npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-026-02465-0

A large-scale benchmark for evaluating large language models on medical question answering in Romanian

Check for updates

Ana-Cristina Rogoz[1,6], Radu Tudor Ionescu[1,6] ✉, Alexandra-Valentina Anghel[2], Ionuț-Lucian Antone-Iordache[2,3], Simona Coniac[2,4] & Andreea Iuliana Ionescu[2,3,5]

We introduce MedQARo, the first large-scale medical QA benchmark in Romanian, alongside a comprehensive evaluation of state-of-the-art large language models (LLMs). We construct a high-quality and large-scale dataset comprising 105,880 QA pairs about cancer patients from two medical centers. The questions regard medical case summaries of 1242 patients, requiring both keyword extraction and reasoning. Our benchmark contains both in-domain and cross-domain (cross-center and cross-cancer) test collections, enabling a precise assessment of generalization capabilities. We experiment with four open-source LLMs from distinct families of models on MedQARo. Each model is employed in two scenarios: zero-shot prompting and supervised fine-tuning. We also evaluate two state-of-the-art LLMs exposed only through APIs, namely GPT-5.2 and Gemini 3 Flash. Our results show that fine-tuned models significantly outperform zero-shot models, indicating that pretrained models fail to generalize on MedQARo. Our findings demonstrate the importance of both domain-specific and language-specific fine-tuning for reliable clinical QA in Romanian.

The development of robust question answering (QA) systems is an important goal in natural language processing (NLP), representing one of the prerequisites for building systems that reach human-level intelligence. Despite the significant advances in the area of Large Language Models (LLMs)[1–5], which are able to perform remarkably well across various QA tasks[6–12], QA remains a challenging task within specialized domains, such as the medical field, or in low-resource languages, such as Romanian. For instance, the task of medical QA comes with unique challenges due to the complexity of clinical narratives and the necessity for deep understanding of medical terms. Moreover, there is a soaring demand for high accuracy, as mistakes can have serious consequences in daily medical practice assisted by AI agents.

Question answering emerged as a fundamental task in NLP, with several datasets being developed to support research across various domains and languages. Popular QA datasets focusing on the English language, such as SQuAD[8] and Natural Questions[13], established the foundation for extractive and generative QA systems. These datasets showed the importance of large-scale high-quality annotations for training robust QA models. Later on, given the need for cross-language understanding, more and more datasets covering several languages started to arise. Some notable multilingual efforts include XQuAD[14], which extends SQuAD to 11 languages, and MLQA[15], which provides QA datasets in seven languages with cross-lingual evaluation capabilities. Furthermore, TyDi QA[16] advanced multilingual QA by covering structurally different languages and focusing on information-seeking questions. However, these efforts did not include Romanian among the target languages.

Medical QA represents a challenging domain that requires deep understanding of a broad range of aspects, such as clinical terminology and medical reasoning. Several English medical QA datasets have been already developed to tackle the particularities of the medical domain. For instance, MedQA[17] provides over 12,000 multiple-choice questions from medical licensing exams across three countries, testing complex clinical reasoning abilities. Additionally, the emrQA dataset[18] introduces one of the largest QA medical resources with over 455,000 question-answer pairs automatically generated from electronic medical records. PubMedQA[19] focuses on biomedical literature comprehension, with questions derived from research article titles and abstracts. MedMCQA[20] expands medical QA to include Indian medical entrance exams, with 193,100 questions across 21 medical subjects, demonstrating the importance of diverse medical knowledge sources. Despite its data source (Indian medical exams), the QA pairs contained by MedMCQA are in English.

[1]University of Bucharest, Bucharest, Romania. [2]Carol Davila University of Medicine and Pharmacy, Bucharest, Romania. [3]Colțea Clinical Hospital, Bucharest, Romania. [4]Hospice Hope Bucharest, Bucharest, Romania. [5]Dan Furtună Medical Center, Bucharest, Romania. [6]These authors contributed equally: Ana-Cristina Rogoz, Radu Tudor Ionescu. ✉e-mail: raducu.ionescu@gmail.com

Table 1 | Comparison between MedQARo and existing QA datasets across three dimensions, namely language coverage, domain focus, and dataset size

Dataset	Language	Domain	# QA pairs
SQuAD[8]	English	General knowledge	107,785
Natural Questions[13]	English	General knowledge	307,373
XQuAD[14]	Multilingual (11 languages)	General knowledge	1190
MLQA[15]	Multilingual (7 languages)	General knowledge	17,000
TyDi QA[16]	Multilingual (11 languages)	Information-seeking	204,000
PubMedQA[19]	English	Medical	212,300
emrQA[18]	English	Clinical cases	455,837
MedQA[17]	English	Medical exams	12,723
MedMCQA[20]	English	Medical exams (India)	193,100
JuRo[21]	Romanian	Legal	10,836
RoITD[22]	Romanian	Technical	9500
LiRo / XQuAD-Ro[23]	Romanian	General knowledge	1190
RoMedQA[24]	Romanian	Biology exams	4127
MedQARo (ours)	**Romanian**	**Clinical cases**	**105,880**

Existing QA datasets range from general-purpose English resources to multilingual datasets and domain-specific collections in areas such as legal, technical and medical fields. MedQARo is the first large-scale benchmark for medical QA in Romanian.

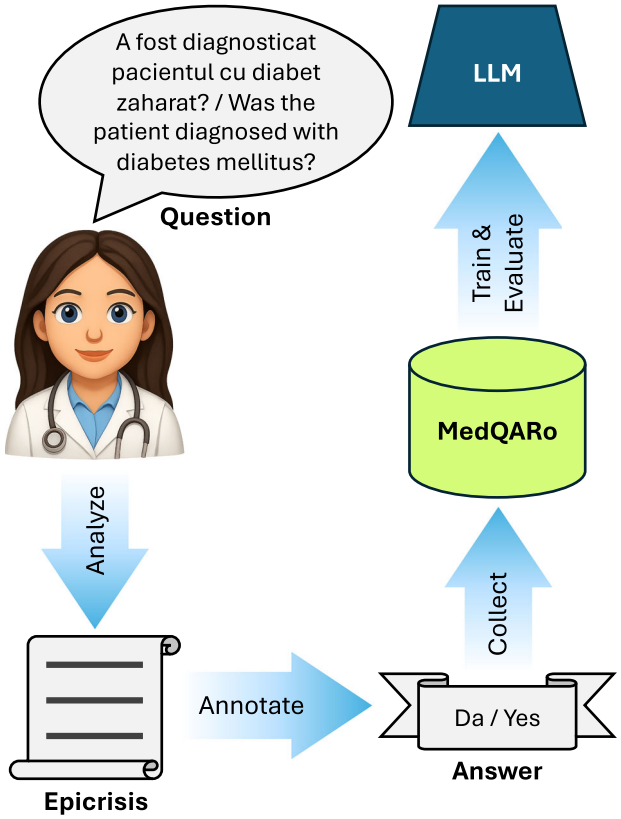

Fig. 1 | An image illustrating our dataset creation and model benchmarking stages. In the annotation stage, physicians analyze epicrises in order to find the correct answer for a given question. Upon collecting question and answer QA pairs, we benchmark zero-shot and fine-tuned LLMs on the medical question answering task, using two evaluation scenarios, namely in-domain and cross-domain.

While large-scale datasets such as PubMedQA[19], emrQA[18], and MedQA[17] have driven significant progress in English medical QA, the availability of similar resources in other languages, especially under-represented ones, remains limited. This hinders the development of LLMs that are robust to language and domain shift. To address this gap, we introduce MedQARo, the first large-scale benchmark for **Med**ical **Q**uestion **A**nswering in **Ro**manian. MedQARo contains original clinically-grounded content rather than translations, addressing a critical gap in current literature.

For the Romanian language, a few QA datasets have been independently developed. Two such datasets focus on domain specific understanding. One is the JuRo dataset[21], which provides a collection of Romanian juridical texts with associated questions, and the other is the RoITD dataset[22], which addresses technical domain questions in the IT sector. Additionally, the LiRo benchmark[23], which provides a comprehensive evaluation platform for Romanian NLP tasks, contributed with a subset of XQuAD[14] for Romanian, consisting of translated versions of English questions. For advanced biology, Dima et al.[24] published a small-scale dataset comprising 4127 single-choice questions. Existing datasets for Romanian QA are typically small (containing between 1000 and 11,000 QA pairs). With 105,880 QA pairs, MedQARo is one order of magnitude larger than the largest dataset for Romanian QA.

In Table 1, we present a comprehensive comparison between MedQARo and existing QA datasets across different languages and domains. MedQARo fills in a gap in the landscape of medical QA resources by providing the first large-scale Romanian medical QA dataset with 105,880 question-answer pairs focused on oncology. While other Romanian QA datasets exist, none of them includes extractive and reasoning questions based on medical case summaries. This positions MedQARo as the first large-scale resource for advancing medical NLP capabilities for Romanian, an underrepresented language in the QA field.

As illustrated in Fig. 1, our benchmark creation process involves two phases, (i) manual data collection and annotation, and (ii) model benchmarking. The dataset comprises 105,880 high-quality QA pairs from real-world clinical records of 1,242 oncology patients (796 patients with breast cancer, 215 patients with lung cancer, and 231 patients with other cancers). The QA pairs are the results of a manual annotation process carried out by physicians specialized in oncology and radiotherapy, who collectively spent nearly 3000 h to generate the pairs, ensuring both linguistic quality and medical correctness. MedQARo includes 76,416 QA pairs about breast cancer patients, 26,230 about lung cancer patients and 3234 about patients with other cancer types, with questions grounded in medical case summaries (epicrises).

To prevent data leakage across splits and support fair model development and evaluation, we provide official training, validation, and test splits at the patient level, i.e. a medical record can appear in only one of the splits. Moreover, we provide both in-domain and cross-domain test collections, where the cross-domain test set comprises patients from a distinct medical center, who suffer from one of five cancer types (head and neck, melanoma, renal cell carcinoma, bladder cancer, hepatocellular carcinoma) that are not available during training. Hence, MedQARo enables a precise assessment of generalization capabilities across different medical centers and cancer types. Oncologists, especially residents, could greatly benefit from using LLMs to extract and reason about the information included in epicrises, when they need to recommend diagnostic tools or treatment lines for their patients. MedQARo provides the necessary means to develop such LLMs, that could be deployed in daily medical practice. We benchmark four models on MedQARo, namely two LLMs specialized on Romanian (RoLLaMA2-7B and RoMistral-7B)[25], a long-context LLM (Phi-4-mini-instruct)[26], and an LLM tuned on biomedical data (LLaMA3-OpenBioLLM-8B)[1]. Each model is assessed in both zero-shot and supervised fine-tuning setups, while testing various prompt formats. In addition, we assess two state-of-the-art LLMs in the zero-shot setting, which are only accessible

through paid APIs, namely GPT-5.2 and Gemini 3 Flash. Our results show that all LLMs significantly benefit from fine-tuning, whereas zero-shot variants perform poorly. These findings emphasize the critical importance of both domain and language adaptation for reliable clinical QA in Romanian. Among all evaluated models, the fine-tuned RoMistral-7B achieves the best performance, outperforming all the other approaches. However, the best-performing model achieves an F1 score of only 0.671 on the in-domain test set, suggesting that MedQARo is a challenging dataset, highlighting the need for the development of more robust models, capable of better generalizing to specialized domains and low-resource languages.

In summary, our contributions are fivefold:

- We introduce **MedQARo**, the first large-scale medical QA dataset in Romanian, comprising 105,880 QA instances obtained with the help of trained physicians.
- We benchmark four LLMs from distinct families of models across multiple configurations and report comprehensive performance metrics under zero-shot and fine-tuning configurations, in both in-domain and cross-domain settings.
- We demonstrate that domain-specific and language-specific adaptation is crucial for reliable clinical QA in low-resource languages such as Romanian.
- We show that state-of-the-art API-based LLMs, such as GPT-5.2 and Gemini 3 Flash, perform far worse than smaller fine-tuned LLMs, further confirming the importance of fine-tuning.
- We release the dataset and code to reproduce the results at https://github.com/ana-rogoz/MedQARo, where patient sensitive data is fully anonymized.

Results

We conduct experiments with the chosen methods on the MedQARo dataset, using our official data split. We start by tuning various hyperparameters involved in the fine-tuning process, aiming to find the optimal hyperparameter setup for all LLMs. We next evaluate two different prompt formats, to determine which prompt format leads to higher performance. For the best performing prompt structure, we conduct several experiments across multiple LLMs, using various prompt lengths. The objectives of our experiments are: (i) to capture how Romanian-adapted models compare with general-purpose and domain-specific alternatives, (ii) to determine how fine-tuning performs against zero-shot inference, (iii) to establish how fine-tuned or zero-shot LLMs compare against trivial baselines, (iv) to investigate how the context length influences model performance, particularly for models with extended context capabilities, and (v) to determine the generalization capacity of LLMs across distinct data distributions, where patients come from different medical centers, and suffer from different cancer types.

Evaluation

We employ four alternative evaluation metrics:

- **F1 score:** The first metric that we take into consideration is the F1 score, which measures precision and recall at the token level. It provides a balanced view of how well models identify certain parts of the answer.
- **Exact match:** We also consider the exact match (EM) score, which represents the percentage of predictions that perfectly match the ground-truth answer. Although this metric is the most demanding, it is crucial in medical contexts where small wording differences can change meaning.
- **BLEU:** Because MedQARo contains answers with more than one token, we consider the BLEU score[27], which measures n-gram overlap to assess fluency and lexical similarity. When the predicted answer includes only a subset of the tokens from the ground-truth answer, BLEU can provide a representative measure by considering partial n-gram matches.
- **METEOR:** Lastly, we include METEOR[28], which offers a more flexible evaluation by accounting for synonyms, stemming, and word order variations. These aspects are particularly valuable given the rich morphology and flexible syntax of Romanian.

Model training setup and hyperparameter tuning

All models are fine-tuned using a LoRA strategy applied to the causal language modeling objective. For RoLLaMA2-7B-Instruct, RoMistral-7B-Instruct, and LLaMA3-OpenBioLLM-8B, LoRA adapters are injected into the self-attention projection layers (query, key, value, and output projections), following the default configuration for these architectures. For Phi-4-mini-instruct, due to architectural differences, LoRA is applied to the fused QKV projection layer and the output projection layer, respectively. In all cases, the original model parameters are kept frozen, and only the LoRA adapters are trainable. Feed-forward layers, token embeddings, layer normalization parameters, language modeling heads, and all bias terms remain frozen throughout training. This attention-only adaptation results in updating approximately 0.04–0.10% of the total number of weights, depending on the core architecture.

We employ consistent hyperparameter settings across all models to ensure a fair comparison among LLM families. All models are fine-tuned using the AdamW[29] optimizer. Due to memory constraints, we set a per-device batch size of one sample with gradient accumulation over 8 steps, resulting in an effective mini-batch size of 8 samples. Training is carried out for a maximum of 2 epochs, using Brain Floating Point (BFLOAT16) precision to optimize memory usage, while maintaining numerical stability. The learning rate, the dropout rate, and the LoRA configuration are established via grid search on the validation set. The range for the initial learning rate is 10^{-3} to 10^{-6}. We employ a cosine learning rate scheduler with 100 warm-up steps. Dropout is applied with drop rates between 0 and 0.2, using a step of 0.05. For the LoRA rank r, we look for an optimal value in the set {4, 6, 8, 10}. For the LoRA scaling factor, we consider values in the set {8, 16, 32}. The optimal configuration is based on a learning rate of $2 \cdot 10^{-5}$, a dropout rate of 0.05, a LoRA rank r of 8, and a LoRA scaling factor α of 16.

Preliminary results on prompt formatting

We conduct a preliminary experiment to assess the impact of the prompt structure on QA performance. Specifically, we compare two input formulations for the RoLLaMA2-7B-Instruct model: epicrisis + question + answer (E+Q+A) vs. question + epicrisis + answer (Q+E+A). Both configurations are evaluated with 4,096 input tokens. As shown in Table 2, the Q+E+A format consistently outperforms the E+Q+A format across all evaluation metrics. As confirmed by other studies[30], the first tokens tend to receive more attention, which might explain why placing the question at the beginning helps the model focus more on the question that needs to be

Table 2 | Preliminary results with alternative prompt formats for RoLLaMA based on 2048 tokens

Model	Prompt format	Validation				In-domain test			
		F1	EM	BLEU	METEOR	F1	EM	BLEU	METEOR
RoLLaMA2-7B	E+Q+A	0.4012	0.3082	0.2820	0.2843	0.4087	0.319	0.2915	0.2975
	Q+E+A	**0.5674**	**0.5572**	**0.5056**	**0.3645**	**0.5759**	**0.5708**	**0.5211**	**0.3688**

We compare two prompt formats: epicrisis + question + answer (E+Q+A) vs. question + epicrisis + answer (Q+E+A). The Q+E+A format yields higher performance across all evaluation metrics. The best results are highlighted in **bold**.

Table 3 | In-domain results of multiple open-source LLMs based on various configurations vs. two trivial baselines, on MedQARo

Model	Fine-tuned	#Tokens	Validation				In-domain test			
			F1	EM	BLEU	METEOR	F1	EM	BLEU	METEOR
Random token selector	N/A	1024	0.0023	0.0006	0.0006	0.0013	0.0019	0.0001	0.0004	0.0011
		2048	0.0021	0.0004	0.0004	0.0013	0.0024	0.0003	0.0005	0.0015
		4096	0.0022	0.0005	0.0006	0.0014	0.0026	0.0003	0.0005	0.0015
Majority answer	N/A	N/A	0.2091	0.1786	0.1831	0.1153	0.2148	0.1838	0.1884	0.1183
RoLLaMA2-7B	✗	2048	0.0188	0.0015	0.0024	0.0146	0.0230	0.0025	0.0023	0.0166
	✓	1024	0.5289	0.5529	0.4991	0.3312	0.5316	0.5551	0.4990	0.3305
	✓	2048	**0.5674**	**0.5572**	**0.5056**	**0.3645**	**0.5759**	**0.5708**	**0.5211**	**0.3688**
	✓	4096	0.3290	0.3055	0.2786	0.1980	0.3267	0.3267	0.3267	0.1980
RoMistral-7B	✗	4096	0.0715	0.0430	0.0234	0.0438	0.0605	0.0350	0.0229	0.0403
	✓	2048	**0.6731**	**0.6977**	**0.6458**	**0.4170**	**0.6716**	**0.6956**	**0.6445**	**0.4180**
	✓	4096	0.6568	0.6882	0.6405	0.4078	0.6587	0.6864	0.6391	0.4068
Phi-4-mini	✗	3072	0.0078	0.0000	0.0007	0.0075	0.0048	0.0002	0.0005	0.0043
	✓	2048	**0.6267**	**0.6575**	**0.5977**	**0.3897**	**0.6291**	**0.6593**	**0.5999**	**0.3916**
	✓	3072	0.6193	0.6493	0.5884	0.3853	0.6251	0.6529	0.5941	0.3876
	✓	4096	0.4806	0.4905	0.4463	0.2886	0.4809	0.4897	0.4477	0.2893
	✓	8192	0.5234	0.5367	0.4912	0.3127	0.5263	0.5374	0.4913	0.3156
	✓	16,384	0.5275	0.5430	0.4963	0.3165	0.5297	0.5415	0.4945	0.3183
	✓	32,768	0.5277	0.5434	0.4970	0.3166	0.5299	0.5418	0.4948	0.3186
OpenBioLLM-8B	✗	2048	0.0140	0.0004	0.0014	0.0156	0.0119	0.0004	0.0015	0.0132
	✓	2048	**0.5465**	**0.5730**	**0.5239**	**0.3389**	**0.5520**	**0.5785**	**0.5287**	**0.3432**
	✓	4096	0.5217	0.5322	0.4873	0.3180	0.5208	0.5338	0.4882	0.3203

For each LLM, there is a zero-shot version and several fine-tuned versions, with various prompt lengths. F1, Exact Match (EM), BLEU, and METEOR scores are reported on both validation and test sets. Fine-tuned models consistently outperform zero-shot versions. The best results for each LLM are highlighted in **bold**.

answered. Given these results, we decided to run all subsequent experiments using the Q+E+A prompt structure.

Main results

In Table 3, we present the results obtained by the chosen open-source LLMs vs. two baselines on MedQARo. We evaluate all four LLMs under both zero-shot prompting and supervised fine-tuning setups, reporting performance metrics on both validation and in-domain test sets.

We first compare baselines vs. LLMs. In the zero-shot setup, the performance levels reached by the various LLMs indicate that out-of-the-box models can surpass the naive random token selector. However, the majority answer baseline, which leverages knowledge about the answer distribution in training data, significantly outperforms all zero-shot variants. This is a first hint that zero-shot models are not able to generalize to the medical QA task in Romanian, regardless of their prior language or domain adaptation. In contrast, each fine-tuned LLM version surpasses the two baseline models, irrespective of prompt length and prior model adaption. This observation holds for all LLM architectures, indicating that the fine-tuning process is generally useful.

We next compare zero-shot vs. fine-tuned LLMs. In general, zero-shot models exhibit significantly lower performance than their fine-tuned counterparts. Furthermore, for various performance metrics and LLM architectures, we observe that the scores can increase by an order of magnitude when going from zero-shot prompting to fine-tuning. The comparison between zero-shot and fine-tuning setups highlights the critical importance of task-specific fine-tuning for clinical QA, confirming that prior language-specific or domain-specific adaptation is insufficient to accurately address the task.

We now analyze the cross-domain evaluation results. In Table 4, we report the results obtained on the cross-domain test set, which includes patients from a different medical center, who suffer from other cancer types than those observed during training, thus representing a substantially harder scenario for fine-tuned models. As expected, all fine-tuned models exhibit a noticeable performance drop with respect to the in-domain test set (see Table 3), confirming that cross-center and cross-diagnosis evaluation poses significant challenges. The majority answer baseline achieves low but non-trivial scores, highlighting that naive priors still provide a competitive lower bound in this setting. Zero-shot variants remain largely ineffective, in some cases performing close to the majority answer baseline, which further emphasizes that neither prior Romanian language specialization nor biomedical pretraining alone is sufficient to handle distribution shifts.

RoMistral-7B demonstrates the strongest robustness among all fine-tuned LLMs, achieving the highest F1 and EM scores with both prompt lengths. Phi-4-mini-instruct, remains the second best performer in both in-domain and cross-domain settings, while experiencing pronounced degradation when going from the in-domain setup to the cross-domain setup. OpenBioLLM-8B benefits from fine-tuning, but fails to outperform the Romanian-adapted RoMistral-7B model or the generic Phi-4-mini-instruct model, suggesting that English biomedical pretraining does not adequately compensate for cross-lingual and cross-center discrepancies. Among the fine-tuned models, we find that RoMistral-7B attains the highest resilience to both institutional and diagnostic variability. The fact that fine-tuned LLMs remain significantly above their zero-shot counterparts indicates that simple fine-tuning can bring useful question answering capabilities that generalize to different data distributions. Still, fine-tuning alone is not sufficient to achieve full generalization capacity across cancer types and medical centers, underlining the opportunity to introduce domain-adaptation strategies.

We further analyze the effect of prompt length. We test the models with various prompt lengths, taking into account the limitation imposed by each

Table 4 | Cross-domain (cross-center and cross-diagnosis) results of multiple open-source LLMs based on various configurations vs. the majority answer baseline, on MedQARo

Model	Fine-tuned	#Tokens	Cross-domain test			
			F1	EM	BLEU	METEOR
Majority answer	N/A	N/A	0.0909	0.0544	0.0544	0.0402
RoLLaMA2-7B	✗	2048	0.0746	0.0087	0.0138	0.0617
	✓	1024	0.2919	0.2313	0.2030	0.1425
	✓	2048	**0.3036**	**0.2319**	**0.2126**	**0.1526**
	✓	4096	0.1205	0.0816	0.0732	0.0614
RoMistral-7B	✗	2048	0.1286	0.0544	0.0469	0.0959
	✓	2048	0.4387	0.3605	0.3688	0.2225
	✓	4096	**0.4454**	**0.3624**	**0.3696**	**0.2248**
Phi-4-mini	✗	2048	0.0120	0.000	0.0020	0.0112
	✓	2048	**0.3648**	**0.3092**	**0.3120**	**0.1865**
	✓	3072	0.3446	0.2993	0.3038	0.1762
	✓	4096	0.3052	0.2669	0.2670	0.1587
	✓	8192	0.3663	0.3101	0.3061	0.1883
	✓	16,384	0.3621	0.3061	0.3004	0.1843
	✓	32,768	0.3613	0.3061	0.3018	0.1838
OpenBioLLM-8B	✗	2048	0.0196	0.0006	0.0029	0.0183
	✓	2048	**0.3165**	**0.2638**	**0.2682**	**0.1606**
	✓	4096	0.2932	0.2440	0.2425	0.1503

For each LLM, there is a zero-shot version and several fine-tuned versions, with various prompt lengths. Models are evaluated in terms of F1, EM, BLEU, and METEOR on a single held-out set comprising 231 patients and 3234 QA pairs. The best results for each LLM are highlighted in **bold**.

Table 5 | Comparison of best-performing model variants against state-of-the-art baselines on a randomly sampled subset of 100 QA pairs from the in-domain test set

Model	Fine-tuned	F1	EM	BLEU	METEOR
GPT-5.2	✗	0.2415	0.1900	0.1339	0.1169
Gemini 3 Flash	✗	0.2041	0.1400	0.0099	0.1039
RoLLaMA2-7B	✗	0.0257	0.0000	0.0009	0.0187
	✓	0.5213	0.5100	0.4730	0.3259
RoMistral-7B	✗	0.0892	0.0700	0.0015	0.0568
	✓	**0.6309**	**0.6700**	**0.6402**	**0.3862**
Phi-4-mini	✗	0.0012	0.0000	0.0007	0.0045
	✓	0.5867	0.5900	0.5413	0.3482
OpenBioLLM-8B	✗	0.0025	0.0000	0.0003	0.0031
	✓	0.5180	0.5400	0.4997	0.3198

State-of-the-art API-based LLMs are evaluated in the zero-shot setting, while the remaining models are either zero-shot prompted or fine-tuned on MedQARo. The best F1, EM, BLEU, and METEOR scores are highlighted in **bold**.

LLM architecture on its input prompt (see Tables 3 and 4). While the average length of an epicrisis is 7,829 tokens (as per Table 9), the LLMs do not necessarily benefit from using extensive prompts. All of our models obtain superior performance levels with 2048 input tokens than with 4096 tokens or longer sequences.

Phi-4-mini-instruct is the only model that accepts very long prompts. Although we explore prompts of up to 32,768 tokens for Phi-4-mini-instruct, the model does not benefit from the extended input.

We further compare open-source vs. API-based LLMs. To compare the open-source LLMs with state-of-the-art API-based models, we randomly select 100 QA pairs from the in-domain test set. The limited test set size is determined by our resource constraints, while the use of zero-shot prompting for the API-based models is due to the access restrictions imposed by OpenAI and Google, respectively. We present the comparative results in Table 5. For a fair comparison, the results of each model are based on an optimal number of input tokens. GPT-5.2 and Gemini 3 Flash substantially outperform the lighter open-source models when both API-based and open-source LLMs are evaluated in the zero-shot setting. This highlights the effectiveness of large-scale general pretraining, even without task-specific supervision. Despite the strong performance of both GPT-5.2 and Gemini 3 Flash, all fine-tuned models consistently outperform the API-based LLMs across all evaluation metrics, confirming the benefits of task-specific and domain-specific fine-tuning. This comparison reinforces the importance of fine-tuning to achieve high performance on specialized clinical QA tasks in low-resource languages.

We next assess the impact of using the first vs. multiple input chunks. Since epicrises are trimmed from the end to fit the desired prompt lengths, the generally low performance levels when using long prompts can be explained by two factors: (i) most questions can be answered just by looking at the beginning part of an epicrisis, or (ii) taking long contexts into account is a challenging learning task, potentially leading to overfitting. To distinguish between these two cases, we conduct another experiment with one of the best performing models, Phi-4-mini-instruct based on 2048 input tokens. In addition to the version that simply trims the epicrisis from the end, we evaluate a version that analyzes all non-overlapping 2048-token chunks from the epicrisis. For the latter approach, we employ majority voting over all chunks to obtain the final answer. To break ties, we rely on the average log-probability over the generated tokens and select the answer with the highest value. The corresponding results are reported in Table 6. While the chunk-based approach enables processing of complete epicrises, it results in substantial performance degradation, suggesting that focusing on the first part of the epicrisis leads to optimal results. Still, as shown by our previous experiments reported in Table 3, using a context that is too short, e.g. 1,024 tokens for RoLLaMA2-7B-Instruct or RoMistral-7B-Instruct, might also lead to performance degradation. In summary, trimming the end part of the epicrisis seems to act as a regularization technique, and finding the optimal context length can be assimilated with the process of tuning the corresponding regularization hyperparameter.

We further showcase the importance of using multiple evaluation metrics. Regarding the employed evaluation metrics, we observe that EM scores exceed F1 scores by small margins across most configurations, suggesting that the models tend to either match answers exactly or miss them entirely, with fewer partial matches. Moreover, METEOR scores are consistently lower than other metrics, likely reflecting the sensitivity of METEOR to synonyms and word variations in Romanian medical terminology, but also its inherent bias against short answers. Nevertheless, for a holistic evaluation of QA models on MedQARo, we promote the use of all four evaluation metrics.

We next analyze the performance per question type. In Fig. 2, we present the results of one of our best performing model, for each of the three question categories. The F1 and EM metrics suggest that binary questions are easier to answer, while reasoning questions are harder. This observation is consistent with our natural expectation. However, the BLEU score seems to promote reasoning questions, while METEOR indicates that the model is more capable on extractive questions. Overall, the analysis shows that some metrics are more appropriate for certain question categories. This represents another strong argument in favor of using a wide variety of metrics to accurately benchmark models on MedQARo.

We further conclude our results with a number of general remarks. Our experimental results reveal several patterns across all models and

Table 6 | Results of Phi-4-mini-instruct on MedQARo, using only the first 2048-token chunk of the epicrisis vs. all 2,048-token chunks

Model	Text chunks	Validation				In-domain test			
		F1	EM	BLEU	METEOR	F1	EM	BLEU	METEOR
Phi-4-mini	First	**0.6267**	**0.6575**	**0.5977**	**0.3897**	**0.6291**	**0.6593**	**0.5999**	**0.3916**
	All	0.4354	0.4806	0.4162	0.2758	0.4530	0.4966	0.4328	0.2839

The former approach significantly outperforms the latter one across all metrics, suggesting that only analyzing the first part of the epicrisis reduces input clutter and prevents overfitting. The best results are highlighted in **bold**.

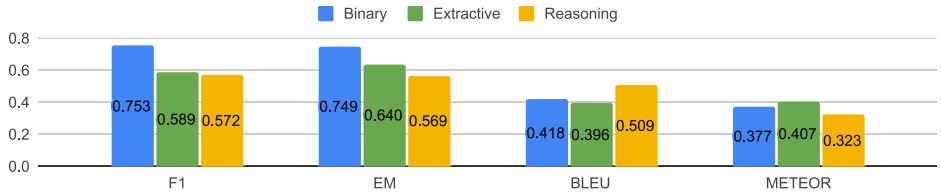

Fig. 2 | Performance per question category (binary, extractive, and reasoning) across four evaluation metrics (F1, EM, BLEU, and METEOR). The values are reported for one of the top scoring models, namely Phi-4-mini-instruct based on 2048 tokens, on the in-domain test set. Blue bars correspond to binary questions. Green bars correspond to extractive questions. Orange bars correspond to reasoning questions.

Table 7 | Examples of test questions with correct and wrong answers provided by the best performing LLM, namely Phi-4-mini-instruct based on 3072 tokens

Question (Romanian / English)	Ground-truth	Predicted	Correct?
Imunoterapia a fost introdusă i^n a doua etapă a schemei terapeutice? / Was immunotherapy introduced in the second stage of the therapeutic regimen?	Nu / No	Nu / No	✓
S-au identificat metastaze la nivelul plămânilor? / Are lung metastases present?	Da / Yes	Da / Yes	✓
Care este stadiul tumoral la diagnostic al pacientului? / What was the tumor stage at the time of diagnosis for this patient?	IIA	IIA	✓
Care este subtipul histologic al pacientului din rezultatul histopatologic? / What is the histological subtype of the patient according to the histopathology report?	ADK	ADK	✓
Ce cadrane ale sânului sunt invadate de tumoră? / Which breast quadrants are invaded by the tumor?	supero-extern / upper-outer	supero-extern / upper-outer	✓
S-au utilizat alte linii de tratament i^n progresia bolii? / Were other lines of treatment used during disease progression?	CHT	Nu / No	✗
S-au identificat metastaze la nivelul oaselor? / Are bone metastases present?	Nu / No	Da / Yes	✗
Sindrom inflamator este prezent la acest pacient? / Is the inflammatory syndrome present for this patient?	prezent / present	absent / absent	✗
Care este stadializarea TNM a pacientului? / What is the TNM staging of the patient?	cT4cT2cN2M0 / pT4bPNx	cT4cN2M0	✗
Care este stadiul tumorii stabilit la i^nceputul tratamentului? / What is the stage of the tumor established at the beginning of treatment?	IV	IIIB	✗

Each row shows a question along with ground-truth and predicted answers. English translations are provided when necessary.

configurations. First, we observe that in-domain validation and test results remain consistent across all experiments, indicating that our patient-level data splits effectively prevent data leakage. All fine-tuned LLMs exceed the naive baselines by large margins, confirming that models learn meaningful clinical question-answering capabilities. We also find that fine-tuned models significantly outperform their zero-shot counterparts across all architectures, emphasizing that neither prior Romanian language adaptation nor English biomedical pretraining is sufficient for the medical QA task. Among the four open-source LLM families, RoMistral-7B-Instruct obtains the highest in-domain performance levels. In contrast, OpenBio-LLM-8B achieves the lowest scores among all evaluated models on the in-domain test set, despite being pretrained on the medical domain. The cross-domain evaluation indicates that fine-tuning is also beneficial for patients enrolled in different medical centers, who suffer from other cancer types. However, fine-tuned LLMs exhibit a visible performance degradation in the cross-center and cross-diagnosis setup, suggesting that fine-tuning alone is not enough to recover the domain gap. Regarding the comparison between the zero-shot API-based LLMs and fine-tuned open-source LLMs, we find that GPT-5.2 and Gemini 3 Flash obtain reasonable performance levels, but they are clearly outperformed by fine-tuned LLMs. This finding emphasizes that fine-tuning open-source LLMs is preferred over zero-shot prompting larger models gated by paywalls.

Qualitative analysis

To better understand the challenges posed by MedQARo and the capabilities of one of the best-performing model, we conduct a qualitative analysis of a random selection of QA pairs with correct and wrong answers predicted by Phi-4-mini-instruct. In Table 7, we showcase 10 examples, 5 that are correctly answered and 5 that are incorrectly answered, respectively.

We start by analyzing correctly answered questions. In the first example, the LLM was able to recognize the sequence of treatment lines and to distinguish between the first line and the second line of therapy, even though the information was not explicitly stated in the epicrisis. In the second example, pulmonary metastases were mentioned under the abbreviation M1PUL. The model recognized that M1PUL denotes the presence of lung metastases, providing the correct answer based on this information. In

the third example, the LLM demonstrates the ability to correctly stage breast cancer, which requires high-level medical reasoning. The model correlated tumor size, regional invasion, and the absence of distant metastases, successfully assigning stage IIA, according to the TNM criteria. In the fourth case, the histological type "adenocarcinom" / "adenocarcinoma" was explicitly stated in the histopathology report, which allowed the LLM to correctly answer with "ADK" to the posed question. In the fifth example, the interpretation of the chest CT explicitly stated "invazia tumorală i^n cadranul supero-extern" / "tumor invasion present in the upper-outer quadrant". The model was able to accurately spot this information and provide the correct answer. Overall, we observe that the model demonstrates the necessary capabilities to answer both reasoning and extractive questions.

We continue by analyzing wrongly answered questions. In the first example, the annotator considered that the presence of "CHT" in the epicrisis indicates a new line of treatment at disease progression. However, the LLM answered "Nu" / "No", either because it did not recognize that "CHT" represented a new line of therapy, or because it required an explicit statement confirming the change of treatment line, which is missing from the epicrisis. In the latter case, the model might have considered that "CHT" is part of the initial treatment, instead of a new line of therapy. Note that this is not exactly a binary question, since the correct answer for this kind of question is a list of treatment lines. When the list is empty, the correct answer is "Nu" / "No". In the second example, the LLM may have been triggered by the presence of the word "os" / "bone". The confusion is caused by an investigation reported in the epicrisis, called "scintigrafie osoasă" / "bone scintigraphy", which is usually recommended for patients with bone metastases. Despite undergoing this investigation, the patient did not have any bone metastases. In the third case, the inflammatory syndrome is not directly mentioned in the epicrisis, although it can be identified by analyzing biological parameters (e.g. elevated ESR, increased CRP, elevated fibrinogen) or secondary diagnoses (e.g. infection, chronic inflammation). The physician recognized these signs and determined the presence of the inflammatory syndrome. Without an explicit mention of "sindrom inflamator" / "inflammatory syndrome", the LLM was not able to extract the correct answer. In the fourth case, the LLM provided only a partially correct answer. It correctly identified the clinical TNM classification, but did not mention the pathological TNM classification. This is a limitation induced by the prompt length of 3,072 tokens, since the pathological TNM classification is usually specified near the end of the epicrisis, which gets removed due to context trimming. In the fifth example, the patient had lymph node metastases. The physician considered the staging to be stage IV, whereas the LLM interpreted the involvement as loco-regional lymph node invasion and classified the case as stage IIIB. Overall, the wrong or partially incorrect answers point out two kinds of limitations: (i) the reasoning abilities of Phi-4-mini-instruct can sometimes hinder the integration of high-level medical information, and (ii) the context trimming procedure might sometimes lead to dropping important clues.

Discussion

In this paper, we introduced MedQARo, the first large-scale benchmark for Romanian medical question answering. The new benchmark is the result of a meticulous data collection and annotation process carried out by experts in the medical domain, who constructed a dataset of 105,880 QA pairs. The QA samples are grounded in a collection of 1242 medical case summaries of cancer patients. We conducted comprehensive experiments with several LLM families to assess the impact of various aspects, namely prompt structure and length, language vs. domain vs. general pretraining, and zero-shot prompting vs. LoRA-based fine-tuning. Our empirical results suggested some interesting findings. First of all, we observed that long prompts are not always helpful, degrading performance even if epicrises are typically long documents. Second of all, we found that language specialization (e.g., RoMistral-7B) and generic pretraining (e.g., Phi-4-mini) offer a better starting point for fine-tuning than medical domain specialization (e.g. OpenBioLLM-8B). Third of all, we determined that fine-tuning on our specific task significantly outperforms zero-shot prompting. Yet, even the best model was only able to correctly answer less than 70% of the in-domain test questions and 40% of the cross-domain test questions, indicating that MedQARo is a challenging benchmark.

In future work, we aim to develop models that integrate retrieval-augmented generation to improve question answering capabilities. More specifically, we will focus on developing a retrieval module to accurately find the sections of the input epicrisis that are likely to contain indicators for the answer. This will allow us to use models with reasonable context lengths, while minimizing the risk of eliminating important information during context trimming. However, since the answer is not always directly mentioned in the epicrisis, finding the right sections from which the answer could be derived requires careful consideration.

Methods

Data collection

MedQARo is obtained through the manual extraction of information from medical documents, namely medical summaries of patients that are registered either in the Oncology Department of Colțea Clinical Hospital, or in the "Dan Furtună" Medical Center, both located in Bucharest, Romania. The epicrises, originally drafted in Word format by the physicians attending each patient, were manually reviewed by a group of seven physicians involved in constructing MedQARo. In this group, two experienced physicians from Colțea Clinical Hospital, who led the data collection process, established the most prevalent cancer sites in their hospital, namely breast and lung. They selected patients with either one of these diagnoses, who received treatment at the hospital between 2019 and 2024. This selection process led to a set of 796 patients with breast cancer and 215 patients with lung cancer. To create a diverse multi-center and multi-cancer benchmark, 231 patients with one of five cancer types (head and neck, melanoma, renal cell carcinoma, bladder cancer, hepatocellular carcinoma) were included in the study. These patients are enrolled in cancer follow-up care programs at "Dan Furtună" Medical Center.

The two experienced physicians further proposed a set of reference questions for each cancer site, comprising 48 questions for breast cancer, 61 questions for lung cancer, and 7 general questions that cover all other types. The other five physicians proposed 8–20 equivalent reformulations of the reference questions. Next, the physicians proceeded by extracting the answers for each unique question from each epicrisis and entering the respective answers into a database. To optimize annotation time, the questions for each epicrisis are grouped and annotated in the same batch. This means that a physician could read an epicrisis once, before answering multiple questions about the respective patient.

Data extraction and completion were carried out via a two-level process. The first level was performed by five medical oncology and radiotherapy residents, who meticulously completed the database of QA pairs following prior training based on the annotation guidelines provided by the experienced physicians. The second level was carried out by two board-certified medical oncology specialists, holders of a good standing certificate issued by the Romanian College of Physicians and with proven experience in good medical and clinical research practices. They verified the accuracy and consistency of the data entered by the residents. In total, the annotation process involved seven physicians, with a cumulative effort of approximately 2170 hours (equivalent to about 310 hours per physician) just to type in the answers. Neither the time required to formulate questions nor the time required to read an epicrisis before answering all questions for a certain patient is taken into account in the time measurement. While the annotation interface did not allow us to measure the extra time required to read epicrises, we estimate that this step required between 600 and 800 additional hours. The time required to formulate all question variants is estimated at 50–70 h. Adding all numbers, we estimate that the entire annotation process took nearly 3000 h. The number of patients processed by each resident ranged from a minimum of 200 to a maximum of 280 cases, depending on individual availability and case complexity. To measure the inter-annotator agreement, a number of 350 randomly chosen QA pairs were annotated by all

five physicians. The average Cohen's κ score between each pair of annotators is 0.9562, which indicates an almost perfect agreement.

Furthermore, each experienced specialist meticulously reviewed the QA pairs associated with about 600–630 patients. This distribution of the workload was planned to ensure a relatively balanced annotation process. For data extraction, the experienced physicians developed an annotation guideline, establishing clear completion rules, including: staging of cancers according to the Tumor Node Metastasis (TNM) classification; grading of adverse events according to the Common Terminology Criteria for Adverse Events (CTCAE) system; standardization of medical abbreviations (e.g. "CHT" for chemotherapy, "RT" for radiotherapy, "NK" for "not known" values); handling incomplete information. In addition to these general rules, the guideline included a detailed list of answer options to be extracted and recorded in a standardized format: Eastern Cooperative Oncology Group (ECOG) performance status, menopausal status, molecular subtype, tumor proliferation index (Ki-67), histological grade, tumor size, family history of cancer, personal history of other cancers, body mass index (BMI), tumor markers, presence/absence of disease progression, number of metastatic sites, metastatic locations, death, etc. The senior physicians also acted as referees to solve annotation conflicts, choosing the final answer between two options based on their own high-skilled assessment.

Ethics approval

Regarding research ethics, we emphasize that all data was fully anonymized prior to processing by removing any direct identifiers (names, dates of birth, personal identification numbers, etc.). The use of anonymized data for research purposes is based on the informed consent of patients, as recorded in the medical charts of Colțea Clinical Hospital and "Dan Furtună" Medical Center, and signed individually by each patient at admission. The database was approved by the Ethics Committee of Colțea Clinical Hospital (Anca Lupu, Ciprian Aldea, Traian Burcoș, Șerban Berteșteanu, Lavinia Coțofană, Letiția Coriu, Gabriela Caragescu, Livia Neacșu) where the research was conducted, in accordance with Decision No. 19091, dated October 5th, 2021. The database was also approved by the Ethics Committee of "Dan Furtună" Medical Center (Bogdan Ioan Furtună, Cipriana Dimitrakopoulos, Alice Corina Fărcaș, Natalia Bărcaru, Daniela Tomescu), as per Decision No. 1323, dated 11th December, 2025. The entire process complies with the requirements of the General Data Protection Regulation (GDPR), the principles of the Declaration of Helsinki, and the standards of good medical and research practices.

Data Preprocessing

To support model training and evaluation, we constructed two domain-specific datasets targeting question answering in oncology given a clinical context: one for breast cancer and one for lung cancer. These are complemented by a generic dataset targeting generic questions covering five other cancer types (not including breast and lung cancer).

We collected data from 796 breast cancer patients, each associated with a complete epicrisis, i.e., a detailed clinical summary for each patient, and a corresponding set of answers to 48 unique medical questions. To simulate natural variation in clinical queries, each question is reformulated by multiple physicians, leading multiple alternative formulations. For every patient-question pair, we randomly sample two question alternatives and generate two QA instances, each structured as a tuple of the form (patient ID, epicrisis, question, answer). This process resulted in a total of 76,416 unique examples.

We compiled data from another 215 lung patients, each accompanied by an epicrisis summarizing their clinical history. For each patient, we curated a set of answers to 61 unique and clinically relevant questions. Similar to the breast cancer setup, each question is manually paraphrased into multiple linguistically diverse formulations to better reflect natural question variability. For each patient-question pair, we randomly sampled two different question variants. This leads to a total of 26,230 unique QA instances, capturing both clinical diversity and linguistic variation.

To construct a benchmark that enables the assessment of generalization capacity across medical centers and cancer types, we created a secondary dataset, which is not used during training. This dataset comprises 231 patients, diagnosed with a different kind of cancer (head and neck, melanoma, renal cell carcinoma, bladder cancer, or hepatocellular carcinoma), all coming from a different center than the patients included in our main dataset. For these patients, there are 7 general questions, relevant for all five cancer types. As for breast and lung cancer patients, we randomly sampled two different question variants for each patient-question pair, leading to a total of 3234 unique QA pairs.

Data partitioning and statistics

We merge the breast and lung cancer patients (from the Colțea Clinical Hospital) and corresponding QA pairs into a single dataset. Then, the merged dataset is split into training, validation, and in-domain test, using a 70%/15%/15% split ratio, ensuring that the division occurs at the patient level. We prefer this split over splitting individual QA pairs because it guarantees that all the data from a single patient, including their epicrisis and associated QA pairs, remains within exactly one subset, thereby preventing data leakage among splits. Furthermore, we keep the patients with other cancers (from the "Dan Furtună" Medical Center) for cross-domain evaluation.

We present the distribution of our dataset across training, validation, and test sets in Table 8. For the breast cancer subset, our split results in 557 patients allocated for training, 119 for validation, and 120 for testing. In terms of QA pairs, the split distributes 53,472 samples for training, 11,424 for validation, and 11,520 for testing, respectively. Similarly, the lung cancer subset is divided into 150 training, 32 validation, and 33 test patients, yielding 18,300, 3904, and 4026 QA pairs, respectively. A separate cross-domain test set contains 231 patients and 3234 QA pairs. In summary, MedQARo contains four data partitions: one for training (707 patients), one for in-domain validation (151 patients), one for in-domain testing (153 patients), and one for cross-domain testing (231 patients). In total, MedQARo comprises 1242 unique patients (796 suffering from breast cancer,

Table 8 | A summary of the number of patients and corresponding QA pairs included in the MedQARo benchmark across cancer types

Patient group	Number of patients				Number of samples			
	Train	Validation	Test	Total	Train	Validation	Test	Total
Breast cancer	557	119	120	796	53,472	11,424	11,520	76,416
Lung cancer	150	32	33	215	18,300	3904	4026	26,230
Other cancers	–	–	231	231	–	–	3234	3234
Total	707	151	384	1242	71,772	15,328	18,780	105,880

For breast and lung cancer, the splits are performed at the patient level, into training (70%), validation (15%), and in-domain test (15%), to prevent information leakage across splits. A cross-domain test set comprises a separate set of 231 patients with one of five cancer types (excluding breast and lung), from a different medical center. We report the number of patients per cancer type per split, and the corresponding number of QA samples generated for each split. The final row aggregates the total number of unique patients and QA pairs across different cancer types for each split. MedQARo comprises a total 105,880 QA pairs for 1242 patients.

Table 9 | A summary of the minimum, maximum, average and total number of tokens for the three kinds of data (epicrisis, question and answer) available in MedQARo

Data type	Number of tokens			
	Min	Max	Mean	Total
Epicrisis	207	34,247	7,829	9,723,618
Question	6	32	16	1,778,675
Answer	1	442	4	432,754

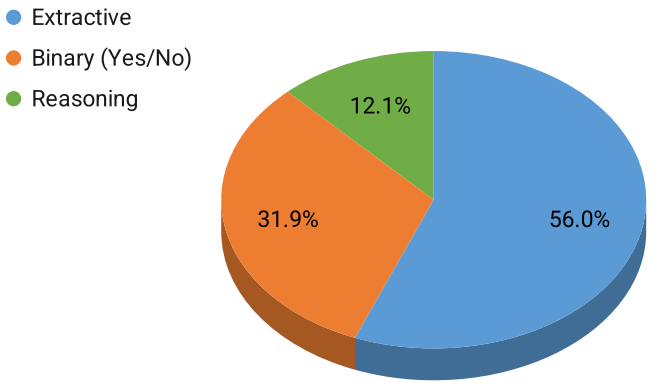

Fig. 3 | Distribution of question types in the MedQARo dataset. The dataset comprises three main categories: binary (yes/no) questions, extractive questions (answers explicitly found in the epicrisis), and reasoning questions (requiring inference beyond explicit mentions). Percentages are computed over the full set of 105,880 QA pairs. Best viewed in color. Blue represents extractive questions. Orange represents binary questions. Green represents reasoning questions.

215 from lung cancer, and 231 from five other cancers) and 105,880 question-answer pairs. Hence, our contribution is a substantial resource for training and evaluating Romanian medical QA systems, in both in-domain and cross-domain scenarios.

In Table 9, we report statistics about the number of tokens in epicrises, questions and answers included in MedQARo. Epicrises range from 207 to 34,247 tokens, with a mean value of 7829 tokens, where shorter documents typically correspond to patients under investigation or at early treatment stages, while longer ones capture extensive treatment histories with multiple interventions. This variation can pose challenges for some models with limited context windows. While epicrises are typically lengthy, questions and answers are much shorter. Questions maintain relative consistency, between 7 and 30 tokens, with a mean value of 15 tokens, reflecting the consistency of clinical query formulations. The answers are represented by predominantly brief texts with a mean of 4 tokens, varying in length from one to 442 tokens. Longer responses are typical for questions involving complex medical reasoning.

We categorize the questions into three types: (i) binary (yes/no); (ii) extractive, where the answer appears explicitly in the clinical context (epicrisis); and (iii) reasoning, where the answer must be inferred from multiple clues within the epicrisis. The distribution of these categories in MedQARo is shown in Fig. 3. Within the binary subset, the label distribution is moderately skewed, with roughly two-thirds of answers being "Nu" ("No") and one-third being "Da" ("Yes"). The slightly skewed distribution resulted naturally from the data collection process. In Table 10, we showcase some representative examples from the three question categories included in MedQARo. Some questions are specific to a certain cancer type, e.g. "Ce cadrane ale sânului sunt invadate de tumoră?" / "Which breast quadrants are invaded by the tumor?" is only relevant for breast cancer, while other questions are valid for all cancer types, e.g., "Care este stadiul tumoral la diagnostic al pacientului?" / "What was the tumor stage at the time of diagnosis for this patient?" is applicable to all cancers. The reformulated questions indicate that semantic equivalence was ensured during the process of manual question generation conducted by the annotators.

Table 10 | Representative examples of binary, extractive, and reasoning questions from MedQARo

Type	Question (Romanian / English)
Binary	S-a realizat o intervenție chirurgicală i^nainte de inițierea imunoterapiei? / Has a surgical intervention been performed before initiating immunotherapy?
	Intervenția chirurgicală a precedat imunoterapia i^n cazul acestui pacient? / Did surgery precede immunotherapy for this patient?
	Pacientul are gena BRCA1? / Does the patient have the BRCA1 gene?
	Rezultatul analizei genetice indică prezența mutației BRCA1? / Does the genetic analysis result indicate the presence of the BRCA1 mutation?
	S-au identificat metastaze la nivelul plămânilor? / Have lung metastases been identified?
	Există determinări secundare pulmonare la acest pacient? / Are there any secondary pulmonary tumors in this patient?
Extractive	Care a fost dimensiunea tumorii la debut? / What was the tumor size at the time of diagnosis?
	Cât de mare era tumora atunci când a fost stabilit diagnosticul? / How large was the tumor when the diagnosis was made?
	Ce cadrane ale sânului sunt invadate de tumoră? / Which breast quadrants are invaded by the tumor?
	Specifică cadranele invadate de tumoră / Specify the quadrants invaded by the tumor
	Ce status ECOG are pacientul? / What is the ECOG status of the pacient?
	Specifică statusul de performanță ECOG / Specify the ECOG performance status
Reasoning	Câte luni sunt de la i^nceputul tratamentului până la apariția progresiei bolii? / How many months have passed from the start of treatment until the onset of disease progression?
	Cât a durat (i^n luni) ca boala să progreseze sub tratament? / How long did it take (in months) for the disease to progress under treatment?
	Câte luni au trecut de la chirurgie până la i^nceperea tratamentului cu radioterapie? / How many months have elapsed from surgery until the initiation of radiotherapy treatment?
	La cât timp după operație a i^nceput tratamentul radioterapic? / How long after surgery did radiotherapy treatment begin?
	Care este stadiul tumoral la diagnostic al pacientului? / What was the tumor stage at the time of diagnosis for this patient?
	Ce stadiu de cancer a fost stabilit la diagnostic? / What stage of cancer was determined at diagnosis?

Each question is shown in Romanian, followed by its English translation. One reformulation is displayed below each distinct question.

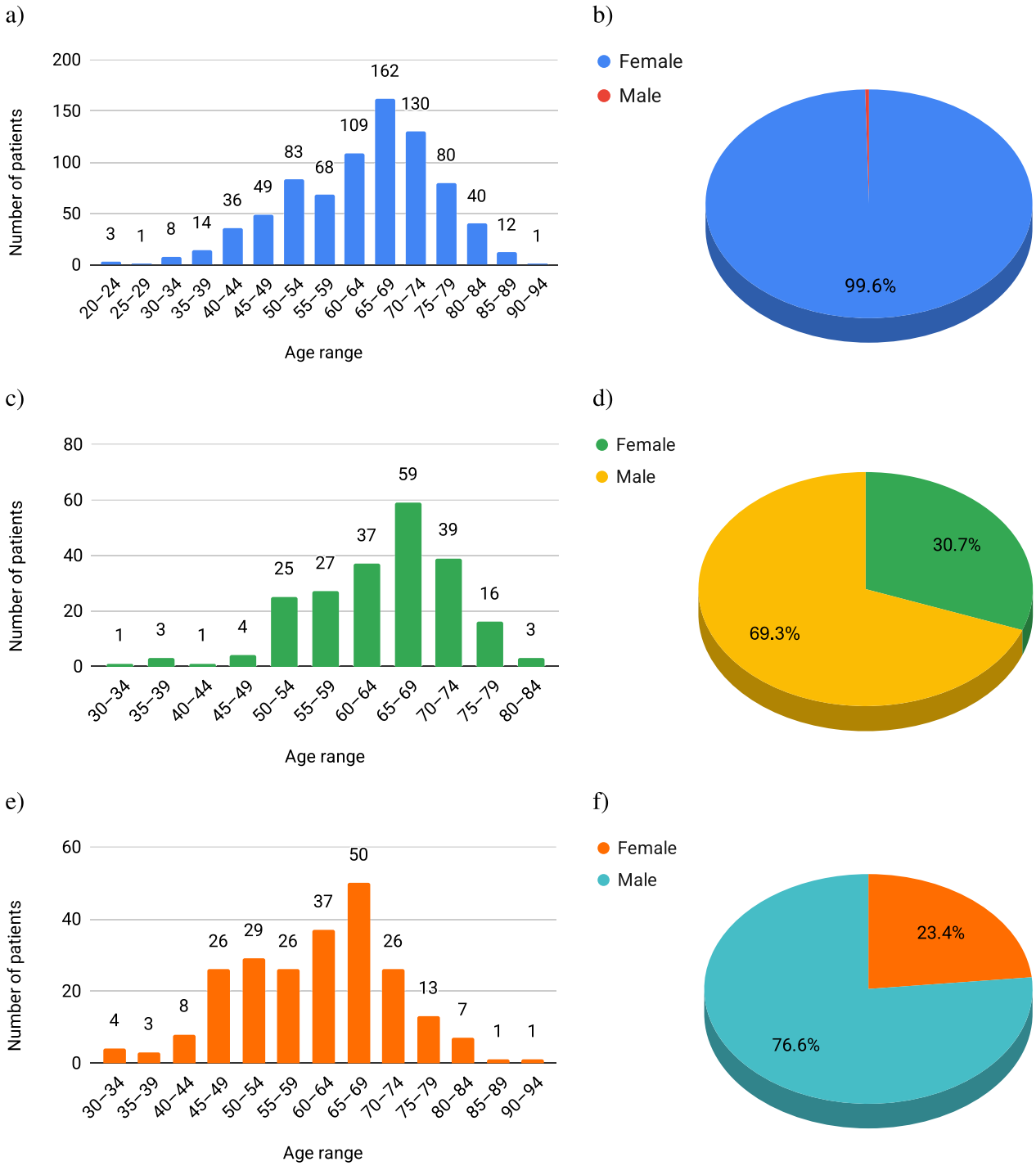

Fig. 4 | Age and gender distributions of patients included in MedQARo. Separate distributions are reported for breast, lung, and other cancer patients. Age distributions are based on 5-year bins (between 20 and 94 years). **a** Age distribution for breast cancer. **b** Gender distribution for breast cancer. **c** Age distribution for lung cancer. **d** Gender distribution for lung cancer. **e** Age distribution for other cancers. **f** Gender distribution for other cancers.

Data demographics

We further examine the age and gender distributions of patients included in MedQARo. In Fig. 4, we report the distributions for the three patient groups, which are separated by cancer type: breast cancer, lung cancer, and other cancers. Across all diagnostic categories, the cohort is predominantly composed of middle-aged and elderly patients, with most cases falling in the 50–79 age range. As naturally expected, most breast cancer patients are females (793 out of 796), as breast cancer is less common in the male population. In contrast, the male population is more affected by lung cancer and other cancers, which can be attributed to the often negligent behavior of males towards maintaining a healthy lifestyle[31,32].

In Fig. 5, we illustrate the distribution of patients diagnosed with other cancer types. The distribution reveals heterogeneous demographic patterns across diagnoses, with head and neck cancers accounting for the largest proportion of cases, followed by melanoma.

These demographic patterns are consistent with real-world clinical populations and suggest that MedQARo captures representative patient characteristics, supporting the etiological validity of the benchmark for clinical question answering.

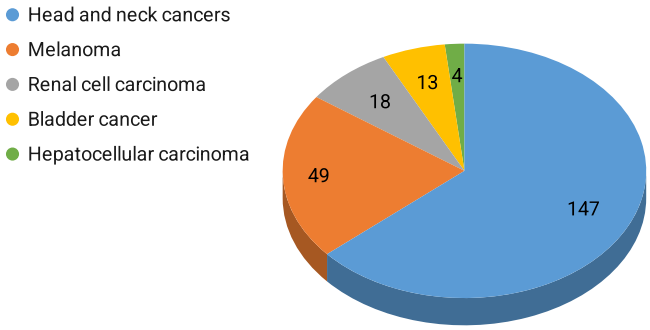

Fig. 5 | Distribution across cancer types for patients included in the cross-domain test set of MedQARo. Blue represents head and neck cancers. Orange represents melanoma. Gray represents renal cell carcinoma. Yellow represents bladder cancer. Green represents hepatocellular carcinoma.

Overview of large language models

To address the task of medical question answering in Romanian, we consider a diverse set of LLMs and use them under various configurations, considering both zero-shot and fine-tuning settings. Our goal is to understand how distinct architectural specializations (e.g. language-specific, domain-specific, or long-context models) and alternative prompt formulations affect performance on the Romanian clinical QA task. We select six distinct LLM families. Two of them are pretrained on Romanian, namely RoLLaMA2-7B[25], which is adapted from LLaMA 2[2], and RoMistral-7B[25], which is adapted from the Mistral architecture[3]. The third model, namely Phi-4-mini-instruct[26], is optimized for long-context inputs. This is useful for the processing of entire epicrises, which can often get to be very long (see Table 9), especially for patients with a long medical history. Given the target domain, the fourth model, LLaMA3-OpenBioLLM-8B[1], is pretrained on biomedical corpora. Finally, we evaluate two state-of-the-art LLMs, GPT-5.2[33] and Gemini 3 Flash[34], that are only accessible via public APIs.

The first four LLMs are employed in two setups, zero-shot prompting and end-to-end fine-tuning. Due to the large number of parameters in the chosen models, we adopt a parameter-efficient fine-tuning strategy based on Low-Rank Adaptation (LoRA)[35] specialized on the causal language modeling task. By comparing the results obtained in the two setups, we can precisely contextualize the impact of fine-tuning on MedQARo. In addition, we also test two alternative prompts, where the order between context (epicrisis) and question is interchanged. To determine the impact of zero-shot prompting, we include a random token selector and a majority answer model as baselines for the QA task. Since access to the architectures and weights of GPT-5.2 and Gemini 3 Flash is not available, we only employ these two models in the zero-shot setup. Next, we describe all evaluated methods in detail.

Romanian-Adapted LLMs

We start from two instruction-tuned language models specifically adapted for Romanian, RoLLaMA2-7B-Instruct and RoMistral-7B-Instruct. Both models are part of the recent efforts to build high-quality Romanian LLMs by further pretraining on Romanian corpora and applying instruction tuning with Romanian datasets, such as RoAlpaca[4,25], RoOpenAssistant[5,25] and RoOrca[25,36]. The two LLMs have been optimized for Romanian morphology and syntax using tokenizers adapted to the language. They have shown strong performance on Romanian natural language understanding and generation tasks[25]. In our study, we further fine-tune both models on MedQARo and compare their performance under different prompt structures and lengths.

As it inherits the architectural constraints of the original LLaMA 2 model[2], RoLLaMA2-7B is limited to a context window of 4096 tokens. We employ a LoRA configuration specialized on the causal language modeling task. The chosen configuration enables updating a small subset of the model parameters (0.0622%) during fine-tuning, significantly reducing training overhead. To enhance the adaptability of the LLM to the clinical QA task, we further consider several configuration alternatives. First, we evaluate on various prompt lengths, considering limits of 1024, 2048, and 4096 tokens. If an epicrisis is longer than the accepted input length, we simply trim it from the end. When evaluating on prompts of 2048 tokens, we also explore two prompt formats: `Q+E+A` (question + epicrisis + answer) and `E+Q+A` (epicrisis + question + answer). The model is trained via token-level cross-entropy, where the task is to predict the answer tokens via next token prediction. For this reason, the answer is always placed at the end of the prompt.

Given the large number of parameters in the base Mistral model, we again employ a LoRA-based fine-tuning strategy, enabling updates to only a small number of parameters. In this case, we fine-tune just 0.0470% of the parameters, using the same LoRA configuration as for RoLLaMA2-7B-Instruct. This approach ensures both memory efficiency and effective domain adaptation to the Romanian clinical QA setting. We conduct experiments with two prompt lengths, namely 2,048 and 4,096. Consistent with the preliminary empirical results obtained with the previous model, the prompt format is set to `Q+E+A` in all the experiments performed with RoMistral-7B-Instruct. We employ token-level cross-entropy to optimize the model.

Long-context LLM

Phi-4-mini-instruct is a compact instruction-tuned transformer model designed for long-context understanding, offering efficient inference for up to 128K tokens. We employ this model primarily due to its ability to process significantly longer input sequences than RoLLaMA2-7B-Instruct and RoMistral-7B-Instruct, respectively. Given that a substantial proportion of the epicrises exceed the 4096-token limit (see Table 9), we hypothesize that Phi-4-mini-instruct could offer competitive performance by leveraging a broader context, despite being a smaller model, and without explicit pretraining on the Romanian language.

To adapt the model to the clinical QA setting, we fine-tune it using LoRA, updating only 0.0956% of the parameters via token-level cross-entropy. We employ the same LoRA configuration as for RoLLaMA2-7B-Instruct and RoMistral-7B-Instruct. Thanks to the architectural support of Phi-4 for extended input lengths, we evaluate the model under increasingly longer prompts, namely 2048, 3072, 4096, 8192, 16,384, and 32,768 tokens. This allows us to assess the relationship between available context and model performance, particularly in scenarios where clinical answers are distributed across epicrises of patients with a long medical history. All prompts follow the `Q+E+A` format.

Biomedical LLM

LLaMA3-OpenBioLLM-8B is a domain-specialized model designed to address biomedical and clinical tasks. It is based on the LLaMA3 architecture and has undergone extensive pretraining on a diverse collection of biomedical corpora, including high-quality resources such as PubMed abstracts, full-text clinical trials, and biomedical textbooks. Through this targeted pretraining, the model achieves a strong understanding of medical terminology, domain-specific reasonin,g and structured clinical knowledge.

Despite being pretrained primarily in English, LLaMA3-OpenBioLLM-8B can still offer a valuable perspective when evaluated on Romanian clinical question answering. To adapt the model to our Romanian QA task, we train the model via LoRA (with the same configuration as for the other LLMs), updating only 3,407,872 parameters, which corresponds to 0.0424% of the full model size. We use the `Q+E+A` prompt format and explore two different prompt lengths, 2048 and 4096.

Baseline models

To assess the effectiveness of our fine-tuned models and to better isolate the impact of domain and task adaptation, we include four baselines:

The **random token selector** is a baselines that randomly selects tokens for the input epicrisis. For this baseline, we run three versions with different input lengths (1024, 2048, and 4096 tokens). For each setup, we randomly

select the starting token index (between 0 and the maximum input length) and a span length (between 1 and 3 tokens) to serve as the predicted answer. The answer is extracted from the input context (epicrisis), while entirely ignoring the question. This baseline serves as a trivial reference point, highlighting the complexity of the clinical QA task and providing a lower-bound performance estimate.

The **majority answer model** is another simple baseline is based on answering with the most prevalent answer in the training set. This model consistently answers with "Nu" ("No"), regardless of the posed question or the epicrisis content. While still trivial, the majority answer model can be a more competitive baseline than the random token selector.

For each of the open-source LLMs (RoLLaMA2-7B-Instruct, RoMistral-7B-Instruct, Phi-4-mini-instruct, and LLaMA3-OpenBioLLM-8B), we assess performance without any task-specific fine-tuning. We use the `Q+E+A` prompt format across all models. The **zero-shot inference** setup is meant to reflect the out-of-the-box capabilities of distinct types of LLMs.

In addition to the open-source models, we evaluate two state-of-the-art LLMs that are only accessible through commercial APIs, namely GPT-5.2 and Gemini 3 Flash. We refer to these models as **zero-shot API-based LLMs**. GPT-5.2 is designed to effectively process long contexts, have strong reasoning capabilities, and support multiple languages. Its reasoning ability is enforced via reinforcement learning, while the multilingual support is ensured by the large-scale pretraining on public data. Gemini 3 Flash is an LLM constructed on top of Gemini 3 Pro, inheriting its multimodal and reasoning capabilities. Its architecture is based on a sparse mixture-of-experts (MoE), which is able to activate only a subset of parameters (experts) per input token. This is achieved by learning to dynamically route tokens to a subset of experts. The thinking levels of Gemini 3 Flash are configured to strike a balance between quality, cost and latency. For both models, we use the same prompt format as before, namely `Q+E+A`, to ensure a fair comparison with the other models. To control evaluation costs, we compare GPT-5.2 and Gemini 3 Flash with the other LLMs on a randomly sampled subset of 100 test instances. The comparative empirical study is intended to provide a reference point for the zero-shot capabilities of current state-of-the-art LLMs on Romanian clinical question answering.

Data availability

The dataset has been made publicly available for non-commercial use under the CC BY-NC-SA 4.0 license agreement, at https://github.com/ana-rogoz/MedQARo.

Code availability

The code has been made publicly available for non-commercial use under the CC BY-NC-SA 4.0 license agreement, at https://github.com/ana-rogoz/MedQARo.

Received: 30 October 2025; Accepted: 12 February 2026;
Published online: 21 February 2026

References

1. Pal, A. & Sankarasubbu, M. OpenBioLLMs: Advancing Open-Source Large Language Models for Healthcare and Life Sciences. *Hugging Face Repository*. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B (2024).
2. Touvron, H. et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971* (2023).
3. Jiang, A. Q. et al. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).
4. Taori, R. et al. Stanford Alpaca: An Instruction-following LLaMA model. GitHub repository. https://github.com/tatsu-lab/stanford_alpaca (2023).
5. Köpf, A. et al. OpenAssistant conversations – Democratizing large language model alignment. In Proceedings of the 37th International Conference on Neural Information Processing Systems, 47669–47681. https://proceedings.neurips.cc/paper_files/paper/2023/hash/949f0f8f32267d297c2d4e3ee10a2e7e-Abstract-Datasets_and_Benchmarks.html (Curran Associates Inc., 2023).
6. Brown, T. et al. Language models are few-shot learners. In Proceedings of the 34th International Conference on Neural Information Processing Systems, vol. 33, 1877–1901. https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html (Curran Associates Inc., 2020).
7. Chowdhery, A. et al. PaLM: Scaling language modeling with pathways. *J. Mach. Learn Res*. **24**, 1–113 (2023).
8. Rajpurkar, P., Zhang, J., Lopyrev, K. & Liang, P. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392 (Association for Computational Linguistics, Austin, Texas, 2016).
9. Hendrycks, D. et al. Measuring massive multitask language understanding. In Proceedings of the International Conference on Learning Representations. https://openreview.net/forum?id=d7KBjmI3GmQ (2021).
10. Achiam, J. et al. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).
11. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
12. Singhal, K. et al. Towards expert-level medical question answering with large language models. *Nat. Med.* **31**, 943–950 (2025).
13. Kwiatkowski, T. et al. Natural questions: A benchmark for question answering research. *Trans. Assoc. Comput Linguist* **7**, 452–466 (2019).
14. Artetxe, M., Ruder, S. & Yogatama, D. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4623–4637 (Association for Computational Linguistics, Online, 2020).
15. Lewis, P., Oguz, B., Rinott, R., Riedel, S. & Schwenk, H. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7315–7330 (Association for Computational Linguistics, Online, 2020).
16. Clark, J. H. et al. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Trans. Assoc. Comput Linguist* **8**, 454–470 (2020).
17. Jin, D. et al. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Appl. Sci.* **11**, 6421 (2021).
18. Pampari, A., Raghavan, P., Liang, J. & Peng, J. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2357–2368 (Association for Computational Linguistics, Brussels, Belgium, 2018).
19. Jin, Q., Dhingra, B., Liu, Z., Cohen, W. & Lu, X. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2567–2577 (Association for Computational Linguistics, Hong Kong, China, 2019).
20. Pal, A., Umapathi, L. K. & Sankarasubbu, M. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. In *Proceedings of the Conference on Health, Inference, and Learning*, vol. 174, 248–260 (PMLR, 2022). https://proceedings.mlr.press/v174/pal22a.html.
21. Crăciun, C.-G., Smădu, R.-A., Cercel, D.-C. & Cercel, M.-C. GRAF: Graph retrieval augmented by facts for Romanian legal multi-choice question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, 12708–12742 (Association for Computational Linguistics, Vienna, Austria, 2025).
22. Nicolae, D. C. & Tufiş, D. RoITD: Romanian IT Question Answering Dataset. In *Proceedings of the 16th International Conference on Lingustic Resources and Tools for Natural Language Processsing*,

105–117. https://conferences.info.uaic.ro/consilr/prevEditions/Consilr_2021.pdf (2021).
23. Dumitrescu, S. D. et al. LiRo: Benchmark and leaderboard for Romanian language tasks. In Proceedings of the 35th Conference on Neural Information Processing Systems. https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/5f93f983524def3dca464469d2cf9f3e-Abstract-round1.html (Curran Associates, Inc., 2021).
24. Dima, G.-A., Avram, A.-M., Craciun, C.-G. & Cercel, D.-C. RoQLlama: A lightweight Romanian adapted language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 4531–4541 (Association for Computational Linguistics, Miami, Florida, USA, 2024).
25. Masala, M. et al. "Vorbeşti Româneşte?" A Recipe to Train Powerful Romanian LLMs with English Instructions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 11632–11647 (Association for Computational Linguistics, Miami, Florida, USA, 2024).
26. Abouelenin, A. et al. Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743* (2025).
27. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318 (Association for Computational Linguistics, 2002).
28. Banerjee, S. & Lavie, A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, 65–72 (Association for Computational Linguistics, Ann Arbor, Michigan, 2005). https://aclanthology.org/W05-0909/.
29. Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization. In *Proceedings of the International Conference on Learning Representations.* https://openreview.net/forum?id=Bkg6RiCqY7 (2019).
30. Barbero, F. et al. Why do LLMs attend to the first token? In *Proceedings of Conference on Language Modeling.* https://openreview.net/forum?id=tu4dFUsW5z (2025).
31. Bonhomme, J. J. Men's health: impact on women, children and society. *J. Mens. Health Gend.* **4**, 124–130 (2007).
32. Dong, X., Simon, M. A. & Evans, D. A. Prevalence of self-neglect across gender, race, and socioeconomic status: Findings from the Chicago Health and Aging Project. *Gerontology* **58**, 258–268 (2012).
33. OpenAI. Update to GPT-5 System Card: GPT-5.2. *Technical Report.* https://cdn.openai.com/pdf/3a4153c8-c748-4b71-8e31-aecbde944f8d/oai_5_2_system-card.pdf (2025).
34. Google. Gemini 3 Flash - Model Card. *Technical Report.* https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf (2025).
35. Hu, E. J. et al. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the International Conference on Learning Representations.* https://openreview.net/forum?id=nZeVKeeFYf9 (2022).
36. Mukherjee, S. et al. Orca: Progressive learning from complex explanation traces of GPT-4. *arXiv preprint arXiv:2306.02707* (2023).

Acknowledgements

This research is supported by the project "Romanian Hub for Artificial Intelligence—HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 351416. This work was also supported by a grant of the Ministry of Research, Innovation, and Digitization, CCCDI—UEFISCDI, project number PN-IV-P6-6.3-SOL-2024-0090, within PNCDI IV.

Author contributions

A.C.R. performed the experiments and wrote the initial draft. R.T.I. designed the study and revised the manuscript. A-.V.A. and I-.L.A.-I. participated in data collection and curation and wrote the initial draft. S.C. and A.I.I. participated in data collection and curation, and revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Radu Tudor Ionescu.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026