct answer but also included additional content that introduced at least one decision-relevant defect.

- **B – Valid alternative interpretation of the question:** The output answered a valid interpretation of the open-ended question, but not the issue implied by the original closed-ended question and its official correct answer.

- **C – Underspecified expected detail:** The output addressed the expected issue but did so with lower specificity or completeness than the official correct answer.

- **D – Non-existent medical terms:** The output contained a term absent from Polish, English, or Latin medical nomenclature, including invalid hybrid terms.

- **E – Misuse of existing medical terms:** The output used an existing Polish, English, or Latin medical term inconsistently with its accepted meaning.

- **F – Incorrect spelling or notation of medical terms:** The output contained an existing Polish, English, or Latin medical term with incorrect spelling, notation, inflection, abbreviation, or written form.

The detailed criteria used for applying these labels are presented in Table 20.

### I.1.2 Supplementary annotation

Of the 74 discrepant cases, 3 were excluded because the official correct answer was inconsistent with current medical knowledge. The senior annotator then applied the six discrepancy patterns to the remaining 71 cases to describe potential sources of disagreement.

| Discrepancy pattern | Human evaluator tendency | AI evaluator tendency |
|---|---|---|
| A | Penalised errors appearing in both necessary and redundant content. | Accepted answers if expected information was present, disregarding errors in redundant content. |
| B | Accepted valid alternative interpretations if they contained no significant errors. | Usually rejected answers that missed information specified in answer key. |
| C | No consistent tendency identified. | No consistent tendency identified. |
| D | Usually penalised non-existent medical terms. | Usually did not penalise non-existent medical terms. |
| E | Too few cases to infer a dominant tendency. | Too few cases to infer a dominant tendency. |
| F | Usually penalised incorrect spelling or notation. | Usually did not penalise incorrect spelling or notation. |

Table 19: Observed tendencies in evaluator approach to discrepancy patterns. See Table 3 for code (A-F) definitions.

| | Discrepancy pattern | Criteria |
|---|---|---|
| A | Defects in redundant content | 1) Output engages with the issue indicated by the official correct answer. *and* 2) Output includes additional content beyond that issue. *and* 3) The additional content features $\geq 1$ defect that could influence the evaluator's decision. |
| B | Valid alternative interpretation of the question | 1) Output addresses a valid interpretation of the open-ended question. *and* 2) This interpretation differs from the issue expressed by the official correct answer. |
| C | Underspecified expected detail | 1) Output addresses the expected issue expressed by the official correct answer. *and* 2) Output addresses this issue with lower specificity or completeness than the official correct answer. |
| D | Non-existent medical terms | 1) Output contains a term absent from Polish, English, or Latin medical nomenclature. *or* 2) Output contains a non-standard hybrid medical term defined as an expression created by combining Polish, English, or Latin elements into a form that is not itself recognised in medical nomenclature. |
| E | Misuse of existing medical terms | 1) Output contains an existing Polish, English or Latin medical term. *and* 2) Output uses this term inconsistently with its accepted meaning. |
| F | Incorrect spelling or notation of medical terms | 1) Output contains an existing Polish, English or Latin medical term. *and* 2) Spelling, notation, inflection or abbreviation of this term is incorrect. |

Table 20: Criteria used for applying discrepancy patterns to LLM answers.

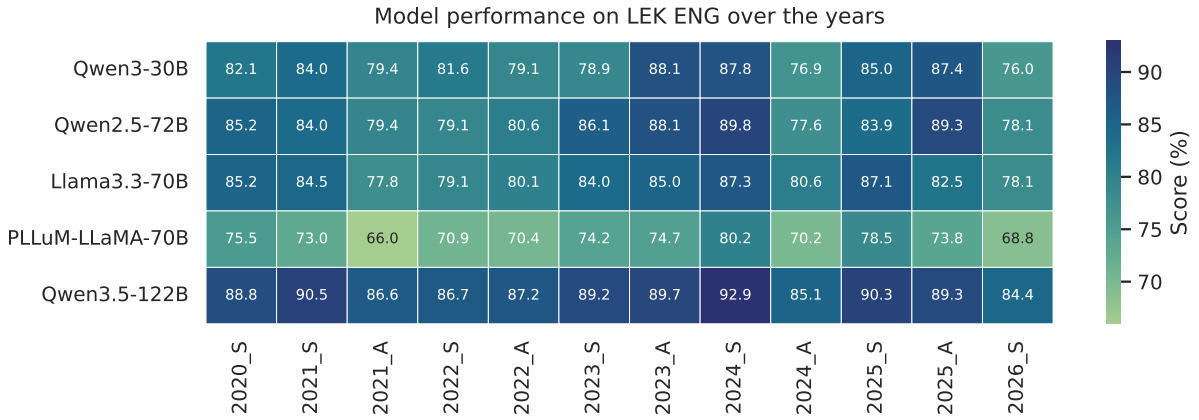



Figure 6: Model performance on LEK in English over the editions Spring-**S** Autumn-**A** over the years

Each answer could receive all applicable labels, provided that their criteria were compatible. Patterns A, B, and C were mutually exclusive, as they described different relations between the output and the issue implied by the original closed-ended question and its official correct answer. Patterns D–F could co-occur with any compatible A–C pattern and with one another; for example, an answer containing a non-existent medical term in an unnecessary explanation could be labelled as both D and A.

#### I.1.3 Pattern-level analysis of evaluator behaviour

After supplementary annotation, we analysed the prevalence of each discrepancy pattern and the distribution of positive and negative human evaluations among answers assigned that pattern. These results served as supporting evidence for the potential influence of each characteristic on human decisions.

For reference, we also prepared an outlier-excluded summary. The excluded annotator's 28 annotations had been rejected because their performance on control questions differed from that of the remaining human annotators. This summary served as a supplementary sensitivity analysis, assessing whether the observed pattern distribution and evaluation outcomes remained broadly similar after removing cases linked to atypical control-case performance.

Tables 21 and 22 present the prevalence of each discrepancy pattern and the distribution of positive and negative human assessments among answers assigned that pattern, with Table 21 including all assessments and Table 22 excluding assessments made by the outlier annotator.

**Limitations** These data should be interpreted cautiously. A single AI-generated answer could contain more than one discrepancy pattern or other defects that were not separately coded. Therefore, the final annotator decision could not be attributed to one pattern alone. Some answers were counted in more than one group. For example, an answer containing a non-existent term in redundant content was counted under both A and D. Similarly, an answer containing a real medical term that was both misused and misspelled was counted under both E and F.

Conclusions about the dominant approaches of human and AI evaluators to these patterns were based not only on pattern frequency and evaluation outcomes, but also on the initial annotators' comments and the senior annotator's qualitative case-by-case observations.

### I.2 Findings

#### I.2.1 Pattern-specific findings

Below we present our findings regarding each discrepancy pattern. Where relevant, we also indicate whether a given pattern appeared to reflect limitations of the original closed-to-open question conversion, the behaviour of the tested LLMs, or differences in how human and AI evaluators approached cases characterised by these features. All data reported in this subsection refer to the analysis, including cases annotated by the outlier annotator

**A – Defects in redundant content** This pattern was identified in 14 AI-generated answers. In two cases, the only defect was an incorrect expression used alongside a correct term as if it were an equivalent. In the remaining cases, defects appeared in unnecessary explanations, elaborations, or specifi-

| | Discrepancy pattern | Answers assigned pattern, n (%) | Positive human assessments, n (%) | Negative human assessments, n (%) |
|---|---|---|---|---|
| A | Defects in redundant content | 14 (20%) | 0 (0%) | 14 (100%) |
| B | Valid alternative interpretation of the question | 20 (28%) | 19 (95%) | 1 (5%) |
| C | Underspecified expected detail | 12 (17%) | 8 (67%) | 4 (33%) |
| D | Non-existent medical terms | 14 (20%) | 2 (14%) | 12 (86%) |
| E | Misuse of existing medical terms | 4 (6%) | 1 (25%) | 3 (75%) |
| F | Incorrect spelling or notation of medical terms | 10 (14%) | 1 (10%) | 9 (90%) |

Table 21: Prevalence of discrepancy patterns and distribution of human assessments among answers assigned each pattern, including assessments made by the outlier annotator

*Note.* Percentages for pattern prevalence use all 71 included answers as the denominator; percentages for positive and negative assessments use answers assigned the given pattern. Patterns were not mutually exclusive.

| | Discrepancy pattern | Answers assigned pattern, n (%) | Positive human assessments, n (%) | Negative human assessments, n (%) |
|---|---|---|---|---|
| A | Defects in redundant content | 14 (33%) | 0 (0%) | 14 (100%) |
| B | Valid alternative interpretation of the question | 7 (16%) | 7 (100%) | 0 (0%) |
| C | Underspecified expected detail | 7 (16%) | 3 (43%) | 4 (57%) |
| D | Non-existent medical terms | 11 (26%) | 1 (9%) | 10 (91%) |
| E | Misuse of existing medical terms | 2 (5%) | 0 (0%) | 2 (100%) |
| F | Incorrect spelling or notation of medical terms | 7 (16%) | 0 (0%) | 7 (100%) |

Table 22: Prevalence of discrepancy patterns and distribution of human assessments among answers assigned each pattern, excluding assessments made by the outlier annotator

*Note.* Percentages for pattern prevalence use all 43 included answers as the denominator; percentages for positive and negative assessments use answers assigned the given pattern. Patterns were not mutually exclusive.

cations.

We interpret this pattern mainly as a consequence of the tested LLMs' tendency to expand answers beyond the issues explicitly requested in the question. However, none of the relevant questions explicitly prohibited such expansion.

Human annotators tended to reject answers containing errors they considered important, regardless of whether the error appeared in the core part of the answer or in redundant content. By contrast, LLM-as-a-judge systems tended to accept answers if they contained the expected information, even when other parts of the answer included significant defects.

**B – Valid alternative interpretation of the question** This pattern was identified in 20 AI-generated answers. In most cases, it appeared to result from loss of context during conversion of closed-ended questions into open-ended form. Without the original answer options, the question could often be answered in more than one reasonable way.

Only one case could be interpreted as a genuine misunderstanding of the question. In the remaining cases, the LLM answer was valid for the open-ended wording of the question but did not match the interpretation implied by the original answer key. This included three questions asking the respondent to identify an incorrect statement. With the answer options removed, the LLMs generated their own examples of false statements.

Human annotators tended to accept answers if they were appropriate to the wording of the open-ended question and did not contain important errors. LLM-as-a-judge systems tended to reject outputs when they did not match the answer key.

**C – Underspecified expected detail** This pattern was identified in 12 answers. In each case, the question did not explicitly define the expected level of detail, such as the level of biological organisation or number of elements required.

We interpret this pattern as another consequence of context loss during conversion from closed-ended to open-ended questions. In the original format, the answer options expressed a predefined level of specificity. After conversion, this information was no longer available.

Evaluator approaches to this pattern were heterogeneous. Some less specific answers were accepted, whereas others were rejected. Variation was also observed between individual human annotators. No clear dominant approach could be identified among either human annotators or LLM-as-a-judge systems.

**D – Non-existent medical terms** This pattern was identified in 14 answers. Human annotators

rejected most answers affected by this pattern, and often treated such defects as important because they directly concerned medical nomenclature and could indicate lack of terminological reliability. In contrast, LLM-as-a-judge systems usually did not appear to penalise such defects, particularly when the intended meaning could be inferred or when the answer otherwise matched the expected content.

**E – Misuse of existing medical terms** This pattern was identified in 4 answers. Because the number of cases was very small, the data were considered insufficient to infer a dominant evaluation tendency among human annotators or AI judges.

This pattern nevertheless remained conceptually important because it captured a distinct type of terminological defect: the use of real medical terms in a meaning inconsistent with accepted medical usage.

**F – Incorrect spelling or notation of medical terms** This pattern was identified in 10 answers. It was usually associated with negative human evaluations and was often explicitly mentioned in annotator comments as a reason for rejection.

LLM-as-a-judge systems did not appear to penalise this pattern in most cases. They often accepted answers despite incorrect spelling or notation of medical terms, especially when the intended term was recognisable and the broader answer overlapped with the expected content.

#### I.2.2 Effects of Outlier Exclusion on Pattern Distribution

The outlier-excluded summary generally supported the main interpretation of evaluator behaviour, although it also showed that some estimates of pattern prevalence were sensitive to the excluded annotator's cases. In both summaries, defects in redundant content, non-existent medical terms, and incorrect spelling or notation were predominantly associated with negative human assessments, whereas valid alternative interpretations of the question were predominantly accepted. The largest change concerned the prevalence of valid alternative interpretations, which decreased after excluding the outlier annotator, indicating that this pattern was overrepresented among that annotator's cases. The mixed distribution of assessments for underspecified expected detail further supported the absence of a consistent evaluation approach for this pattern.

### I.3 Interpretation and Implications

Overall, the disagreement analysis indicates that discrepancies between human annotators and AI judges were driven mainly by systematic differences in how they handled context loss, redundant content, and terminology-related defects, while some variability was also observed among human annotators. Human annotators were generally more sensitive to clinically or linguistically important defects, whereas LLM-as-a-judge systems tended to prioritise overlap with the answer key and often overlooked errors outside the expected content.

These findings suggest that at least some of the observed differences may be mitigable through more explicit evaluation instructions for both human annotators and LLM-as-a-judge systems, specifically targeting the recurring sources of discrepancy identified in this analysis.