Artificial Intelligence In Medicine 155 (2024) 102938

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Research Paper

MedExpQA: Multilingual benchmarking of Large Language Models for Medical Question Answering

Iñigo Alonso, Maite Oronoz, Rodrigo Agerri *

HiTZ Center - Ixa, University of the Basque Country UPV/EHU, Spain

ARTICLE INFO

Keywords:
Large Language Models
Medical Question Answering
Multilinguality
Retrieval Augmented Generation
Natural Language Processing

ABSTRACT

Large Language Models (LLMs) have the potential of facilitating the development of Artificial Intelligence technology to assist medical experts for interactive decision support. This potential has been illustrated by the state-of-the-art performance obtained by LLMs in Medical Question Answering, with striking results such as passing marks in licensing medical exams. However, while impressive, the required quality bar for medical applications remains far from being achieved. Currently, LLMs remain challenged by outdated knowledge and by their tendency to generate hallucinated content. Furthermore, most benchmarks to assess medical knowledge lack reference gold explanations which means that it is not possible to evaluate the reasoning of LLMs predictions. Finally, the situation is particularly grim if we consider benchmarking LLMs for languages other than English which remains, as far as we know, a totally neglected topic. In order to address these shortcomings, in this paper we present MedExpQA, the first multilingual benchmark based on medical exams to evaluate LLMs in Medical Question Answering. To the best of our knowledge, MedExpQA includes for the first time reference gold explanations, written by medical doctors, of the correct and incorrect options in the exams. Comprehensive multilingual experimentation using both the gold reference explanations and Retrieval Augmented Generation (RAG) approaches show that performance of LLMs, with best results around 75 accuracy for English, still has large room for improvement, especially for languages other than English, for which accuracy drops 10 points. Therefore, despite using state-of-the-art RAG methods, our results also demonstrate the difficulty of obtaining and integrating readily available medical knowledge that may positively impact results on downstream evaluations for Medical Question Answering. Data, code, and fine-tuned models will be made publicly available.[1]

1. Introduction

We are currently seeing a dramatic increase in research on how to apply Artificial Intelligence (AI) to the medical domain with the aim of generating decision support tools to assist medical experts in their everyday activities. This has been further motivated by rather strong claims about Large Language Models (LLMs) in medical Question Answering (QA) tasks, such as that they obtain passing marks for medical licensing exams like the United States Medical Licensing Examination (USMLE) [1,2].

Assisting medical experts by answering their medical questions is a natural way of articulating human-AI interaction as it is usually considered that Medical QA involves processing, acquiring and summarizing relevant information and knowledge and then reasoning about how to apply the available knowledge to the current context given by a clinical case. For example, a resident medical doctor preparing for the licensing exams may want to know what and why is the correct treatment or diagnosis in the context of a clinical case [3,4]. This means that a LLM should be able to automatically identify, access and correctly apply the relevant medical knowledge, and that it will be capable of elucidating between the variety of symptoms, each of which may be indicative of multiple diseases. Finally, it is also assumed that the model will interact with the resident medical doctor in a natural manner, ideally using natural language. Therefore, developing the required AI technology to help, for example, resident medical doctors to prepare their licensing exams remains a far from trivial endeavour.

Nonetheless, and as a crucial first step to address this challenge, the AI ecosystem has seen an explosion of LLMs (both general purpose and specific to the medical domain) reporting high accuracy results on Medical QA tasks thereby demonstrating that LLMs are somewhat capable of encoding clinical knowledge [1]. State-of-the-art models include

* Corresponding author.
E-mail addresses: inigoborja.alonso@ehu.eus (I. Alonso), maite.oronoz@ehu.eus (M. Oronoz), rodrigo.agerri@ehu.eus (R. Agerri).

[1] https://huggingface.co/datasets/HiTZ/MedExpQA.

https://doi.org/10.1016/j.artmed.2024.102938
Received 1 April 2024; Received in revised form 8 June 2024; Accepted 26 July 2024
Available online 31 July 2024

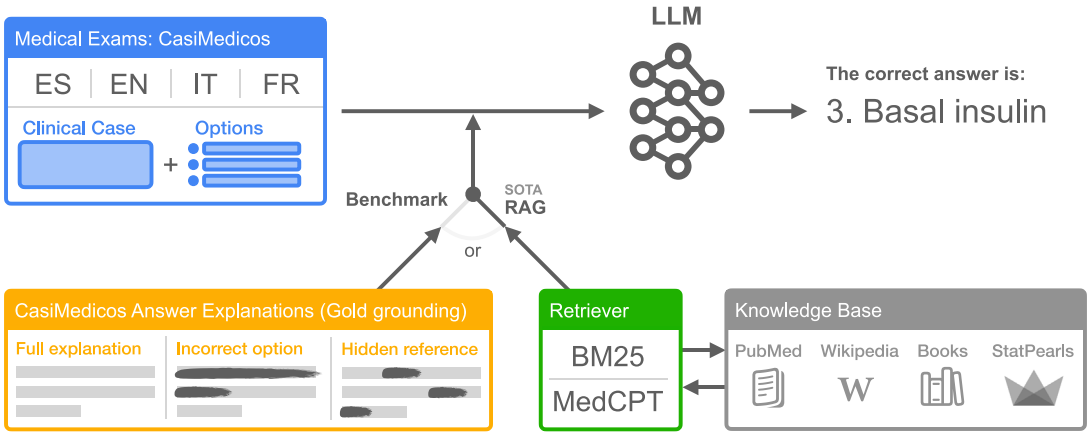


Fig. 1. Graphical description of the MedExpQA benchmark in which various types of gold and external medical knowledge are added to Large Language Models in order to find the correct answer in the CasiMedicos dataset.

publicly available ones such as LLaMA [5] and the medical-specific PMC-LLaMA [6], Mistral [7] and its medical version BioMistral [8], and proprietary models such as MedPaLM [9] and GPT-4 [2], among many others.

While their published high-accuracy scores on Medical QA may seem impressive, these LLMs still present a number of shortcomings. First, LLMs usually generate factually inaccurate answers that seem plausible enough for a non-medical expert (known as hallucinations) [10,11]. Second, their knowledge might be outdated as the pre-training data used to train the LLMs may not include the latest available medical knowledge. Third, the Medical QA benchmarks [1, 11] on which they are evaluated do not include gold reference explanations generated by medical doctors that provide the required reasoning to support the model's predictions. Finally, and to the best of our knowledge, evaluations have only been done for English, which makes it impossible to know how well these LLMs fare for other languages.

Retrieval Augmented Generation (RAG) techniques have been specifically proposed to address the first two issues, namely, the lack of up-to-date medical knowledge and the tendency of these models to hallucinate [11]. Their MedRAG approach obtains clear zero-shot improvements for two of the five datasets on their MIRAGE benchmark, while for the rest the obtained gains are rather modest. Still, MedRAG proves to be an effective technique to improve Medical QA by incorporating external medical knowledge [11].

In this paper we present MedExpQA (Medical Explanation-based Question Answering), which is, to the best of our knowledge, the first multilingual benchmark for Medical QA. Furthermore, and unlike previous work, our new benchmark also includes gold reference explanations to justify why the correct answer is correct and also to explain why the rest of the options are incorrect. Written by medical doctors, these high-quality explanations help to assess the model's decisions based on complex medical reasoning. Moreover, our MedExpQA benchmark leverages the reference explanations as *gold knowledge* to establish various upperbounds for comparison with results obtained when applying automatic MedRAG methods. By doing so, we aim to address all four shortcomings of LLMs for Medical QA listed above.

Although by design independent of the specific source data used, for this work we leverage the Antidote CasiMedicos dataset [4,12], which consist of Resident Medical Exams or *Médico Interno Residente* in Spanish, an exam similar to other licensing examinations such as USMLE, to setup MedExpQA. In addition to a short clinical case, a question and the multiple-choice options, CasiMedicos includes gold reference explanations regarding both the correct and incorrect options. Originally in Spanish, CasiMedicos was translated and annotated in English, French and Italian [4].

Fig. 1 provides an overview of the MedExpQA benchmark. Taking CasiMedicos as the data source, the basic input, without any additional

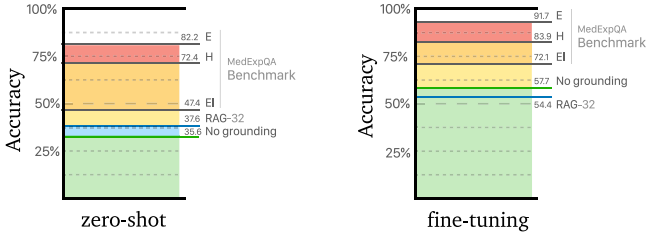


Fig. 2. Overview of averaged results in MedExpQA for gold and automatically knowledge grounding based on Retrieval Augmented Generation (RAG). *E*: gold explanations written by medical doctors; *H*: E with explicit references to the possible answers hidden; and *EI*: gold explanations about the incorrect options; *RAG-32*: automatically retrieved knowledge grounding (details in Section 5); *no-grounding*: baseline model with no external knowledge.

knowledge, to the LLM consists of a clinical case and the multiple-choice options. Furthermore, the model can also be provided with three types of gold reference explanations (or gold knowledge grounding) extracted from the CasiMedicos explanations: (i) the full gold explanation as written by the medical doctors; (ii) only the explanations regarding the incorrect answers and, (iii) the full gold explanation with explicit references to the possible answers hidden. Finally, we can also apply automatic knowledge retrieval approaches such as MedRAG to provide LLMs with automatically obtained up-to-date medical knowledge. Thus, in MedExpQA it is possible to compare not only whether the MedRAG methods improve over the basic input with no external knowledge added, but also to establish the differences in performance of LLMs (with or without RAG) with respect to results obtained when gold reference explanations are available. An additional benefit of MedExpQA being multilingual is that we get to compare LLMs performance not only for English, but also on popular languages such as Spanish, French or Italian.

Fig. 2 shows that comprehensive multilingual experimentation on MedExpQA using four state-of-the-art LLMs including LLaMA [5] PMC-LLaMA [6], Mistral [7] and BioMistral [8], demonstrate that LLMs performance, even when improved with external knowledge from MedRAG (corresponding to RAG-32 in Fig. 2), still has a long way to go to get closer to the performance obtained when the external knowledge available to the LLM is based on gold reference explanations (*E* and *H* in Fig. 2). Another interesting point is that fine-tuning results in huge performance increases across settings and models but at the cost of making MedRAG redundant. In other words, MedRAG only has a positive impact in zero-shot settings. We believe that this illustrates the difficulty of automatically retrieving and integrating readily available knowledge in a way that may positively impact final downstream results on Medical

QA. Finally, results are substantially lower for French, Italian and Spanish, which suggests that more work is needed to improve LLMs performance for languages different to English. Summarizing, the main contributions of our work are the following:

1. MedExpQA: the first multilingual benchmark for MedicalQA including gold reference explanations.
2. Comprehensive study on the role of medical knowledge to answer medical exams by leveraging gold reference explanations written by medical doctors as upper bound with respect to automatically retrieved knowledge using state-of-the-art RAG techniques.
3. Experimental results demonstrate that fine-tuning clearly outperforms querying the LLMs in zero-shot, making redundant the external knowledge obtained via RAG.
4. Overall performance of LLMs with or without RAG still has large room for improvement when compared with any of the results obtained using gold reference explanations.
5. Performance for French, Italian and Spanish substantially lower for every LLM in every evaluation setting, which stresses the urgent need of advancing the state-of-the-art for Medical QA in languages different to English.
6. Data, code and fine-tuned models available to facilitate reproducibility of results and benchmarking of LLMs in the medical domain[2].

In the rest of the paper we first discuss the related work and then in Section 3 we describe the Large Language Models (LLM) and the Retrieval Augmented Generation method used for experimentation. Section 4 provides a description of the MedExpQA benchmark, including the Antidote CasiMedicos dataset. The experimental setup is explained in Section 5 and results are reported in Section 6. Section 7 offers a discussion of the main issues raised by the empirical results obtained. We finish with some concluding remarks and future work in Section 8.

2. Related work

We are currently seeing a vertiginous rhythm in the development of Large Language Models (LLMs) which is having a great impact on Natural Language Processing for the medical domain. This is particularly true on Medical Question Answering tasks where LLMs have been successfully applied to generate answers to highly specialized medical questions. Thus, the performance improvements on Abstractive Medical Question Answering of general purpose LLMs such GPT-4 [2] and GPT-3 [13], PaLM [14], LLaMa [5] or Mistral [7], has resulted in a huge interest to adapt or to generate LLMs specialized for medical text processing.

Some of these models are based on the encoder–decoder architecture, such as SciFive [15], and English T5 model adapted to the scientific domain, or Medical-mT5, a multilingual model built by fine-tuning mT5 on a multilingual corpus of 3B tokens [16]. However, the large majority of the LLMs specially generated for medical applications are autorregresive decoder models such as BioGPT [17], ClinicalGPT [18], Med-PaLM [1], MedPaLM-2 [9], PMC-LLaMA [6], and more recently, BioMistral [8].

These models have been reporting high-accuracy scores on various medical QA benchmarks, which generally consist of exams or general medical questions. Several of the most popular Medical QA datasets [19–24] have been grouped into two multi-task English benchmarks, namely, MultiMedQA [1] and MIRAGE [11] with the aim of providing an easier comprehensive experimental evaluation benchmark of LLMs for Medical QA.

Despite recent improvements on these benchmarks that had led to claims about the capacity of LLMs to encode clinical knowledge [1], these models remain hindered by well known issues related to: (i) their tendency to generate plausible-looking but factually inaccurate answers and, (ii) working with outdated knowledge as their pre-training data may not be up-to-date to the latest available medical progress; (iii) the large majority of these benchmarks do not include gold reference explanations to help evaluate the reasoning capacity of LLMs to predict the correct answers; (iv) they have mostly been developed for English, which leaves a huge gap regarding the evaluation of the abilities of LLMs for other languages.

Regarding the first issue listed above, it should be considered that these LLMs are not restricted to the input context to generate the answer as they are able to produce word by word output by using their entire vocabulary in an auto-regressive manner [25]. This often results in answers that are apparently plausible and factually correct, when in fact they are not always factually reliable. With respect point (ii), while LLMs are pre-trained with large amounts of texts, they may still lack the specific knowledge required to answer highly specialized questions or it may simply be in need of an update.

Recent work [26] has proposed Retrieval Augmented Generation (RAG) [27] to mitigate these limitations. This method involves incorporating relevant external knowledge into the input of these LLMs with the aim of improving the final generation. By doing so, it increases the probability of generated responses being grounded in the automatically retrieved evidence, thereby enhancing the accuracy and quality of the output. Some of the most common retrieval methods employed include TF-IDF, BM25 [28], and others more specific to the medical domain such as MedCPT [29]. With the aim of providing an exhaustive evaluation of RAG methods for the medical domain, the MIRAGE benchmarch includes 5 well-known English Medical QA datasets which are used to compare zero-shot performance of various LLMs whenever automatically retrieved knowledge is available via their MedRAG method or in the absence of it. According to the authors, MedRAG not only helps to address the problem of hallucinated content by grounding the generation on specific contexts, but it also provides relevant up-to-date knowledge that may not be encoded in the LLM [11]. By employing MedRAG they are able to clearly improve the zero-shot results of some of the LLMs tested, although for others results are rather mixed.

Finally, and to the best of our knowledge, no Medical QA benchmark currently addresses the last two shortcomings, namely, the lack of gold reference explanations and multilinguality. Motivated by this, we propose MedExpQA, a multilingual benchmark including gold reference explanations written by medical doctors that can be leveraged to setup various upperbound results to be compared with the performance of LLMs enhanced by automatic RAG methods.

3. Materials and methods

In this section we describe the main resources used in our experimentation with MedExpQA, namely, the Large Language Models (LLMs) tested on our benchmark and MedRAG, the Retrieval Augmented Generation method proposed by Xiong et al. [11] to automatically retrieve medical knowledge.

3.1. Models

We selected two open source state-of-the-art LLMs in the MedicalQA domain at the time of writing: PMC-LLaMA [6] and BioMistral [8].

PMC-LLaMA is based on LLaMA [5], one of the most popular LLMs currently available. PMC-LLaMA is an open-source language model specifically designed for medical applications. This model was first pre-trained on a combination of PubMed-related English academic papers from the S2ORC corpus [30] and from medical textbooks. It was then further fine-tuned on a dataset of instruction-based medical texts. For our experiments we pick the 13B parameter variant of this model which

[2] https://huggingface.co/datasets/HiTZ/MedExpQA

outperforms LLaMA-2 [5], Med-Alpaca [31], and Chat-Doctor [32] in various Medical QA tasks including MedQA [23], MedMCQA [24], and PubMedQA [19].

BioMistral [8] is a suite of open-source models based on Mistral [7] further pre-trained using English textual data from PubMed Central Open Access[3]. They released a set of 7b parameter models following merging techniques like TIES [33], DARE [34], and SLERP [35]. In this paper we use the DARE variant of BioMistral as it is the best performing model on the MedQA benchmark, outscoring other state-of-the-art LLMs on Medical QA evaluations, including PMC-LLaMA.

Additionally, and in order to contrast their performance against their general purpose counterparts, we also test LLaMA-2 and Mistral. Thus, for both PMC-LLaMa and LLaMA-based models we use the 13 billion parameter variants. As BioMistral is only available in the 7b version, we also pick the Mistral model of 7b parameters.

Every zero-shot and fine-tuning experiment with LLMs are performed via the HuggingFace API [36].

3.2. Retrieval-augmented generation (RAG)

We apply MedRAG as the Retrieval-Augmented Generation (RAG) state-of-the-art technique especially developed for the medical domain [11]. RAG approaches are mostly composed of three components: the LLM, the retrieval method and the data source from which to retrieve the knowledge. MedRAG includes four retrievers and four different corpora as data sources.

With respect the retrievers, we use both BM25 [28] and MedCPT [29] to perform the retrieval and fuse the retrieved candidate lists into one using Reciprocal Rank Fusion (RRF) [37]. BM25 is a ranking function used in Information Retrieval to rank documents based on their relevance to a given query. It combines Term Frequency (TF) and Inverse Document Frequency (IDF) to calculate the relevance score of a document to a query taking into account the document length for normalization. MedCPT is a Contrastive Pre-trained Transformer model trained with PubMed search logs for zero-shot biomedical information retrieval. This model retrieves the relevant documents in the knowledge base considering relationships between different medical entities and concepts in the query.

Regarding the data sources, we use MedCorp, a combination of the four corpora available in MedRAG: PubMed, Textbooks [23] for domain-specific knowledge, StatPearls[4] for clinical decision support, and Wikipedia for general knowledge. According to the MIRAGE results [11], using MedCorp was the only realistic option for MedRAG to systematically improve results over the baseline for most of the LLMs and retriever methods evaluated.

4. MedExpQA: A new multilingual benchmark for medical QA

Although independently designed with respect to any specific dataset, in this paper we setup MedExpQA, introduced in Section 4.2, on the Antidote CasiMedicos dataset [4,12], which is described in detailed in Section 4.1.

4.1. Antidote CasiMedicos dataset

Every year the Spanish Ministry of Health releases the previous year's Resident Medical exams or *Médico Interno Residente* (MIR) which, as depicted in Table 1, include a clinical case (**C**), the multiple choice options (**O**), and the correct answer (**A**). The MIR exams are then commented every year by the CasiMedicos MIR Project 2.0[5] which means that CasiMedicos medical doctors voluntarily write gold reference explanations (full gold explanation **E** in Table 1) providing reasons for both correct (**EC**) and incorrect options (**EI**).

The Antidote CasiMedicos dataset [4,12] consists of the original Spanish commented exams which were cleaned, structured and manually annotated to link the relevant textual parts in the gold reference explanation (**E**) with the correct (**EC**) or incorrect options (**EI**). Once the Spanish version of the dataset was created, parallel translated annotated versions were generated for English, French, and Italian.

A quantitative description of the multilingual Antidote CasiMedicos dataset is given in Table 2. The average number of tokens in the clinical cases is 137, being quite similar for Spanish and Italian (140.3 and 142.2 respectively), while for English the average is smaller (115.4 tokens) while the French one is the largest (150.1 tokens). The average length in tokens of the multiple choice options (79.6 tokens in average) is quite high but with a high variability. The multiple choice options may consist of short drug names (the minimum number of words is around 15–17) to long descriptions of treatments or medical claims as illustrated by the example shown in Table 3. The full gold reference explanations that professional medical doctors write can be quite long (170.25 tokens in average) but it should be noted that some documents lack the explanation about the correct answer.

The complexity of some of the clinical case questions can be appreciated in the example shown in Table 3 where the possible answers (section **O**) describe disorders (option (1)), treatments (options (2) and (3)) or medical statements (options (4) and (5)). Furthermore, while in the majority of the cases the question is about the correct answer, sometimes the required option is the incorrect one, as shown in Tables 1 and 3.

The final Antidote CasiMedicos Dataset consists of 622 documents per language [4,12]. The dataset official distribution already provide train, validation and test splits[6] (depicted in Table 4), which we use for the all the experiments presented in Section 6.

Finally, we examined the distribution of correct answers in each of the three splits (train, validation and test) to consider the possibility that an unbalanced distribution might condition the results of the tested models. Fig. 3 shows that, although most of the exams have the option 3 as the correct answer, the distribution among the correct answers in the three subsets is quite balanced. This suggests that this particular issue should not influence the final experimental results.

4.2. The MedExpQA benchmark

MexExpQA is a multilingual benchmark to evaluate LLMs in Medical Question Answering. Unlike previous work, MedExpQA includes reference gold explanations written by medical doctors which are leveraged to setup a benchmark with three types of gold knowledge: (i) the full gold reference explanation (part **E** in Table 1); (ii) the full gold reference explanation corresponding to the incorrect options only (**EI**) and (iii), the full gold reference explanation masking the explicit references in the text to the multiple-choice options.

In other words, and as illustrated in Fig. 1, we use these three types of high-quality explanations written by medical doctors as a proxy of relevant gold knowledge that may be used by LLMs to answer medical questions. Thus, the results obtained by LLMs with each type of gold knowledge can be seen as the upperbound results provided by our benchmark to establish how well LLMs can perform according to the different types of specialized gold knowledge readily available. In the following we describe in detail each of the three types of gold reference explanations that we generate to setup our benchmark.

[3] PMC Open Access Subset. Available from https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

[4] https://www.statpearls.com/

[5] https://www.casimedicos.com/mir-2-0/

[6] https://huggingface.co/datasets/HiTZ/casimedicos-exp

Table 1
Document in the Antidote CasiMedicos dataset with the correct and incorrect explanations manually annotated. C: Clinical case and question; O: Multiple-choice options; A: Correct answer; E: Full gold reference explanation written by medical doctors; EC: Explanation about the correct answer; EI: Explanation about the incorrect answers.

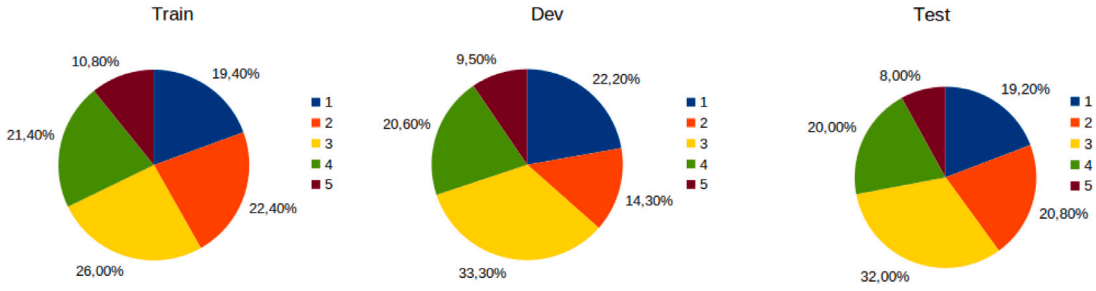
C	30-year-old man with no past history of interest. He comes for consultation due to the presence of small erythematous-violaceous lesions that on palpation appear to be raised in the pretibial region. The analytical study shows a complete blood count and coagulation study without alterations, and in the biochemistry, creatinine and ions are also within the normal range. The urinary sediment study shows hematuria, for which the patient had already been studied on other occasions, without obtaining a definitive diagnosis. Regarding the entity you suspect in this case, it is FALSE that
O	(1) In 20 to 50% of cases there is elevation of serum IgA concentration. (2) In the renal biopsy the mesangial deposits of IgA are characteristic. (3) It is frequent the existence of proteinuria in nephrotic range. (4) It is considered a benign entity since less than 1/3 of patients progress to renal failure. (5) The cutaneous biopsy allows to establish the diagnosis in up to half of the cases.
A	**3**
E	They are talking to us with high probability of a mesangial IgA glomerulonephritis or Berger's disease. Therefore, we are going to discard options one by one: 1: True. Serum IgA elevation is found in up to 50% of cases. 2: True. Mesangial IgA deposits are present in almost 100% of cases. 3: This option is false, because this glomerulonephritis is classically manifested with nephritic and not nephrotic syndrome (although in some rare cases proteinuria in nephrotic range does appear, but in the MIR they do not ask about these rare cases). 4: At the beginning this option generated doubts in me, but looking in the literature, it is true that the evolution to renal failure (according to last series) occurs in about 25% of the cases, so this option is true. 5: Skin biopsy, because it is easier to perform than renal biopsy, is the diagnostic technique of choice (the skin lesions that constitute Schonlein-Henoch purpura, so frequently associated with this entity and which the patient in the case presents, are biopsied).
EC	3: This option is false, because this glomerulonephritis is classically manifested with nephritic and not nephrotic syndrome (although in some rare cases proteinuria in nephrotic range does appear, but in the MIR they do not ask about these rare cases).
EI	1: True. Serum IgA elevation is found in up to 50% of cases. 2: True. Mesangial IgA deposits are present in almost 100% of cases. 4: At the beginning this option generated doubts in me, but looking in the literature, it is true that the evolution to renal failure (according to last series) occurs in about 25% of the cases, so this option is true. 5: Skin biopsy, because it is easier to perform than renal biopsy, is the diagnostic technique of choice (the skin lesions that constitute Schonlein-Henoch purpura, so frequently associated with this entity and which the patient in the case presents, are biopsied).

Table 2
Quantitative description of the multilingual CasiMedicos dataset. Number of tokens in the clinical case including: the question (C), the multiple-choice options (O), the explanation about the correct answer (EC) and the full gold reference explanation (E) including argumentation about the correct and incorrect answers.

	Number of tokens	Average	Min	Max
Spanish	Clinical Case (C)	140.3 ± 62.4	41	504
	Multiple choice options (O)	77.0 ± 47.0	15	297
	Explanation about the correct (EC)	58.9 ± 37.7	0	483
	Full explanation (E)	174.1 ± 147.8	9	982
English	Clinical Case (C)	115.4 ± 52.8	34	419
	Multiple choice options (O)	64.7 ± 37.1	15	217
	Explanation about the correct (EC)	47.3 ± 30.4	0	382
	Full explanation (E)	139.1 ± 117.7	4	784
Italian	Clinical Case (C)	142.2 ± 64.5	35	539
	Multiple choice options (O)	79.0 ± 50.1	17	284
	Explanation about the correct (EC)	60.6 ± 38.4	0	500
	Full explanation (E)	179.1 ± 150.6	8	1013
French	Clinical Case (C)	150.1 ± 68.6	39	586
	Multiple choice options (O)	83.0 ± 52.8	16	319
	Explanation about the correct (EC)	63.9 ± 41.2	0	535
	Full explanation (E)	188.7 ± 158.9	8	1076
Avg. ALL	Clinical Case (C)	137		
	Multiple choice options (O)	79.6		
	Explanation about the correct (EC)	57.6		
	Full explanation (E)	170.25		

Table 3
Example of a document in the CasiMedicos dataset with very different types of response options. (1) diagnosis; (2) and (3) treatments; and (4) and (5) correspond to medical statements.

	Example of a document from the CasiMedicos Dataset
C	A 63-year-old woman comes to the emergency department reporting severe headache with signs of meningeal irritation, bilateral visual disturbances and ophthalmoplegia. A CT scan showed a 2 cm space-occupying lesion in the sella turcica compatible with pituitary adenoma with signs of intratumoral hemorrhage, with deviation of the pituitary stalk and compression of the glandular tissue. Mark which of the following answers is WRONG:
O	(1) Diagnostic suspicion is pituitary apoplexy. (2) Treatment with high-dose corticosteroids should be initiated and the evolution observed, since this treatment could reduce the volume of the lesion and avoid intervention. (3) Treatment with glucocorticoids should be considered to avoid secondary adrenal insufficiency that would compromise the patient's vital prognosis. (4) The presence of ophthalmoplegia and visual defects are indications for prompt intervention by urgent surgical decompression. (5) After resolution of the acute picture, the development of panhypopituitarism is frequent.
A	**4**



(a) Distribution of correct answers in the **train** dataset.

(b) Distribution of correct answers in the **validation** dataset.

(c) Distribution of correct answers in the **test** dataset.

Fig. 3. Distribution of correct answers in the train, validation and test splits. The percentage in blue indicates the proportion of exams with the first option, number 1, as correct answer; orange corresponds to option 2; yellow to option 3; green to option 4; and brown to option 5. Note that not every document includes 5 possible options. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Number of documents in CasiMedicos train, validation and test splits.

	Train	Validation	Test
Clinical cases	434	63	125
Total	622		

4.2.1. Full reference gold explanations

The full explanation (**E**) about the correct and incorrect answers is given as context to the LLM, in what we assume to be gold specific knowledge for the model to answer the medical questions of CasiMedicos. Being the full gold reference explanation, we consider this to be the best possible form of gold knowledge that we can provide the LLM with. In other words, the performance obtained in MedExpQA using this type of knowledge will mark the upperbound for this particular benchmark. Table 5 provides an example of the full gold reference explanation for the same document already discussed in Table 1.

4.2.2. Explanation of the incorrect options

As shown in Table 6, for this particular type of gold knowledge we only use the part of the full gold reference explanation corresponding to the explanations about the incorrect options (**EI**). This type gold knowledge aims to test the capacity of LLMs to correctly answer the medical question by knowing which options are incorrect.

Depending on the nature of the question, sometimes medical doctors consider sufficient to only explain the correct answer. Thus, it should be noted that not every document in CasiMedicos includes the gold reference explanations about the incorrect options. On average, 20.5% of the explanations correspond in their entirety to the correct answer (17.7% in the train set, and 22.2% and 21.6% in the validation and test, respectively), while 26.7 include the explanations for all the possible options. Obviously, as CasiMedicos is a multilingual parallel dataset, this phenomenon occurs across the four languages: English, French, Italian and Spanish.

4.2.3. Full gold explanation with explicit references hidden

As it can be appreciated in the full gold reference explanations discussed above, most of the time medical doctors provide explicit textual references regarding the correct or incorrect options. In order to analyze the impact of these explicit signals or patterns on the LLMs performance, we decided to mask those explicit references to establish how well LLMs could answer with actual gold knowledge but without the easy clues in the text pointing to the correct or incorrect answers.

In order to avoid the manual annotation of 2488 documents, we prompt GPT-4[7] [38] with a set of rules and in-context-learning examples to automatically mask the specific areas of text that may point the model at the correct or incorrect answer without any further reasoning. The prompt can be found in A, Fig. A.10.

A small manual analysis of a subset of GPT-4-generated texts revealed a strong correlation with human annotations. To further validate the efficacy of our method, we randomly selected 80 documents (20 per

[7] `gpt-4-1106-preview`

Table 5
Full explanation (E) of the example in Table 1. The explanation about the correct answer is marked in blue and the remaining 4 explanations for the incorrect options in green.

E	They are talking to us with high probability of a mesangial IgA glomerulonephritis or Berger's disease. Therefore, we are going to discard options one by one: 1: True. Serum IgA elevation is found in up to 50% of cases. 2: True. Mesangial IgA deposits are present in almost 100% of cases. 3: This option is false, because this glomerulonephritis is classically manifested with nephritic and not nephrotic syndrome (although in some rare cases proteinuria in nephrotic range does appear, but in the MIR they do not ask about these rare cases). 4: At the beginning this option generated doubts in me, but looking in the literature, it is true that the evolution to renal failure (according to last series) occurs in about 25% of the cases, so this option is true. 5: Skin biopsy, because it is easier to perform than renal biopsy, is the diagnostic technique of choice (the skin lesions that constitute Schonlein-Henoch purpura, so frequently associated with this entity and which the patient in the case presents, are biopsied).

Table 6
Explanation of the Incorrect Options (EI) which corresponds to the full explanation (E) of the example in Table 1 with the explanation of the correct answer removed.

EI	They are talking to us with high probability of a mesangial IgA glomerulonephritis or Berger's disease. Therefore, we are going to discard options one by one: 1: True. Serum IgA elevation is found in up to 50% of cases. 2: True. Mesangial IgA deposits are present in almost 100% of cases. 5: Skin biopsy, because it is easier to perform than renal biopsy, is the diagnostic technique of choice (the skin lesions that constitute Schonlein-Henoch purpura, so frequently associated with this entity and which the patient in the case presents, are biopsied).

Table 7
Full gold reference explanation with explicit references hidden (H). Process performed by GPT-4 with the prompt in A Fig. A.10. In this example the segments *'This option is false'*, *'so this option is true'* and *'is the diagnostic technique of choice'* are hidden.

H	They are talking to us with high probability of a mesangial IgA glomerulonephritis or Berger's disease. Therefore, we are going to discard options one by one: 1: True. Serum IgA elevation is found in up to 50% of cases. 2: True. Mesangial IgA deposits are present in almost 100% of cases. 3: [HIDDEN], because this glomerulonephritis is classically manifested with nephritic and not nephrotic syndrome (although in some rare cases proteinuria in nephrotic range does appear, but in the MIR they do not ask about these rare cases). 4: At the beginning this option generated doubts in me, but looking in the literature, it is true that the evolution to renal failure (according to last series) occurs in about 25% of the cases, [HIDDEN]. 5: Skin biopsy, because it is easier to perform than renal biopsy, is the [HIDDEN] (the skin lesions that constitute Schonlein-Henoch purpura, so frequently associated with this entity and which the patient in the case presents, are biopsied).

language) and measured performance across the four languages. This resulted in an average F1 score of 0.85 with a standard deviation of 0.02.

Thus, this method allowed us to perform this rather precise multilingual redacting process over the 2488 documents in a fast and cost effective manner. Table 7 shows how every explicit reference to the correct or incorrect answers discussed previously now appear as **[HIDDEN]**.

The results obtained by LLMs in MedExpQA using the three types of gold knowledge described above can then be compared with other automatic knowledge retrieval approaches based, for example, on Retrieval-Augmented Generation techniques for the medical domain such as MEDRAG, introduced in the previous section. Furthermore, we should stress that MedExpQA as a benchmark is independent of any dataset, as the only requirement is for it to include gold reference explanations of the possible answers.

5. Experimental setup

For our experiments we selected top performing state-of-the-art models for Medical Question Answering described in Section 3.1, namely, PMC-LLaMA, LLaMA-2, BioMistral, and Mistral.

We test these models in both zero-shot (see prompts in Figs. A.6–A.9) and fine-tuned settings to contrast their out-of-the-box performance against a more adjusted performance to our dataset. The models were fine-tuned using Low-Rank Adaptation (LoRA) [39], using adapters with a rank of 8 and a scaling factor (alpha) of 16 across all models (details about parameters used with LoRA are provided in C).

The choice of hyperparameters was based on previous work using the same LLMs we use in this papers. Moreover, satisfactory results were confirmed in a preliminary round of experiments. Although these models would benefit from an exhaustive grid search of hyperparameters tailored to each model and evaluation setting, the compute required to do so exceeds the capacity of our lab. Full details of hyperparameter settings are available in B. Each model was fine-tuned for 10 epochs, with checkpoints saved at the end of each. Experiments were undertaken in a NVIDIA A100 GPU (C offers information about computation times). At the end of the fine-tuning process, the checkpoint with the highest performance was selected. All models underwent monolingual training using the dataset corresponding to each specific language. We will measure the impact on MedExpQA of the different types of knowledge that LLMs may use:

(i) Gold grounding knowledge:

  (1) **E**: Full gold reference explanations as written by the medical doctors.
  (2) **EI**: Gold explanations about the Incorrect Options.
  (3) **H**: Full gold explanations with [HIDDEN] explicit references to the multiple-choice options.

(ii) Automatically obtained grounding knowledge:

  (1) **None**: Answering the medical question with no additional external knowledge.
  (2) **RAG-7**: Automatically obtained knowledge by applying MEDRAG to retrieve the k=7 most relevant documents.
  (3) **RAG-32**: Automatically obtained knowledge by applying MEDRAG to retrieve the k=32 most relevant documents.

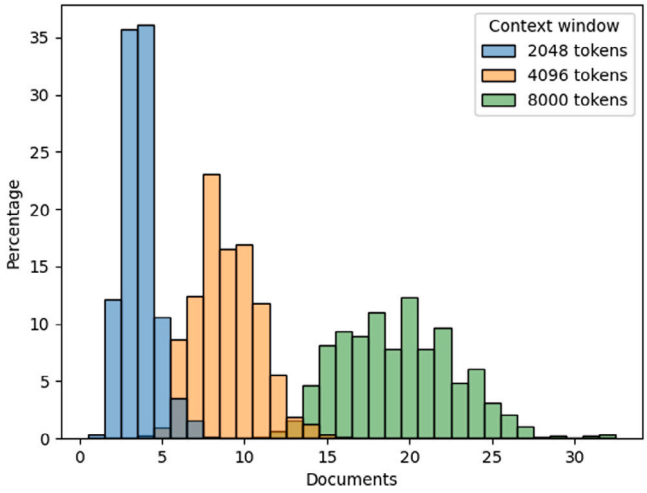


Fig. 4. Distribution of retrieved documents across different context windows. Three different histograms are shown that depict the maximum number of documents that can be accommodated within various context windows across dataset examples: 2,048 tokens (PMC-LLaMA), 4,096 tokens (LLaMA2), and 8,192 tokens (Mistral and BioMistral).

We use the entire clinical case, question, and multiple-choice options to generate the query for all 6 different evaluation settings. Gold knowledge grounding is leveraged as explained in the previous section. With respect to the methods to automatically obtained external knowledge, we take into account the results obtained in the MIRAGE benchmark [11] and apply MEDRAG by using the RRF-2 of two retrieval algorithms, namely, BM25 and MedCPT, over the MEDCORP corpus. We use the entire clinical case, question, and multiple-choice options to generate the query to retrieve the k = 7 most relevant documents. We define k = 7 by computing the average token length of MedCorp documents; if we consider that 85% of our prompts can be represented under 400 tokens, this leaves 1648 tokens for knowledge grounding, which amounts to 7 documents on average. This configuration is used to define **RAG-7**.

Furthermore, as MEDRAG obtained best results for most of the benchmarks when retrieving at most 32 documents, we also experimented with this setting. Nevertheless, it should be considered that the context window of each model, namely, the maximum amount of word tokens that each LLM can pay attention to in the input, will determine how many of these documents are actually fed into the LLM at each forward pass. Hence, when the combination of both the retrieved documents and the prompt exceed the context window, then we truncate the amount of documents to ensure that the prompt is not affected. Fig. 4 illustrates the distribution of documents corresponding to different context window sizes. Specifically, it shows the number of examples in the dataset that align with varying numbers of retrieved documents for context windows of 2048, 4096, and 8000 tokens. In the results reported in the next section, **RAG-32** for both zero-shot and fine-tune settings helps us to evaluate the impact of retrieving more or less relevant documents as external knowledge.

5.1. Evaluation

We ask LLMs to generate not only the index number of the predicted correct option but also the full textual answer. However, accuracy is calculated by comparing the first generated character after the prompt following *"The correct answer is: "*[8]. We verify that this character always corresponds to one of the options in the exams' possible answers. A provides an example of the prompts used for each language and for every model.

[8] And equivalent prompts for French, Italian and Spanish.

6. Results

We report the main results of the experiments performed in the MedExpQA benchmark in Table 8 for zero-shot while the fine-tuning accuracy scores are presented in Table 9.

Zero-shot results. They show that Mistral consistently achieves the highest accuracy across every evaluation setting and language, even outscoring the medical specific BioMistral. Among the gold knowledge results, we can see that removing the explanation of the correct answer (**EI**) really hinders performance. However, using the full gold reference answer helps LLMs to obtain excellent marks. Moreover, differences between using **E** and **H** are quite large, especially for languages different to English.

It should be noted that the best automatic method still fares very badly with respect to any of the gold knowledge results, which shows that retrieval methods for the medical domain still have large room for improvement. While the best automatic method corresponds to **RAG-7**, differences in performance are not that great with respect to **None** or **RAG-32**.

We hypothesize that the lack of substantial improvement when using 32 snippets for knowledge grounding may indicate that a saturation point may be reached beyond which additional snippets do not provide any additional benefit. To analyze this more precisely, we conducted an evaluation of the zero-shot performance of the 4 LLMs when feeding the model from 0 to up to 32 snippets, following a power of two sequence of snippets. Thus, Fig. 5 illustrates that a positive trend exists when increasing the number of snippets. However, we can see how this improvement tanks at around 8 snippets in most of the models. This result correlates to our findings in Tables 8 and 9.

Finally, performance on English was substantially higher for every models and RAG configurations. This manifests the English-centric focus of most LLMs while showcasing the urgent need of dedicating resources and effort to developing multilingual LLMs which could then compete across all languages included in multilingual benchmarks such as MedExpQA.

Fine-tuning results. They show that fine-tuning the LLMs on the CasiMedicos dataset help to greatly increase performance for every evaluation setting, language and LLM. BioMistral seems to obtain the best overall scores but that is due to its high scores on the full gold reference explanation setting (**E**). Thus, if we look at the rest of the evaluation settings, Mistral, as it happened in the zero-shot scenario, remains the best performing LLM on the MedExpQA benchmark.

The superior results of **None** with respect to RAG scores demonstrate that fine-tuning makes any external knowledge automatically retrieved using RAG methods redundant. Finally, while scores for French, Italian and Spanish remain lower than those obtained for English, performance for those languages greatly benefit from fine-tuning, especially if we compare them with their zero-shot counterpart results.

Overall results. Overall, results demonstrate that the gold reference explanations leveraged as knowledge for Medical QA help LLMs to obtain almost perfect scores, especially when fine-tuning the models. Fine-tuning particularly benefits **EI**, which obtains as good results as **H** applied in zero-shot settings.

Our results allow us to draw several more conclusions. First, that despite using state-of-the-art RAG methods for the medical domain [11], their results are rather disappointing. Both in zero-shot when compared with the results based on any kind of gold knowledge, and in fine-tuning in which RAG methods score worse than not using any additional knowledge.

Second, our MedExpQA benchmark suggests that overall performance of even powerful LLMs such as Mistral still have a huge room for improvement to reach scores comparable to those obtained when gold knowledge is available.

We calculated a McNemar [40] test of statistical significance to establish whether the *RAG-7* and *RAG-32* results were significantly

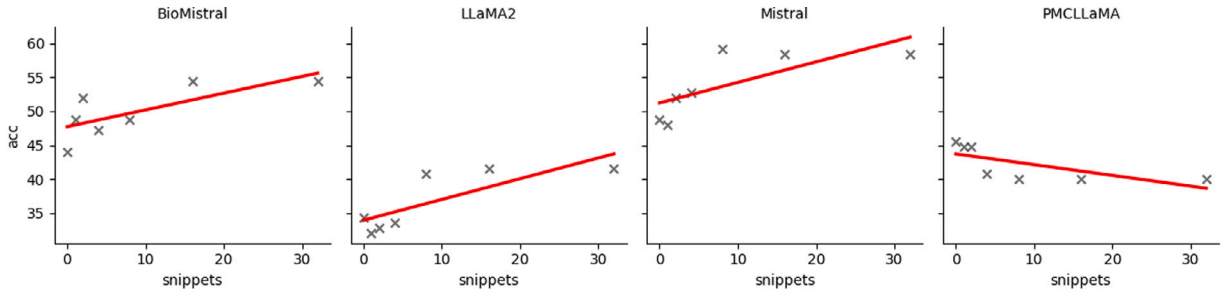


Fig. 5. Performance of different models in a zero-shot setting with up to 0, 2, 4, 8, 16, and 32 retrieved snippets.

Table 8
Zero-shot results. E: Full gold explanation. EI: Gold Explanations of the Incorrect Options; H: Full gold explanation with Hidden explicit references to the correct/incorrect answer; None: model without any additional external knowledge; RAG-7: Retrieval Augmented Generation with k = 7; RAG-32: Retrieval Augmented Generation with k = 32; underline: best result per type of knowledge; **bold**: best result overall.

	PMC-LLaMA (13B)				LLaMA2 (13B)				Mistral (7B)				BioMistral (7B)				Avg.
	EN	ES	IT	FR	EN	ES	IT	FR	EN	ES	IT	FR	EN	ES	IT	FR	ALL
E	83.2	77.6	76.8	80.0	81.6	77.6	77.6	75.2	**89.6**	**88.0**	**87.2**	**88.0**	88.8	83.2	80.8	80.8	**82.2**
EI	60.0	42.4	43.2	46.4	44.0	31.2	39.2	44.8	59.2	53.6	52.0	52.8	50.4	44.0	46.4	49.6	47.4
H	78.4	63.2	72.0	70.4	68.8	64.8	63.2	65.6	82.4	75.2	77.6	78.4	80.8	74.4	69.6	74.4	72.4
None	45.6	36.8	33.6	30.4	34.4	18.4	12.8	27.2	48.8	41.6	40.8	39.2	44.0	39.2	35.2	41.6	35.6
RAG-7	40.0	30.4	28.0	24.8	42.4	36.0*	30.4*	32.0	55.2	44.0	38.4	42.4	44.8	40.0	40.8	36.8	37.9
RAG-32	40.0	30.4	28.0	24.8	41.6	31.2*	32.8*	26.4	58.4*	41.6	41.6	42.4	54.4	37.6	31.2	39.2	37.6
Avg.	57.9	46.8	46.9	46.1	52.1	43.2	42.7	45.2	**65.6**	**57.3**	**56.3**	**57.2**	60.5	53.1	50.7	53.7	–

* Results that are statistically significant at $\alpha = .05$ wrt to their None baseline.

Table 9
Fine-tuning results. E: Full gold explanation. EI: Gold Explanations of the Incorrect Options; H: Full gold explanation with Hidden explicit references to the multiple choice options; None: model without any additional external knowledge; RAG-7: Retrieval Augmented Generation with k = 7; RAG-32: Retrieval Augmented Generation with k = 32; underline: best result per type of knowledge; **bold**: best result overall.

	PMC-LLaMA (13B)				LLaMA2 (13B)				Mistral (7B)				BioMistral (7B)				Avg.
	EN	ES	IT	FR	EN	ES	IT	FR	EN	ES	IT	FR	EN	ES	IT	FR	ALL
E	92.0	89.6	89.6	88.8	90.4	90.4	89.6	92.0	**94.4**	92.8	91.2	92.8	**94.4**	**93.6**	**92.0**	**93.6**	**91.7**
EI	69.6	67.2	67.2	68.0	73.6	70.4	66.4	70.4	81.6	78.4	75.2	76.8	73.6	72.0	71.2	71.2	72.1
H	82.4	76.0	80.0	82.4	83.2	85.6	84.0	81.6	88.0	84.8	88.8	88.0	83.2	82.4	86.4	84.8	83.9
None	58.4	48.8	49.6	53.6	57.6	50.4	53.6	54.4	68.0	63.2	56.8	66.4	61.6	58.4	56.8	65.6	57.7
RAG-7	56.8	35.2	44.8	38.4	60.8	56.8	48.8	51.2	69.6	59.2	56.8	64.8	64.8	57.6	61.6	59.2	55.4
RAG-32	56.8	35.2	44.8	38.4	60.8	52.0	51.2	49.6	75.2	55.2	52.0	60.0	65.6	57.6	55.2	60.8	54.4
Avg.	69.3	58.7	62.7	61.6	71.1	67.6	65.6	66.5	**79.5**	**72.3**	70.1	**74.8**	73.9	70.3	**70.5**	72.5	–

better than their respective *None* baselines. As it can be seen in Tables 8 and 9, only five zero-shot scores (out of 64) marked with an asterisk in Table 8 are statistically significant at $\alpha = .05$. Finally, performance for languages different to English is much lower for every model and evaluation setting. This points out to an urgent necessity to invest in the development and research of LLMs which may be optimized not only for English, but for other world languages too. Obviously, the evaluation of such LLMs would in turn require multilingual evaluation benchmarks which may be deployed to provide a comprehensive and realistic overview of their performance. We hope that contributing MedExpQA may serve as encouragement to the AI and medical research communities to generate more benchmarks of its kind for many of the world languages.

7. Discussion

The results discussed in the previous section show that even when performing fine-tuning with the full gold reference explanations LLMs still remain several points below perfect scores. Furthermore, the statistical analysis of the obtained results indicates that, despite differences compared to the **None** models, the performance gains (when that is the case) of models using *RAG-7* or *RAG-32* are, in 61 out 64 cases, not statistically significant. In contrast, the statistical analysis found out that the results using gold knowledge (**E**, **EI**, **H**) were all statistically significant at $\alpha = .05$

Apart from the evaluation results, and in order to better understand the dataset on which the MedExpQA is setup, we performed several analysis regarding the quality and quantity of the explanations provided by the CasiMedicos medical doctors.

Regarding the quality of the explanations, we found several examples such as the one depicted in Table 10. Instead of directly answering the question, the medical doctor (psychiatry resident) writing the explanation gives information that is not relevant to explain the correct answer (marked in red). We hypothesize that such explanations, which lack any relevant medical information, may have a negative impact on the final LLMs performance.

It should be noted that, despite CasiMedicos being a high-quality dataset written voluntarily by medical doctors, sometimes (i) their explanations may not follow a repetitive formal structure and, (ii) they are not always subjected to a second review by an auditor as it usually happens in specialized textual books.

Regarding the quantity of the explanations, around 5% of the full gold reference explanations in the CasiMedicos dataset do not contain any explicit explanation regarding the correct answer. Sometimes the

Table 10
Example of a gold full explanation (E) with irrelevant and not medical comments.

E	Another simple question with an immediate answer, which offers no doubt. It describes a patient worried about a non-existent physical defect, whose concern distresses him and prevents him from leaving the house. As a psychiatry resident, I wish the MIR questions in my specialty were a bit more thought-provoking and in-depth, although I know that the seconds you will have saved by marking **the fourth** one directly are very valuable.

medical doctor explains the incorrect options, hoping that the reader may indirectly reach the correct conclusion, or sometimes they are cases such as the one discussed above.

In any case, while it is possible to filter out such examples, we thought it useful to leave them with the aim of analyzing in the future the performance of LLMs and RAG methods for these specific cases. After all, we would like LLMs to be able to also generalize in situations in which the knowledge is provided in a non-standard structured manner, as it is the case in the large majority of the full gold reference explanations provided in CasiMedicos.

We would like to give a final word on multilinguality. Results have shown that performance for French, Italian and Spanish is worse across the board and we believe that this topic has a lot of interesting questions for future research. Are these results a consequence of the pre-training of the LLMs? For the RAG experiments, how much, positive or negative, influence has the fact that the extracted knowledge from MedCorp is in English? Would it be better to prompt the model only in English and then translate the answers into each of the target languages, in what is usually known as a *translate-test* approach? We believe that a benchmark such as MedExpQA would help to investigate these research questions which may be crucial to develop robust multilingual medical QA approaches.

8. Concluding remarks

In this paper we present MedExpQA, the first multilingual benchmark for Medical QA. As a new feature, our new benchmark also includes gold reference explanations to justify why the correct answer is correct and also to explain why the rest of the options are incorrect. The high-quality gold explanations have been written by medical doctors and they allow to test the LLMs when different types of gold knowledge is available. Comprehensive experimentation has demonstrated that automatic state-of-the-art RAG methods still have a long way to go to get near the scores obtained by LLMs when fed with gold knowledge. Furthermore, our benchmark has made explicit the lower overall performance of LLMs for languages other than English for Medical QA.

We think that MedExpQA may contribute to the development of AI tools to assist medical experts in their everyday activities by providing a robust multilingual benchmark to evaluate LLMs in Medical QA. Future work may involve evaluating LLMs not only regarding their accuracy in predicting the correct answer, but also on the quality of the explanations generated to justify such prediction. Of course, these approaches may pose new evaluation challenges that have not been yet contemplated in this work.

CRediT authorship contribution statement

Iñigo Alonso: Writing – original draft, Visualization, Validation, Software, Investigation, Data curation, Conceptualization. **Maite Oronoz:** Writing – original draft, Visualization, Supervision, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Rodrigo Agerri:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Methodology, Investigation, Funding acquisition, Data curation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Rodrigo Agerri reports financial support was provided by Spain Ministry of Science and Innovation. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank the CasiMedicos Proyecto MIR 2.0 for their permission to share their data for research purposes. This work has been partially supported by the HiTZ Center and the Basque Government, Spain (Research group funding IT1570-22). We are also thankful to several MCIN/AEI/10.13039/501100011033 projects: (i) Antidote (PCI2020-120717-2), and by European Union NextGenerationEU/PRTR; (ii) DeepKnowledge (PID2021-127777OB-C21) and ERDF A way of making Europe; (iii) Lotu (TED2021-130398B-C22) and European Union NextGenerationEU/PRTR; (iv) EDHIA (PID2022-136522OB-C22); (v) DeepMinor (CNS2023-144375) and European Union NextGenerationEU/PRTR. We also thank the European High Performance Computing Joint Undertaking (EuroHPC Joint Undertaking, EXT-2023E01-013) for the GPU hours.

Appendix A. Prompts

In this appendix, we provide the specific prompts used to interact with the Large Language Models of this work.

Appendix B. Hyperparameters

In this appendix we list some of the hyperparameters used in this work (see Table B.11).

Appendix C. Efficiency metrics

In this work we only use or apply the LLMs to establish our benchmark, be that in zero-shot or fine-tuning. As such, we do not perform any modification in the way the LLMs work. Therefore, for efficiency and architectural issues the original papers of Llama2, PMC-Llama, Mistral and BioMistral could be inspected. Our contributions are focused on (i) establishing a multilingual benchmark for Medical QA, (ii) experimenting with state-of-the-art RAG methods and (iii) providing gold reference explanations as a form of "gold" RAG that can be used to compare the LLMs with. Having said that, below we offer detailed information about some efficiency metrics. All the metrics have been calculated using a NVIDIA A100 Graphics Processing Unit (GPU).

- The total number of parameters updated through Low Rank Adaptation (LoRA) during Parameter-Efficient Fine-Tuning (PEFT) are the reported in Table C.12.
- Table C.13 shows the number of **samples per second** processed when using Mistral (7B) and LLaMA2 (13B) in a NVIDIA A100 GPU. The performance in the other two models, BioMistral (7B) and PMC-LLaMA (13B) is the same.
- Table C.14 shows **the time in minutes and hours** when processing data with Mistral (7B) and LLaMA2 (13B). The other two models, BioMistral (7B) and PMC-LLaMA (13B), showcase the same times.

```
======================== Prompt English ========================

You are a helpful medical expert, and your task is to answer a
multi-choice medical question using the relevant documents. Please
choose the answer from the provided options. Your responses will
be used for research purposes only, so please have a definite
answer.
Here are the relevant documents:
{context}
Here is the question:
{question}
Here are the potential choices:
{options}
The correct answer is:
```

Fig. A.6. Prompt used for models in English.

```
======================== Prompt Spanish ========================

Eres un experto médico y tu tarea consiste en responder a una
pregunta médica de test utilizando tu conocimiento y los
siguientes documentos relevantes. Por favor, elige la respuesta
entre las opciones proporcionadas. Tus respuestas se utilizarán
únicamente con fines de investigación, así que te rogamos que
proporciones una respuesta definitiva.
Estos son los documentos relevantes:
{context}
Aquí está la pregunta:
{question}
Aquí están las posibles opciones:
{options}
La opción correcta es:
```

Fig. A.7. Prompt used for models in Spanish.

```
======================== Prompt Italian ========================

Sei un medico esperto e il tuo compito consiste nel rispondere a
una domanda di test medico utilizzando le tue conoscenze e i
documenti successivi rilevanti. Per favore, scegli la risposta tra
le opzioni fornite. Le tue risposte verranno utilizzate
esclusivamente con fini di indagine, quindi ti chiediamo di
fornirti una risposta definitiva.
Questi sono i documenti rilevanti:
{context}
Ecco la domanda:
{question}
Ecco le opzioni possibili:
{options}
L'opzione corretta è:
```

Fig. A.8. Prompt used for models in Italian.

```
======================== Prompt French ========================

Vous êtes un expert en médecine et votre tâche consiste à répondre
à une question d'examen médical en utilisant vos connaissances et
les documents suivants. Veuillez choisir la réponse parmi les
options proposées. Vos réponses seront utilisées uniquement à des
fins de recherche, veuillez donc fournir une réponse claire.
Voici les documents pertinents:
{context}
Voici la question:
{question}
Voici les options possibles:
{options}
La bonne option est:
```

Fig. A.9. Prompt used for models in French.

```
======================== Prompt Redacting ========================
In the following text, remove all references that clearly state
that any of the options 1, 2, 3, {"4 or 5" if
example_contains_5_options else "or 4"} are either correct or
false. Don't change the original text and don't write linebreaks;
only replace with the tag [HIDDEN] the text that says that
something is the correct or incorrect option if there is are any.
Don't replace text that doesn't specifically imply that certain
something is the right or wrong answer. For example: the text
"option {correct_option_index} is correct." should be "[HIDDEN]",
the text "Option {random.choice(incorrect_option_indexes)} is less
likely because this and that" should be "[HIDDEN] this and that",
the text "answer blablabla is the right answer because whatever"
should be "answer blablabla is [HIDDEN] whatever". Here is the
text: {full_answer}
```

Fig. A.10. Prompts to remove explicit references to the multiple-choice options.

Table B.11
Hyperparameters used in the configuration of the experiments.

Hyperparameter	Value
Optimizer	adamw_torch_fused
Learning rate	0.00015
Weight decay	0.0
ADAM ϵ	1e−7
Epochs	10
Train batch size	16
Evaluation batch size	8
Floating Point 16-bit precision training	False
Brain Float 16-bit precision training	True
Maximum #tokens in input	
PMCLLaMA	2048
LLaMA2	4096
Mistral	8000
BioMistral	8000
Maximum #tokens in generation	
PMCLLaMA	2048
LLaMA2	4146
Mistral	8050
BioMistral	8050
Low-Rank Adaptation (LoRA)	
R parameter	8
LoRA α	16
LoRA Dropout	0.05

Table C.12
Trainable parameters: Number of parameter in training using the LoRA model; All parameters: total of parameters used in the LoRA model; Trainable %: number of trainable parameters of the total number of parameters in the LoRA model.

	7B parameter models		
	Trainable parameters	All parameters	Trainable %
Mistral and BioMistral	20,971,520	3,773,042,688	0.555825
	13B parameter models		
	Trainable parameters	All parameters	Trainable %
LLaMA2	31,293,440	6,703,272,960	0.466838
PMC-LLaMa	31,293,440	6,703,283,200	0.466838

Table C.13
Samples processed by second in a NIVIDIA A100 GPU. E: Full gold explanation. H: Full gold explanation with Hidden explicit references to the correct/incorrect answer; None: model without any additional external knowledge; RAG-7: Retrieval Augmented Generation with k = 7; RAG-32: Retrieval Augmented Generation with k = 32.

Samples per second	Train		Inference	
	7B	13B	7B	13B
E	1.981	1.270	7.681	4.757
H	1.998	1.282	7.676	4.76
None	3.248	2.116	11.375	6.956
RAG-7	1.031	0.629	3.637	2.081
RAG-32	0.191	0.281	0.744	1.013

Table C.14
Time in minutes (m) and hours (h) when processing data in a NIVIDIA A100 GPU. E: Full gold explanation. H: Full gold explanation with Hidden explicit references to the correct/incorrect answer; None: model without any additional external knowledge; RAG-7: Retrieval Augmented Generation with k = 7; RAG-32: Retrieval Augmented Generation with k = 32.

Time for training	7B	13B
E	1 h 4 m	2 h 1 m
H	1 h 9 m	2 h 9 m
None	47 m	1 h 39 m
RAG-7	1 h 42 m	3 h 2 m
RAG-32	7 h 34 m	5 h 31 m

References

[1] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023;620(7972):172–80.

[2] Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of gpt-4 on medical challenge problems. 2023, arXiv preprint arXiv:2303.13375.

[3] Safranek CW, Sidamon-Eristoff AE, Gilson A, Chartash D. The role of large language models in medical education: Applications and implications. JMIR Med Educ 2023;9:e50945.

[4] Goenaga I, Atutxa A, Gojenola K, Oronoz M, Agerri R. explanatory argument extraction of correct answers in resident medical exams. 2023.

[5] Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: Open foundation and fine-tuned chat models. 2023, arXiv:2307.09288.

[6] Wu C, Lin W, Zhang X, Zhang Y, Wang Y, Xie W. PMC-LLaMA: Towards Building Open-source Language Models for Medicine. 2023, arXiv:2304.14454.

[7] Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas Ddl, et al. Mistral 7B. 2023, arXiv preprint arXiv:2310.06825.

[8] Labrak Y, Bazoge A, Morin E, Gourraud P-A, Rouvier M, Dufour R. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. 2024, arXiv:2402.10373.

[9] Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards expert-level medical question answering with large language models. 2023, arXiv preprint arXiv:2305.09617.
[10] Xie Q, Schenck EJ, Yang HS, Chen Y, Peng Y, Wang F. Faithful AI in medicine: A systematic review with large language models and beyond. medRxiv 2023.
[11] Xiong G, Jin Q, Lu Z, Zhang A. Benchmarking Retrieval-Augmented Generation for Medicine. 2024, arXiv preprint arXiv:2402.13178.
[12] Agerri R, Alonso I, Atutxa A, Berrondo A, Estarrona A, García-Ferrero I, et al. HiTZ@Antidote: Argumentation-driven Explainable Artificial Intelligence for Digital Medicine. In: SEPLN 2023: 39th International Conference of the Spanish Society for Natural Language Processing. 2023.
[13] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst 2020;33:1877–901.
[14] Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. PaLM: Scaling Language Modeling with Pathways. J Mach Learn Res 2022;24:240:1–240:113.
[15] Phan LN, Anibal JT, Tran H, Chanana S, Bahadroglu E, Peltekian A, et al. SciFive: a text-to-text transformer model for biomedical literature. 2021, CoRR arXiv:2106.03598.
[16] García-Ferrero I, Agerri R, Salazar AA, Cabrio E, de la Iglesia I, Lavelli A, et al. Medical mT5: An Open-Source Multilingual Text-to-Text LLM for the Medical Domain. In: Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING). 2024.
[17] Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform 2022;23(6).
[18] Wang G, Yang G, Du Z, Fan L, Li X. ClinicalGPT: Large Language Models Finetuned with Diverse Medical Data and Comprehensive Evaluation. 2023, arXiv preprint arXiv:2306.09968.
[19] Jin Q, Dhingra B, Liu Z, Cohen W, Lu X. PubMedQA: A Dataset for Biomedical Research Question Answering. In: Inui K, Jiang J, Ng V, Wan X, editors. Proceedings of EMNLP-IJCNLP. Association for Computational Linguistics; 2019, p. 2567–77.
[20] Abacha AB, Shivade C, Demner-Fushman D. Overview of the MEDIQA 2019 Shared Task on Textual Inference, Question Entailment and Question Answering. In: Proceedings of the 18th bioNLP workshop and shared task. 2019, p. 370–9.
[21] Vilares D, Gómez-Rodríguez C. HEAD-QA: A Healthcare Dataset for Complex Reasoning. In: Proceedings of the ACL. Florence, Italy: Association for Computational Linguistics; 2019, p. 960–6.
[22] Abacha AB, Mrabet Y, Sharp M, Goodwin TR, Shooshan SE, Demner-Fushman D. Bridging the Gap Between Consumers' Medication Questions and Trusted Answers. In: MedInfo. 2019, p. 25–9.
[23] Jin D, Pan E, Oufattole N, Weng W-H, Fang H, Szolovits P. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Appl Sci 2021;11(14):6421.
[24] Pal A, Umapathi LK, Sankarasubbu M. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In: Flores G, Chen GH, Pollard T, Ho JC, Naumann T, editors. Proceedings of the conference on health, inference, and learning. Proceedings of machine learning research, Vol. 174, PMLR; 2022, p. 248–60.
[25] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res 2020;21(1):5485–551.
[26] Zakka C, Shad R, Chaurasia A, Dalal AR, Kim JL, Moor M, et al. Almanac — Retrieval-augmented language models for clinical medicine. NEJM AI 2024;1(2). AIoa2300068.
[27] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Adv Neural Inf Process Syst 2020;33:9459–74.
[28] Robertson S, Zaragoza H. The probabilistic relevance framework: BM25 and beyond. Found Trends Inf Retr 2009;3(4):333–89.
[29] Jin Q, Kim W, Chen Q, Comeau DC, Yeganova L, Wilbur WJ, et al. MedCPT: Contrastive Pre-trained Transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. Bioinformatics 2023;39(11):btad651.
[30] Lo K, Wang LL, Neumann M, Kinney R, Weld D. S2ORC: The semantic scholar open research corpus. In: Jurafsky D, Chai J, Schluter N, Tetreault J, editors. Proceedings of the ACL. Association for Computational Linguistics; 2020, p. 4969–83.
[31] Han T, Adams LC, Papaioannou J-M, Grundmann P, Oberhauser T, Löser A, et al. MedAlpaca–An Open-Source Collection of Medical Conversational AI Models and Training Data. 2023, arXiv preprint arXiv:2304.08247.
[32] Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. Cureus 2023;15(6).
[33] Yadav P, Tam D, Choshen L, Raffel C, Bansal M. TIES-Merging: Resolving Interference When Merging Models. In: Thirty-seventh conference on neural information processing systems. 2023.
[34] Yu L, Yu B, Yu H, Huang F, Li Y. Language models are super mario: Absorbing abilities from homologous models as a free lunch. 2023, arXiv preprint arXiv:2311.03099.
[35] Shoemake K. Animating rotation with quaternion curves. In: Proceedings of the 12th annual conference on computer graphics and interactive techniques. SIGGRAPH '85, New York, NY, USA: Association for Computing Machinery; 1985, p. 245–54.
[36] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: State-of-the-art natural language processing. In: Liu Q, Schlangen D, editors. Proceedings of EMNLP: System Demonstrations. Association for Computational Linguistics; 2020, p. 38–45.
[37] Cormack GV, Clarke CLA, Buettcher S. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval. SIGIR '09, New York, NY, USA: Association for Computing Machinery; 2009, p. 758–9.
[38] OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 Technical Report. 2024, arXiv:2303.08774.
[39] Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: Low-Rank Adaptation of Large Language Models. In: International Conference on Learning Representations. 2022.
[40] Dietterich TG. Approximate statistical test for comparing supervised classification learning algorithms. Neural Comput 1998;10(7):1895–923.