CasiMedicos-Arg: A Medical Question Answering Dataset Annotated with Explanatory Argumentative Structures

Anonymous ACL submission

Abstract

Explaining Artificial Intelligence (AI) decisions is a major challenge nowadays in AI, in particular when applied to sensitive scenarios like medicine and law. However, the need to explain the rationale behind decisions is a main issues also for human-based deliberation as it is important to justify *why* a certain decision has been taken. Resident medical doctors for instance are required not only to provide a (possibly correct) diagnosis, but also to explain how they reached a certain conclusion. Developing new tools to aid residents to train their explanation skills is therefore a central objective of AI in education. In this paper, we follow this direction, and we present, to the best of our knowledge, the first multilingual dataset for Medical Question Answering where correct and incorrect diagnoses for a clinical case are enriched with a natural language explanation written by doctors. These explanations have been manually annotated with argument components (i.e., premise, claim) and argument relations (i.e., attack, support). The Multilingual CasiMedicos-arg dataset consists of 558 clinical cases (English, Spanish, French, Italian) with explanations, where we annotated 5021 claims, 2313 premises, 2431 support relations, and 1106 attack relations. We conclude by showing how competitive baselines perform over this challenging dataset for the argument mining task.

1 Introduction

There is an increasingly large body of research on Artificial Intelligence (AI) applied to the medical domain with the objective of developing technology to assist and support medical doctors in explaining their decisions or how they have reached a certain conclusion. For example, resident medical doctors preparing for licensing exams may get an AI support to explain what and why is the treatment or diagnosis correct given some background information (Safranek et al., 2023; Goenaga et al., 2023).

A prominent example of this is the recent proliferation of Medical Question Answering (QA) datasets and benchmarks, in which the task often involves processing and acquiring relevant specialized medical knowledge to be able to answer a medical question based on the context provided by a clinical case (Singhal et al., 2023a; Nori et al., 2023; Xiong et al., 2024).

The development of Large Language Models (LLMs), both general purpose and specialized in the medical domain, has enabled rapid progress in Medical QA tasks which has led in turn to claims about LLMs being able to pass official medical exams such as the United States Medical Licensing Examination (USMLE) (Singhal et al., 2023b; Nori et al., 2023). Thus, publicly available LLMs such as LLaMA (Touvron et al., 2023) or Mistral (Jiang et al., 2023) and their respective medical-specific versions PMC-LLaMA (Wu et al., 2023) and BioMistral (Labrak et al., 2024), or proprietary models such as MedPaLM (Singhal et al., 2023b) and GPT-4 (Nori et al., 2023), to name but a few, have been reporting high-accuracy scores in a variety of Medical QA benchmarks[1](Singhal et al., 2023a,b; Xiong et al., 2024).

While these results constitute impressive progress, currently the Medical QA research field still presents a number of shortcomings. First, experimentation has been mostly focused on providing the correct answer in medical exams, usually in a multiple-choice setting. However, as doctors are also required to explain and argue about their predictions, research on Medical QA should also address the generation of argumentative explanations. Unfortunately, and to the best of our knowledge, no Medical QA dataset, that currently exists,

[1] https://huggingface.co/blog/leaderboard-medicalllm

includes correct and incorrect diagnoses enriched with natural language explanations written by medical doctors. Second, the large majority of Medical QA benchmarks are available only in English (Singhal et al., 2023a; Xiong et al., 2024), which makes it impossible to know the ability of current LLMs for Medical QA in other languages.

In this paper we address these issues by presenting CasiMedicos-Arg, the first Multilingual (English, French, Italian, Spanish) dataset for Medical QA with manually annotated gold explanatory argumentation about incorrect and correct predictions written by medical doctors. More specifically, the corpus consists of 558 documents with reference gold doctors' explanations which are enriched with manual annotations for argument components (5021 claims and 2313 premises) and relations (2431 support and 1106 attack). This new resource will make it possible, for the first time, to research not only on Argument Mining but also on generative techniques to argue about and explain predictions in Medical QA settings. Finally, strong baselines on argument component detection, a challenging sequence labelling task, using encoder (Devlin et al., 2019; He et al., 2021), encoder-decoder (García-Ferrero et al., 2024) and decoder-only LLMs (Jiang et al., 2023; Touvron et al., 2023) demonstrate the validity of our annotated resource. Data, code and fine-tuned models will be made publicly available upon publication.

2 Related Work

In this section we will focus on reviewing datasets for Medical QA and on Explanatory Argumentation, the two main features of our main contribution, CasiMedicos-Arg.

2.1 Medical Question Answering

Several of the most popular Medical QA datasets (Jin et al., 2019; Abacha et al., 2019b,a; Jin et al., 2021; Pal et al., 2022) have been grouped into three multi-task English benchmarks, namely, MultiMedQA (Singhal et al., 2023a), MIRAGE (Xiong et al., 2024), and the Open Medical-LLM Leaderboard (Pal et al., 2024), with the aim of providing comprehensive experimental evaluation benchmarks of LLMs for Medical QA.

MultiMedQA includes MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), LiveQA (Abacha et al., 2019b), MedicationQA (Abacha et al., 2019a), MMLU clinical topics (Hendrycks et al., 2020) and HealthSearchQA (Singhal et al., 2023a). Except the last one, all of them consist of a multiple-choice format and MedQA, MedMCQA and MMLU's source data comes from licensing medical exams. In terms of size, MedQA includes almost 15K questions, MedMCQA 187K while the rest of them are of more moderate sizes, namely, 500 QA pairs in PubMedQA, around 1200 in MMLU, 738 in LiveQA and 674 in MedicationQA.

While every dataset except MedQA and HealthSearchQA includes long form correct answers, they are not considered really usable for benchmarking LLMs because they were not optimally constructed as a *ground-truth* by medical doctors or professional clinicians (Singhal et al., 2023a).

Regarding the Open Medical-LLM Leaderboard, it also includes MedQA, MedMCQA, PubMedQA and MMLU clinical topics. General purpose LLMs such as GPT-4 (Nori et al., 2023), PaLM (Chowdhery et al., 2022), LLaMa (Touvron et al., 2023) or Mistral (Jiang et al., 2023) report high-accuracy scores on these Medical QA benchmarks, although recently a number of specialized LLMs for the medical domain are appearing, sometimes with even stronger performances. Some popular models include Med-PaLM (Singhal et al., 2023a), MedPaLM-2 (Singhal et al., 2023b), PMC-LLaMA (Wu et al., 2023), and more recently, BioMistral (Labrak et al., 2024).

The MIRAGE benchmark includes subsets of MedQA, MedMCQA, PubMedQA, MMLU clinical topics and adds the BioASQ-YN dataset (Tsatsaronis et al., 2015) with the aim of evaluating Retrieval Augmented Generation (RAG) techniques for LLMs in Medical QA tasks. According to the authors, their MEDRAG method 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 (Xiong et al., 2024). By employing MEDRAG, they are able to clearly improve the zero-shot results of some of the tested LLMs, although the results for others are rather mixed.

Summarizing, no Medical QA dataset currently provides reference gold argumentative explanations regarding the incorrect and correct predictions. Furthermore, and with the exception of Vilares and Gómez-Rodríguez (2019), they have been mostly developed for English, leaving a huge gap regarding the evaluation LLMs in Medical QA for other languages. Motivated by this we present CasiMedicos-Arg, the first Medical QA dataset including gold reference explanations which has been manually annotated with argumentative structures, including argument components (premises and claims) and their relations (support and attack).

2.2 Explanatory Argumentation in the Medical Domain

Explanatory argumentation in natural language refers to the process of generating or analyzing explanations within argumentative texts. In recent years, natural language explanation generation has gained significant attention due to the advancements of generative models that are leveraged to develop specialized explanatory systems. The need for explanation generation is also driven by the predominant use of non-transparent algorithms which lack interpretability, thus being unsuitable for sensitive domains as medical.

Camburu et al. (2018) tackle the task of explanation generation by introducing an extension of the Stanford Natural Language Inference (SNLI) dataset (Bowman et al., 2015), which includes a new layer of annotations providing explanations for the entailment, neutrality, or contradiction labels. The generation of these explanations is addressed with a bi-LSTM encoder trained on the new e-SNLI dataset. e-SNLI (Camburu et al., 2018) is also exploited to generate explanations for a NLI method, which first generates possible explanations for predicted labels (Label-specific Explanations) and then takes a final label decision (Kumar and Talukdar, 2020). The authors employ GPT-2 (Radford et al., 2019) for label-specific generation and classify explanations using RoBERTa (Liu et al., 2019).

Narang et al. (2020) focus on generating complete explanations in natural language following a prediction step, utilizing a T5 model. The model is trained to predict both the label and the explanation. Li et al. (2021) also propose to generate explanations along with predicting NLI labels. The generation step is leveraged for the question-answering task exploiting domain-specific or commonsense knowledge, while the NLI step allows to predict relations between a premise and a hypothesis.

In the medical domain, Molinet et al. (2024) propose generating template-based explanations for medical QA tasks. Their system incorporates medical knowledge from the Human Phenotype Ontology, making the explanations more verifiable and sound for the medical domain.

Despite the extensive research proposing various approaches to generate explanations, these approaches are not grounded on any argumentation model. This is particularly important in sensitive domains like medicine, where sound and well-founded explanations are essential to justify the taken decision. Moreover, medical explanations require verified medical knowledge at their core, which the described methods lack, as discussed in (Molinet et al., 2024).

3 CasiMedicos-Arg Annotation

The Spanish Ministry of Health yearly publishes the Resident Medical or *Médico Interno Residente* (MIR) licensing exams including the correct answer. Every year the CasiMedicos MIR Project 2.0[2] takes the published exams by the ministry and provide gold explanatory arguments written by volunteer Spanish medical doctors to reason about the correct and incorrect options in the exam.

The Antidote CasiMedicos corpus consists of the original Spanish commented exams by the CasiMedicos doctors which were cleaned, structured and freely released for research purposes (Agerri et al., 2023). The original Spanish data was automatically translated and manually revised into English, French, and Italian. The corpus includes 622 documents each with a short clinical case, the multiple-choice questions and the explanations written by medical doctors[3].

In the rest of this section we describe the process of manually annotating argumentative structures in the raw Antidote CasiMedicos dataset.

3.1 Argumentation Annotation Guidelines

In line with the guidelines proposed by Mayer et al. (2021) for Randomized Controlled Trials (RCT) annotation, we identify two main argument components: Claims and Premises, and their relations, Support and Attack. Furthermore, we also propose to annotate Markers and labels specific to the medical domain, namely, Disease, Treatment and Diagnostics. In the following, we define and describe the annotation of each label.

Claim is a concluding statement made by the author about the outcome of the study (Mayer et al., 2021):

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

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

1. *The patient's presenting picture is presumably erythema nodosum. (CasiMedicos)*

2. *We propose immunotherapy with thymoglobulin and cyclosporine as a proper treatment. (CasiMedicos)*

Premise corresponds to an observation or measurement in the study, which supports or attacks another argument component, usually a claim. It is important that they are observed facts, therefore, credible without further evidence (Mayer et al., 2021):

3. *In addition, pancytopenia is not observed. (CasiMedicos)*

4. *What is important is that the eye that has received the blow does not go up, and therefore there is double vision in the superior gaze. (CasiMedicos)*

Analyzing the CasiMedicos dataset, we found certain ambiguity between claims and premises. Thus, statements representing general medical knowledge about a disease, symptoms, or treatments must be annotated as claims. Although these statements may support or attack the main claim, they are not premises since they do not involve case-specific evidence but represent medical facts:

5. *[The patient's presenting picture is presumably erythema nodosum]. [About 10% of cases of erythema nodosum are associated with inflammatory bowel disease, both ulcerative colitis and Crohn's disease]. [As mentioned, in most cases, erythema nodosum has a self-limited course]. [When associated with inflammatory bowel disease, erythema nodosum usually resolves with treatment of the intestinal flare, and recurs with disease recurrences. Local measures include elevation of the legs and bed rest]. (CasiMedicos)*

Here the first statement in square brackets represents a claim that asserts the patient's diagnosis (*erythema nodosum*). The following ones represent information about the diagnosis, its symptoms and its possible treatment. They are not based on the evidences given in the case, but on general medical knowledge available to the doctor. Therefore, these examples should be annotated as Claims.

Additionally, long statements with multiple self-contained pieces of evidence must be divided into single premises to differentiate their relations to specific claims. For example, a given evidence in a sentence may support a claim while others may attack it. To preserve these distinctions, such sentences should be split into independent premises.

As well as Claims and Premises we annotate **Markers** – discourse markers that are relevant for arguments as they help to identify the spans of argument components and the type of argumentative relations. In the following examples markers are written in bold:

6. ***Other causes** related to this picture are autoimmune diseases **leading to** transverse myelitis (Behcet's, FAS, SLE,...) **or** inflammatory diseases such as sarcoidosis, **although** our patient does not seem to meet the criteria for them. (CasiMedicos)*

7. ***Although** this usually gives a subacute **or** chronic picture. (CasiMedicos)*

The possible answers proposed in the CasiMedicos multiple-choice options corresponds to predicting a **Disease**, a **Treatment** or a **Diagnosis**. We decided to also annotate them as they help to identify the type of doctor's arguments (whether to look justification of a diagnosis or about a possible treatment) and the type of argumentative relations.

For advanced reasoning comprehension, we need to explore argumentative relations connecting argument components (claims and premises) and forming a structure of an argument (Mayer et al., 2021). Here we provide the definitions of support and attack relations, as well as real examples illustrating them.

Support. All statements or observations justifying the proposition of a target argument component are considered as supportive (Mayer et al., 2021):

8. *In the examination there is a clear dissociation with thermoalgesic anesthesia and preservation of arthrokinetic and vibratory. [1] Reflexes are normal, neither abolished nor exalted. [2] In addition, the rest of the examination is strictly normal. [3] **With all this I believe that the correct answer is 5, that is a syringomyelic lesion, whose initial characteristic is the sensitive dissociation with anesthesia for the thermoalgesic and conservation of the posterior chordal**. (CasiMedicos)*

This example provides premises (in italic) that justify a claim (bold) which they are related to. The supportive nature is highlighted by the marker *With all this I believe...* .

Attack. An argument component is attacking another one if (i) it contradicts the proposition of a target component or (ii) it undercuts its implicit assumption of significance or relevance, for example, stating that the observations related to a target component are not significant or not relevant (Mayer et al., 2021):

9. ***It might be tempting to answer 3 Fracture of the superior wall of the orbit with entrapment of the superior rectus muscle****. However, muscles trapped in a fracture do not automatically lose their muscular action. (CasiMedicos)*

10. *The palpebral hematoma and hyposphagma (subconjunctival hemorrhage) does not give us the key data. (CasiMedicos)*

These examples represent premises (in italic) which either contradict their claims (bold) in Example 9 or which are not considered significant to justify or reject target components (Example 10).

3.2 Annotation Process and Results

The annotation process consisted of three stages: training, reconciliation, and complete dataset annotation. During training, annotators worked on 10 CasiMedicos cases. We then calculated inter-annotator agreement (IAA) results of the training phase to highlight any weak spots, guideline flaws, and any issues in the dataset needing further analysis.

At the reconciliation phase, the descriptions of Claim and Premise labels were discussed and agreed upon. After this, we started the complete dataset annotation. As mentioned earlier, the original CasiMedicos dataset included 622 medical cases, but 64 cases were excluded during the annotation phase. Some of them did not have gold explanations while others were cases with confusing relations: the correct answer is a wrong disease, treatment, or diagnosis as asked in a question, thus, it is attacked by its premises instead of being supported. Therefore, the final number of annotated cases is 558. In the following subsections we present the IAA of the entire dataset (3.3), annotation results and their description (3.4).

3.3 Inter-Annotator Agreement (IAA)

The IAA is calculated over a random batch of 100 CasiMedicos cases. Since one instance (e.g. a claim) is usually an entire self-contained sentence, we measured the IAA at both instance level and at token level. In other words, we compute agreement over entire instances and over the tokens of each instance.

Label	Mean F1
Claim	0.765
Premise	0.659
Marker	0.642
Disease	0.639
Treatment	0.586
Diagnostics	0.527

Table 1: Instance-based F1 agreement.

Label	Mean F1
Claim	0.915
Premise	0.891
Marker	0.634
Disease	0.738
Treatment	0.777
Diagnostics	0.638

Table 2: Token-based F1 agreement.

Table 1 illustrates the IAA at instance level. Since instances are very long, annotators may be uncertain about which elements to include, leading to lower agreement scores for some labels. However, the major labels Claim and Premise have relatively good results with scores of 0.765 and 0.659, respectively. The mean F1 over all labels is 0.669.

Table 2 shows the IAA at the token level. Here we compute the agreement over tokens of each instance. The highest agreement score is of a Claim label being 0.915, while the lowest is of a Diagnostics label accounting for 0.638. The mean F1 over all tokens is 0.880.

3.4 Annotation Results

In this part we report the stats about label distribution over entire cases (documents) and the label distribution over the doctor's explanations only. Additionally, we also discuss the distribution of argumentative relations.

Table 3 reports the total number of entities over the dataset and the average number of entities per case. Table 4 shows the label distributions only for the explanations, namely, the total number of entities in explanations and the averaged number of entities per explanation. In both tables we notice that the discrepancy between the average number

Label	Total	Mean per case
Claim	5021	8.998
Premise	2313	4.145
Marker	1117	2.0
Disease	1791	3.21
Treatment	1278	2.29
Diagnostics	786	1.40

Table 3: Label Distribution over Entire Cases.

Label	Total	Mean per explanation
Claim	3003	5.948
Premise	470	0.935
Marker	974	1.833

Table 4: Label Distribution in Explanations.

of claims per explanation and of premises per explanation is rather high. This may seem strange since premises are needed to accept or reject claims in order to complete one argumentation unit.

However, there are plausible reasons for such distribution. First, there is a certain number of cases where the explanation is based on evidences from doctor's knowledge rather than clinical facts described in the case itself. Such explanations take into account the information given about the patient (e. g. age, symptoms, vital signs), but do not repeat any of these facts (as in *Example 1* in Appendix A). Second, explanations that do not repeat evidences from the case are frequent, e. g. *"Here we must suspect ... disease. All the symptoms fall perfectly within the picture"*; *"This is a fairly easy epidemiology question, in adults without other data, Pneumococcus is the 1st"*). Last but not least, there is a group of cases with implicit premises or implicit warrants: the explanation presents claims (e. g. a conclusion about a disease and a treatment) implying that some evidences from the case text and implying certain medical knowledge to align evidences with a disease and a choice of treatment (as in *Example 2* in Appendix A).

In Table 5 we present the distribution of argumentative relations. Support relations appear twice as much as Attack ones, making this argumentation pattern frequent and probably more convincing. In cases where the conclusion is made solely excluding wrong propositions by attacking them there is a lack of confidence about the claim.

As a result, we present CasiMedicos-Arg, a multi-layer argument-based annotation of the English version of CasiMedicos consisting of **558** clinical cases with explanations. In the following sections we describe the experiments performed on argument component detection (claims and premises) to establish strong baselines on the task and validate our annotations.

Relation	Total	Mean per case
Support	2431	4.357
Attack	1106	1.982

Table 5: Distribution of Argumentative Relations.

4 Experimental Setup

We first describe the process of projecting the manually annotated argumentation labels from the source English data to the other three target languages, namely, French, Italian and Spanish. This process will result in the Multilingual Casimedicos-Arg which will then be leveraged to produce strong baselines on argument component detection using a variety of LMs, including encoders (Devlin et al., 2019; He et al., 2021), encoder-decoders (García-Ferrero et al., 2024) and decoder-only LLMs (Touvron et al., 2023; Jiang et al., 2023).

4.1 Multilingual CasiMedicos-Arg

Taking the manually annotated English CasiMedicos-Arg as starting point, we first needed to project the annotations to Spanish, French, and Italian following the method described in Yeginbergenova and Agerri (2023). Second, and to ensure that the projection method correctly leveraged the annotations to the new data we additionally performed an automatic post-processing step of the newly generated data to correct any misalignments. Finally, and to guarantee the quality of annotations and the validity of our evaluations, the translated and projected data is manually revised by native speakers.

Label projection is performed using word alignments calculated by AWESOME (Dou and Neubig, 2021) and Easy Label Projection (García-Ferrero et al., 2022) to automatically map the word alignments into sequences (argument components) and project them from the source (English) to the target language (French, Italian and Spanish).

A particular feature of argument of argument components is that the sequences could span over the entire length of the sentences. Therefore, after revising the automatically projected data, an extra post-processing step was performed by correcting the projections in the sequences where some annotations were placed incorrectly. The most common correction was fixing articles at the beginning of the argument components, which were systematically missed out during the automatic projection step. Other sequences were labeled only by half instead of the whole sequence. This post-processing step was essential to minimize the human labor during manual correction. The number of corrections introduced during the post-processing step can be found in Appendix B.

The final manual correct step involved checking the translation quality and projected labels by native expert annotators fixing any misprojections or errors in the translation. The result of this process is the Multilingual CasiMedicos-Arg dataset, obtained by projecting the manual annotations from English to Italian, French and Spanish.

4.2 Sequence Labelling with LLMs

We leverage Multilingual CasiMedicos-Arg to perform crosslingual and multilingual argument component detection, a task that, due the heterogeneity and length of the sequences, is usually a rather challenging task (Stab and Gurevych, 2017; Eger et al., 2018; Yeginbergenova and Agerri, 2023). Furthermore, In addition to classic encoder-only models like mBERT (Devlin et al., 2019) and mDeBERTa (He et al., 2021), we decided to also perform the task using encoder-decoder and decoder-only models. For the encoder-decoder category, we chose two variants of Medical mT5, a multilingual text-to-text model adapted to multilingual medical texts: med-mT5-large and med-mT5-large-multitask (García-Ferrero et al., 2024). For the decoder-only architecture, we selected the LLaMA2 (Touvron et al., 2023) and Mistral (Jiang et al., 2023) models with 7B parameters. The domain-specific versions of these models produced less promising results, so we opted to report the results of the aforementioned models.

Previous work in sequence labeling with LLMs has demonstrated that discriminative approaches based on encoder-only models still outperform generative techniques based on LLMs (Wang et al., 2023). The motivation behind it is usually the nature of the sequence labeling task that even though LLMs possess some linguistic knowledge they suffer from a number of problems, notably, hallucinated content. In this paper we use the LLMs for Sequence Labelling library to fine-tune the generative models with unconstrained decoding[4].

We structure the experiments as follows. First, we perform *monolingual* experiments in which we train and test for each language separately. Note that for English we use the gold standard annotations, while for French, Italian and Spanish we are fine-tuning the models on *projected* data, what in crosslingual transfer research is usually called *data-transfer*. Additionally, we also report results of *model-transfer* (fine-tuning the models in English and predict in the rest of the target languages). Finally, we experiment with *multilingual* data augmentation by pooling the training data of all four languages and then evaluate in each language separately.

Since each model has its own way of learning due to the architecture, namely, some models learn better over longer iterations and others perform at a good level in less time, we report the best results yielded from the models under different hyperparameters. Multilingual BERT and mDeBERTa were fine-tuned for 3 epochs, while Medical mT5 required 20 epochs; the rest of the hyperparameters are based on previous related work (Yeginbergenova and Agerri, 2023) and (García-Ferrero et al., 2024), respectively. Regarding LLaMA2 and Mistral, they were fine-tuned for 5 epochs leaving the rest of the hyperparameters as default.

Model	Monolingual	Multilingual
mBERT	76.24(0.59)	<u>77.14</u>(0.97)
mDeBERTa	77.08(0.89)	<u>77.30</u>(0.59)
med-mT5-large	80.43(0.22)	<u>82.37</u>(0.21)
med-mT5-large-multitask	80.93(0.26)	<u>82.03</u>(0.32)
LLaMA2-7B	81.49(0.82)	<u>83.07</u>(0.11)
Mistral-0.1-7B	**83.27**(0.48)	83.24(0.73)

Table 6: F1-scores and their standard deviations for argument component detection in English CasiMedicos-Arg; **bold**: best overall result; <u>underlined</u>: best result per model across the two language settings.

5 Empirical Results

In this section, we report the results obtained after performing the steps described in Section 4. All the results and standard deviations reported in this section are obtained by averaging three randomly initialized runs. We evaluate using sequence level F1-macro score, a common metric for argument component detection.

[4] https://github.com/ikergarcia1996/Sequence-Labeling-LLMs

Model	Spanish	French	Italian	Avg.
	monolingual data-transfer			
mBERT	75.39(0.49)	73.66(0.66)	74.78(0.59)	74.61
mDeBERTa	77.39(0.83)	76.35(0.29)	76.98(0.76)	<u>76.91</u>
med-mT5-large	80.79(0.19)	80.12(0.59)	80.32(0.04)	80.41
med-mT5-large-multitask	80.69(0.65)	80.13(0.56)	80.70(0.08)	80.51
LLaMA2-7B	80.39(0.52)	80.89(0.54)	80.69(0.46)	80.66
Mistral0.1-7B	81.71(0.29)	81.38(0.52)	81.56(0.44)	81.55
	multilingual data-transfer			
mBERT	75.08(0.89)	74.92(0.62)	74.95(1.38)	<u>74.98</u>
mDeBERTa	76.06(1.42)	76.22(0.89)	77.06(0.65)	76.45
med-mT5-large	82.07(0.12)	80.85(0.26)	80.89(0.72)	<u>81.27</u>
med-mT5-large-multitask	82.09(0.26)	80.83(0.28)	80.57(0.49)	<u>81.16</u>
LLaMA2-7B	81.56(0.28)	81.03(0.49)	81.16(0.20)	<u>81.25</u>
Mistral-0.1-7B	82.40(0.12)	82.10(0.33)	81.41(0.69)	<u>**81.97**</u>
	cross-lingual model-transfer			
mBERT	72.75(0.24)	71.47(1.27)	72.49(0.09)	72.24
mDeBERTa	76.05(0.14)	74.63(0.53)	75.22(0.32)	75.30
med-mT5-large	79.91(1.26)	78.51(1.20)	79.41(0.87)	79.28
med-mT5-large-multitask	79.81(0.83)	77.96(0.13)	77.07(0.34)	78.28
LLaMA2-7B	75.31(0.68)	68.56(1.07)	73.86(0.51)	72.58
Mistral-0.1-7B	79.27(0.42)	70.62(7.37)	78.36(0.37)	76.08

Table 7: F1-scores and their standard deviations of data-transfer (monolingual and multilingual), and cross-lingual model-transfer experiments using Spanish, French, and Italian data; **bold**: best overall result; <u>underlined</u>: best result per model across the three language settings.

We first show the results on monolingual (using the manually annotated English data) and multilingual (fine-tuning on all four languages and evaluating in English) in Table 6. Overall, it can be observed that the decoder-only generative models outperform the rest, though the Medical mT5 models are nearly as effective. Furthermore, the *multilingual* method of pooling all languages into a single dataset proves to be beneficial for every model, improving over the results obtained when training using the gold standard English data only.

The results for Spanish, French and Italian are displayed in Table 7. As for the English results, it can be seen that the *multilingual data-transfer* approach is the most effective setting, even with LLMs which are supposedly pre-trained on English data only. Among all the models, Mistral achieves the highest F1-macro scores. However, while for all the other models the multilingual training was advantageous no substantial improvement was observed in a similar setting with Mistral. Finally, it can be seen that *crosslingual model transfer* is the least optimal of the settings, even when using state-of-the-art multilingual LMs such as mDeBERTa (He et al., 2021). An interesting point to note is that for *crosslingual model transfer* the best results are obtained by the Medical mT5 models, which may be due to this model being trained on multilingual medical data (García-Ferrero et al., 2024).

Summarizing, in this section we present competitive baselines for argument component detection on CasiMedicos-Arg, validating both the manual annotations and the strategy of projecting English labels to other languages to facilitate the application of crosslingual and multilingual techniques.

6 Conclusion

In this paper we present CasiMedicos-Arg, a multilingual (French, English, Italian and Spanish) Medical QA dataset including gold reference explanations written by medical doctors which has been annotated with argumentative structures. This dataset aims to bridge a glaring gap in the Medical QA ecosystem by facilitating the evaluation of explanations generated to argue or justify a given prediction.

The final dataset includes 558 documents (parallel in four languages) with reference gold doctors' explanations which are enriched with manual annotations for argument components (5021 claims and 2313 premises) and relations (2431 support and 1106 attack).

Both interannotator agreement results and the baselines provided for argument component detection demonstrate the validity of our annotations. Furthermore, experiments show the advantage of performing argument component detection from a *multilingual data-transfer* perspective.

Limitations

We consider two main limitations in our work that we would like to address in the short term future. First, the choice of languages. We would have liked to include languages from different language families and with different morphological and grammatical characteristics, but we were limited by the native expertise available to us to perform the manual corrections of the projected labels and translations. Second, the size of the dataset (558 documents) could be larger.

Regarding the first limitation, we still think that our experiments demonstrate the superiority of performing *multilingual data-transfer* over *cross-lingual model transfer*, at least with the LLMs currently available. With respect to the size of the dataset, we would like to point out that its size is similar to other datasets reviewed in Section 2, which are being widely used to benchmark LLMs for Medical QA.

Another issue worth considering in the future is the need to further research the generation of explanations for the predictions while taking into account a crucial unsolved issue, namely, the evaluation explanation generation in the highly specialized medical domain.

Acknowledgments

References

Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis R Goodwin, Sonya E Shooshan, and Dina Demner-Fushman. 2019a. Bridging the Gap Between Consumers' Medication Questions and Trusted Answers. In *MedInfo*, pages 25–29.

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019b. 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*, pages 370–379.

Rodrigo Agerri, Iñigo Alonso, Aitziber Atutxa, Ander Berrondo, Ainara Estarrona, Iker García-Ferrero, Iakes Goenaga, Koldo Gojenola, Maite Oronoz, Igor Perez-Tejedor, German Rigau, and Anar Yeginbergenova. 2023. Hitz@antidote: Argumentation-driven explainable artificial intelligence for digital medicine. In *SEPLN 2023: 39th International Conference of the Spanish Society for Natural Language Processing.*

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *NeurIPS*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. *arXiv preprint arXiv:2101.08231*.

Steffen Eger, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2018. Cross-lingual argumentation mining: Machine translation (and a bit of projection) is all you need! In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 831–844, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Iker García-Ferrero, Rodrigo Agerri, and German Rigau. 2022. Model and data transfer for cross-lingual sequence labelling in zero-resource settings. In *In Findings of EMNLP*.

Iker García-Ferrero, Rodrigo Agerri, Aitziber Atutxa Salazar, Elena Cabrio, Iker de la Iglesia, Alberto Lavelli, Bernardo Magnini, Benjamin Molinet, Johana Ramirez-Romero, German Rigau, Jose Maria Villa-Gonzalez, Serena Villata, and Andrea Zaninello. 2024. Medical mt5: An open-source multilingual text-to-text llm for the medical domain. *Preprint*, arXiv:2404.07613.

Iakes Goenaga, Aitziber Atutxa, Koldo Gojenola, Maite Oronoz, and Rodrigo Agerri. 2023. Explanatory argument extraction of correct answers in resident medical exams. *arXiv preprint arXiv:2312.00567*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577. Association for Computational Linguistics.

Sawan Kumar and Partha Talukdar. 2020. NILE : Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online. Association for Computational Linguistics.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *Preprint*, arXiv:2402.10373.

Dongfang Li, Jingcong Tao, Qingcai Chen, and Baotian Hu. 2021. You can do better! if you elaborate the reason when making prediction. *arXiv preprint arXiv:2103.14919*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tobias Mayer, Santiago Marro, Elena Cabrio, and Serena Villata. 2021. Enhancing evidence-based medicine with natural language argumentative analysis of clinical trials. *Artificial Intelligence in Medicine*, 118:102098.

Benjamin Molinet, Santiago Marro, Elena Cabrio, and Serena Villata. 2024. Explanatory argumentation in natural language for correct and incorrect medical diagnoses. *Journal of Biomedical Semantics*, 15.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Ankit Pal, Pasquale Minervini, Andreas Geert Motzfeldt, Aryo Pradipta Gema, and Beatrice Alex. 2024. openlifescienceai/open_medical_llm_leaderboard. `https://huggingface.co/spaces/openlifescienceai/open_medical_llm_leaderboard`.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Conrad W Safranek, Anne Elizabeth Sidamon-Eristoff, Aidan Gilson, and David Chartash. 2023. The role of large language models in medical education: Applications and implications. *JMIR Med Educ*, 9:e50945.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023b. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16:1–28.

David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A Healthcare Dataset for Complex Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966, Florence, Italy. Association for Computational Linguistics.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Towards building open-source language models for medicine. *Preprint*, arXiv:2304.14454.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*.

Anar Yeginbergenova and Rodrigo Agerri. 2023. Cross-lingual argument mining in the medical domain. *arXiv preprint arXiv:2301.10527*.

A Appendix. CasiMedicos Real Cases

Example 1:

QUESTION TYPE: DERMATOLOGY
CLINICAL CASE:

A 62-year-old man with a history of significant alcohol abuse, carrier of hepatitis C virus, treated with Ibuprofen for tendinitis of the right shoulder, goes to his dermatologist because after spending two weeks on vacation at the beach he notices the appearance of tense blisters on the dorsum of his hands. On examination, in addition to localization and slight malar hypertrichosis. The most likely diagnosis is:

1- Epidermolysis bullosa acquisita.
2- Porphyria cutanea tarda.
3- Phototoxic reaction.
4- Contact dermatitis.
5- Acute intermittent porphyria.

CORRECT ANSWER: 2

Porphyria Cutanea Tarda: 60% of patients with PCT are male, many of them drink alcohol in excess, women who develop it are usually treated with drugs containing estrogens. Most are males with signs of iron overload, this overload reduces the activity of the enzyme uroporphyrinogen decarboxylase, which leads to the elevation of uroporphyrins. HCV and HIV infections have been implicated in the precipitation of acquired PCT. There is a hereditary form with AD pattern. Patients with PCT present with blistering of photoexposed skin, most frequently on the dorsum of the hands and scalp. In addition to fragility, they may develop hypertrichosis, hyperpigmentation, cicatricial alopecia and sclerodermal induration.

Example 2:

QUESTION TYPE: PEDIATRICS
CLINICAL CASE:

6-month-old infant presenting to the emergency department for respiratory distress. Examination: axillary temperature 37.2°C, respiratory rate 40 rpm, heart rate 160 bpm, blood pressure 90/45 mmHg, SatO2 95% on room air. He shows moderate respiratory distress with intercostal and subcostal retraction. Pulmonary auscultation: scattered expiratory rhonchi, elongated expiration and slight decrease in air entry in both lung fields. Cardiac auscultation: no murmurs. It is decided to keep the patient under observation in the hospital for a few hours. What do you consider the most appropriate attitude at this time with regard to the complementary tests?

1- Request venous blood gas, leukocyte count and acute phase reactants.
2- Request chest X-ray.
3- Request arterial blood gases and acute phase reactants.
4- Do not request complementary tests.

CORRECT ANSWER: 4

The patient probably presents with bronchiolitis. At this stage, no additional tests should be performed unless there is a clinical worsening.

Set (Language)	Number of corrections
Train (ES)	450
Test (ES)	153
Dev (ES)	64
Train (FR)	378
Test (FR)	109
Dev (FR)	49
Train (IT)	336
Test (IT)	117
Dev (IT)	55

Table 8: Number of corrections introduced in the post-processing step after automatic label projection.

B Number of corrections after annotation projection

The number of corrections required after automatically projecting the annotations.