When Multilingual Models Compete with Monolingual Domain-Specific Models in Clinical Question Answering

Vojtěch Lanz and Pavel Pecina
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{lanz,pecina}@ufal.mff.cuni.cz

Abstract

This paper explores the performance of multilingual models in the general domain on the clinical Question Answering (QA) task to observe their potential medical support for languages that do not benefit from the existence of clinically trained models. In order to improve the model's performance, we exploit multilingual data augmentation by translating an English clinical QA dataset into six other languages. We propose a translation pipeline including projection of the evidences (answers) into the target languages and thoroughly evaluate several multilingual models fine-tuned on the augmented data, both in mono- and multilingual settings. We find that the translation itself and the subsequent QA experiments present a differently challenging problem for each of the languages. Finally, we compare the performance of multilingual models with pretrained medical domain-specific English models on the original clinical English test set. Contrary to expectations, we find that monolingual domain-specific pretraining is not always superior to general-domain multilingual pretraining. The source code is available at `https://github.com/lanzv/Multilingual-emrQA`.

1 Introduction

Medical professionals spend considerable time going through (long) clinical documents to find answers to specific questions about particular patients (Demner-Fushman et al., 2009). This process can be simplified using natural language processing models designed for Question Answering (QA), either by searching for relevant evidence to answer the question or directly providing a precise answer that does not even need to be present in the context texts (Tsatsaronis et al., 2015). Patients would directly benefit from this more efficient process through better quality care. In addition, such QA systems can be designed specifically for patients, allowing them to ask direct questions about their discharge summaries or about other aspects of their medical records (Soni and Demner-Fushman, 2025).

...
Lungs : R lower 01-20 with coarse BS and rales ; L side clear ; no wheezing Abd : thin, nd, nt, soft, no masses palpable Ext : thin, no edema, multiple old well-healed scars on R leg Skin : warm and dry, no rash or breakdown noted though could not examine sacrum Neuro : reactive to pain, otherwise

Pertinent Results : 2014-01-20 05:30 AM BLOOD WBC - 10.9 RBC - 4.63 Hgb - 13.6 * Hct - 40.3 # MCV - 87 MCH - 29.3 MCHC - 33.7 RDW - 14.0 Plt Ct - 393 # 2014-01-20 05:30 AM BLOOD Neuts - 82.6 * Lymphs - 14.5 * Monos - 2.2 Eos - 0.2 Baso - 0.4 2014-01-20 02:08 PM BLOOD PT - 13.2 PTT - 27.4 INR (PT)- 1.2 2014-01-20 05:30 AM BLOOD Plt Ct - 393 # 2014-01-20 05:30 AM BLOOD Glucose - 334 *
...

Figure 1: Clinical text sample from emrQA dataset (Pampari et al., 2018), after filtration by Yue et al. (2020).

Finding specific evidence supporting an answer in discharge summaries is a crucial step for two reasons: First, given the sensitive nature of the data and the current inability to guarantee that models will not hallucinate, the model must point to the specific part of the text that it used to generate its response. This allows a physician to verify the answer directly. Second, discharge summaries are typically lengthy documents, which pose challenges for large language models (LLMs) (Premasiri et al., 2023; Luo et al., 2024). Extracting relevant evidence from the text and incorporating it into prompts within a Retrieval-Augmented Generation setup offers a potential solution to this problem (Abdelghafour et al., 2024).

Currently, most medical research data and related QA models are conducted predominantly in English (Jin et al., 2019; Henry et al., 2019; Johnson et al., 2023) although most medical institutions use their local language to produce clinical texts, and models trained on English data are not applicable to documents in other languages.

In contrast, general-domain multilingual models

Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), pages 69–82
May 4, 2025 ©2025 Association for Computational Linguistics

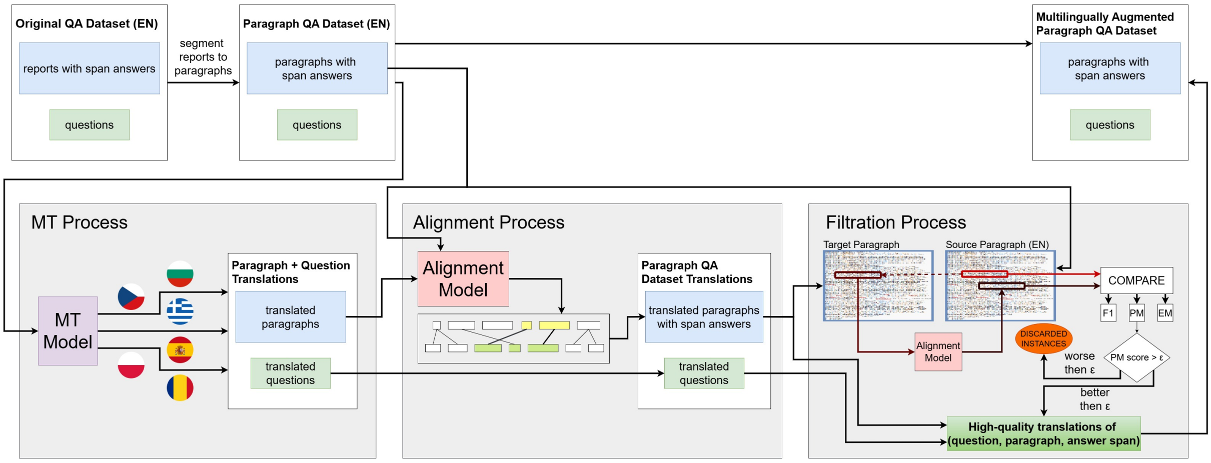

Original QA Dataset (EN)
reports with span answers
questions
segment reports to paragraphs
Paragraph QA Dataset (EN)
paragraphs with span answers
questions
Multilingually Augmented Paragraph QA Dataset
paragraphs with span answers
questions
MT Process
MT Model
Paragraph + Question Translations
translated paragraphs
translated questions
Alignment Process
Alignment Model
Paragraph QA Dataset Translations
translated paragraphs with span answers
translated questions
Filtration Process
Target Paragraph
Source Paragraph (EN)
COMPARE
F1
PM
EM
Alignment Model
DISCARDED INSTANCES
PM score > ε
worse then ε
better then ε
High-quality translations of (question, paragraph, answer span)

Figure 2: Multilingual data augmentation pipeline for the emrQA dataset.

(Devlin et al., 2018; Sanh et al., 2019; Conneau et al., 2019) are available for QA tasks in various languages. This raises two questions: How do such models, which have never been exposed to clinical data, perform clinical QA tasks? How important is the pretraining of the clinical domain?

To enhance the performance of multilingual models and expose them to more clinical data during fine-tuning, this study explores the impact of multilingual data augmentation. Several previous works have shown that multilingual data augmentation generally improves the performance of multilingual models (Liu et al., 2021; Bornea et al., 2021). However, it remains unclear whether the same holds in the clinical domain, which often differs from the standard language (Henriksson et al., 2014) (see Figure 1 for an illustration).

In this paper, we explore this idea by translating an English QA dataset derived from the emrQA dataset (Pampari et al., 2018) into six European languages: Bulgarian (BG), Czech (CS), Greek (EL), Spanish (ES), Polish (PL), and Romanian (RO) (as shown in Figure 2). We present a systematic approach to machine translation of a QA dataset that produces multilingual data for the task of finding evidence in clinical text that answers a given question. We exploit these translations for fine-tuning and evaluation of various models in monolingual and multilingual settings to investigate the impact of such multilingual data augmentation. Following Yue et al. (2020) and Lanz and Pecina (2024), we use two subsets from the emrQA dataset – *Medication* and *Relations*

We first describe the Machine Translation (MT) pipeline, which involves translating clinical reports, translating questions, and projecting the answer evidence substring into the translated text. Next, we discuss some poor-quality translated samples and propose how to deal with them. We then use these translations to fine-tune several Transformer-based models on the QA task. Based on that, we investigate how multilingual data augmentation improves the models' performance. Finally, we compare the performance of multilingual models with the clinically pretrained domain-specific models and discuss whether the clinical pretraining is necessary for this task.

This paper presents the following contributions:

- We propose a pipeline for augmentation of the clinical QA dataset into other languages.
- We introduce a novel unsupervised forward-backward substring alignment evaluation method that allows a more accurate assessment of substring alignment quality between languages without the need for labeled data.
- We demonstrate the performance of multilingual models on clinical QA tasks, highlighting the benefits of multilingual data augmentation and showing that clinical pretraining does not have to be more beneficial than general-domain multilingual pretraining.

2 Related Work

The task of QA involving the retrieval of the answer evidence substrings for a given question in a provided context has been extensively explored through various datasets. Among the most prominent are general purpose QA datasets such as SQuAD (Rajpurkar et al., 2016), which has also been already translated into several European languages via MT methods (Macková and Straka, 2020; Carrino et al., 2020; Cattan et al., 2021; Staš et al., 2023; Nuutinen et al., 2023). In addition to these, the clinical QA domain has gained attention with the emrQA dataset (Pampari et al., 2018), derived from the n2c2 challenge dataset (Henry et al., 2019).

Considerable work was done on the emrQA dataset with notable contributions by Yue et al. (2020), who adapted two emrQA subsets into a SQuAD-like format for more general use. Lanz and Pecina (2024) proposed segmentation of reports into paragraphs for better QA performance.

Various medical datasets exist in multiple languages, and the Khresmoi data set (Dušek et al., 2017) stands out as a parallel corpus of medical sentences in several European languages. Furthermore, there is a growing trend towards the development of datasets focused on extracting information from clinical documents in languages other than English (López-García et al., 2023; Zaghir et al., 2024; Richter-Pechanski et al., 2024). Furthermore, Gaschi et al. (2023) extended the n2c2 dataset by translating it into French and German (and we build on this work). This process involved aligning named entities using methods such as FastAlign (Dyer et al., 2013) and Awesome (Dou and Neubig, 2021). They also used machine translation systems such as Opus-MT (Tiedemann and Thottingal, 2020) and FAIR (Ng et al., 2019). However, the most recent MT systems are currently NLLB (Costa-jussà et al., 2022) and MadLad (Kudugunta et al., 2023).

In their multilingual experiments, Gaschi et al. (2023) tested a range of multilingual models, including mBERT (Devlin et al., 2018), distilmBERT (Sanh et al., 2019), and XLM-R (Conneau et al., 2019). However, these models are not pretrained on medical/clinical data, unlike BioBERT (Lee et al., 2019) or ClinicalBERT (Alsentzer et al., 2019), which were already used for emrQA experiments on English data (Yue et al., 2020; Lanz and Pecina, 2024). Despite the existence of LLMs trained on predominantly English medical data, such as MediTron (Chen et al., 2023) and BioMistral (Labrak et al., 2024), Lanz and Pecina (2024) demonstrated that the application of LLMs to answer substring-based evidence QA tasks is not straightforward, often computationally expensive without providing proportional benefits.

	Medication	Relations
Number of reports	262	426
Number of paragraphs	5 081	9 482
Number of questions	232 347	987 965

Table 1: Statistics of the *Medication* and *Relations* subsets segmented into paragraphs (each question has at least one answer in a paragraph).

3 Machine Translation of QA Dataset

This section outlines the MT methodology for the *Medication* and *Relations* subsets of the emrQA dataset, filtered and normalized by Yue et al. (2020). The process includes two phases: First, clinical reports and questions are translated using multilingual LLMs. Second, for each answer evidence, we find the corresponding substring in the translated text.

Clinical reports often pose a challenge for MT due to the size and complexity of their text. In addition, aligning answer evidences in such large texts would be challenging and error-prone. Therefore, we begin with segmenting the reports into paragraphs proposed by Lanz and Pecina (2024) which reduce the size of the context while preserving all necessary information (see statistics in Table 1).

3.1 Translation Process

Several recent works have presented highly robust MT models for general domains (Popel et al., 2020; Costa-jussà et al., 2022; Kudugunta et al., 2023). However, it is unclear how these models perform on clinical data. Following Gaschi et al. (2023), the performance of several MT models was evaluated in the Khresmoi medical domain data set (Dušek et al., 2017) (the results are reported in the Appendix B). For subsequent experiments, we chose MadLad-3B, which performs best or is very similar to the best results, but is significantly smaller and thus more time and memory efficient.

Translations of the questions in our dataset were done sentence by sentence. Translating (sometimes much) longer paragraphs turned out to be more challenging. Therefore, long paragraphs were divided into shorter parts. The paragraphs that exceed 750 characters were split into two parts of about the same length – preferably at the end of the sentence identified by the regular expression[1] closest to the middle of the entire paragraph. If such a split were not feasible, we split the segment at the whitespace

[1] `[a-z]{2}\.\s+[A-Z][a-z]`

closest to the middle of a paragraph. After translation, all segments within the paragraph are joined in their original order.

MadLad-3B sometimes tends to hallucinate when translating clinical reports, especially when they contain abundant medical abbreviations, acronyms, and figures. To address this, we propose the following solution: We append the phrase "Based on medical reports." after the end of each segment to be translated, providing the model with explicit context that the text is related to a clinical text (which is not always obvious from the segment content itself). If a correct translation of this phrase appears in a newly translated segment, it is removed along with any surrounding whitespaces. Otherwise, the text is translated again, with additional spaces inserted between the segment and the prompted medical phrase to make the difference even more explicit. In case of an increase in the limit of translation attempts, the standard translation using the MT model without any additional phrases was chosen. We refer to this method as the Prompted Medical Phrase (PMP) approach and compare it with the standard MT. The list of alternative translations of the phrase added to the prompt in all languages is provided in Appendix C. An example of the PMP approach is provided in Appendix D.

3.2 Answer Evidence Alignment

After translating the paragraphs, the answer evidence for each question must be found in the translated text. Due to the synthetic nature of evidence substrings in emrQA, these evidence segments often lack structure, sometimes appearing as incomplete sentences. Additionally, clinical texts frequently contain repetitive patterns (e.g., "mg," "q.p.m."), making the alignment crucial to correctly identify key clinical terms. However, these concepts are often very specific and the model may not have encountered them in alignment-based approaches during training. See Figure 3 for examples of evidence substrings from emrQA.

To align the answer evidence substring in the translated text, we could translate the original substring and locate it in the translated paragraph, as done for SQuAD (Macková and Straka, 2020; Cattan et al., 2021; Staš et al., 2023). However, due to the complexity of clinical data, identical translation cannot be guaranteed. Since SQuAD evidence is usually short (such as a person's name or location), the problem is not so complex. Therefore, this paper opts for word alignment methods, similarly to Gaschi et al. (2023) and Zaghir et al. (2024). Specifically, this work considers two alignment models: the statistical model FastAlign (Dyer et al., 2013) and the Transformer-based model Awesome (Dou and Neubig, 2021) to project evidence from the source to the target language.

> Question: Has the patient ever tried heparin?
> Context:
> ...
> HOSPITAL COURSE: The patient was admitted on 0/25/95 with concern for her being at high risk of skin breakdown and infection. Plans were made for panniculectomy, however, prior to this procedure, she was admitted for IV Ancef t.i.d.; Hibiclenz showers and sub-q Heparin. Preoperatively, her pulmonary function was assessed and found to have an FEV-1 of 53% of predicted; FVC of 57% of predicted and an FEV-1/FVC of 93% of predicted. Chest x-ray showed no active cardiopulmonary disease ...

> Question: Has the patient had allopurinol in the past?
> Context:
> ...
> MEDICATIONS: At the time of admission were as follows: 1. Lantus 100 mg q.p.m. 2. Humalog 20 units q.p.m. 3. Humalog sliding scale. 4. Neurontin 300 mg t.i.d. 5. Lisinopril 40 mg q.d. 6. Allopurinol 300 mg q.d . 7. Hydrochlorothiazide 25 mg q.d. 8. Zocor 20 mg q.d. 9. TriCor 50 mg b.i.d. 10. Atenolol 25 mg q.d. 11. Eyedrops prednisolone and atropine. 12. The patient was on iron supplementation.
> ...

Figure 3: Examples of emrQA evidence substrings, highlighted as colored spans showing alignment challenges.

Awesome is a pretrained aligner, while FastAlign requires additional training. For this purpose, we use the parallel corpus NLLB (Costa-jussà et al., 2022), selecting the first 44.6 million sentences paired with English for each of the languages involved in our work. Since we have the same amount of data for each language, we can directly compare alignments across languages. Alignment is performed on the same segments as described in Section 3.1. Based on the predicted alignment, the counterparts of the source answer evidence are found in the translated paragraph. The alignment of the first and last words determines the boundaries of the target answer evidence substring.

As observed by Gaschi et al. (2023), the choice of an aligner is not straightforward. They noted that performance in the general domain is not always indicative of behavior on clinical data sets, leading to an initial suboptimal choice in their study. To objectively compare the performance of Awesome and FastAlign, this work introduces the unsupervised forward-backward substring alignment evaluation method. This method involves a double answer evidence substring alignment, once from the source language to the target language and then back to the source. As a result, there are two substrings in the source language: the original answer evidence

	BG			**CS**			**EL**			**ES**			**PL**			**RO**		
	EM	**F1**	**PM**	**EM**	**F1**	**PM**	**EM**	**F1**	**PM**	**EM**	**F1**	**PM**	**EM**	**F1**	**PM**	**EM**	**F1**	**PM**
FastAlign	32.1	83.2	82.4	50.0	86.6	86.0	28.6	81.6	80.9	54.6	90.9	90.5	48.3	89.0	88.4	34.2	86.7	85.3
Awesome	46.0	82.9	82.4	64.0	89.8	89.4	24.8	70.3	69.8	71.2	93.7	93.5	57.1	89.3	89.1	64.7	90.9	90.4
FastAlign PMP	41.0	88.9	88.2	53.1	91.4	91.0	**41.9**	**87.9**	**87.2**	56.3	93.8	93.4	50.1	**90.8**	90.2	35.7	89.6	88.1
Awesome PMP	**59.3**	**89.2**	**88.8**	**66.8**	**93.0**	**92.8**	36.5	76.2	75.7	**72.9**	**96.3**	**96.1**	**58.8**	90.6	**90.5**	**68.0**	**93.8**	**93.5**

Table 2: Comparison of FastAlign and Awesome and impact of the PMP translation approach on *Medication* subset.

	BG			**CS**			**EL**			**ES**			**PL**			**RO**		
	EM	**F1**	**PM**	**EM**	**F1**	**PM**	**EM**	**F1**	**PM**	**EM**	**F1**	**PM**	**EM**	**F1**	**PM**	**EM**	**F1**	**PM**
FastAlign	54.9	89.9	89.1	61.2	91.5	90.9	55.8	91.1	90.6	66.7	93.6	93.4	62.7	92.2	91.5	53.3	90.0	89.2
Awesome	60.7	86.3	86.0	66.0	91.0	90.8	40.2	77.3	77.0	77.0	95.1	95.2	59.5	88.3	87.9	72.3	91.8	91.5
FastAlign PMP	61.1	**92.9**	**92.1**	67.0	**94.0**	**93.5**	**60.6**	**92.1**	**91.7**	71.0	95.3	95.1	**66.7**	**93.9**	**93.2**	57.0	91.9	91.2
Awesome PMP	**66.8**	89.4	89.0	**70.2**	93.2	93.0	44.9	79.7	79.5	**79.3**	**97.0**	**97.2**	62.6	90.1	89.8	**76.2**	**94.3**	**94.1**

Table 3: Comparison of FastAlign and Awesome and impact of the PMP translation approach on *Relations* subset.

substring and a two-step alignment projection of the answer evidence substring, both included in the same source paragraph. Ideally, the two substrings should be identical.

If the substring changes (expands, shrinks, shifts, etc.) during the two-step alignment projection, the alignment is considered inaccurate. An incorrect answer evidence substring alignment in the forward step is likely to carry over to the backward projection, leading to further errors. In contrast, successful alignment in both directions serves as a reliable indicator of accurate projection from the source language to the translation language. Of course, the projection of the substring alignment from the source language to the target language could be correct, but the second projection back to the source language was problematic. So, this evaluation method is stricter than directly measuring the quality of the newly generated answer evidence substrings. Furthermore, it also indirectly evaluates the quality of the MT from the previous stage described in Section 3.1. Poor translation would hinder accurate alignment, allowing this method to compare the performance of the straightforward MT and the PMP approach.

In the unsupervised forward-backward substring alignment evaluation, we compare two English substrings and aim for identity. To measure string similarity, we use SQuAD metrics — Exact Match (EM) and F1 score. However, evaluating the correctness of the projected substring position, not just the word similarity, may be valuable. Thus, in addition to Exact Match (EM) and F1, we introduce Position Match (PM) computed as:

$$\mathrm{PM} = \frac{2 \times O_P \times O_T}{O_P + O_T} \tag{1}$$

where $O_P = \frac{\text{Overlap Length}}{\text{Predicted Length}}$ is the predicted overlap ratio, and $O_T = \frac{\text{Overlap Length}}{\text{True Length}}$ is the true overlap ratio. The overlap is the common span between the predicted and original substring positions.

The final scores, averaged over all aligned answer evidence substrings, are shown in Tables 2 and 3. The PMP approach improves the performance of the standard MT model. The *Relations* subset is easier to process for the MT and alignment stages compared to the *Medication* subset, achieving F1 scores higher than 90% for most languages. The EM metric shows that approximately two-thirds of the answer evidence substrings in almost every language were perfectly projected without change. The *Medication* subset is more challenging but still exhibits good results. For both subsets, the Transformer-based aligner Awesome excels in Romance languages, while FastAlign outperforms in Greek. For Slavic languages, Awesome performs better in the *Medication* subset, but the results in the *Relations* subset are less clear. Only for Polish, FastAlign outperforms Awesome in all metrics. The differences between FastAlign and Awesome may be due to the fact that we trained FastAlign on all our languages, whereas Awesome was fine-tuned for word alignment only on the Romanian-English language pair relevant to our study. This could explain the performance disparities between Romance languages and others. However, since Awesome is based on mBERT, which has seen all these languages during pretraining, and Dou and Neubig (2021) showed that Awesome performs well even without fine-tuning, the impact of fine-tuning should not be pronounced.

3.3 Evaluation on Full Clinical Reports

Building on the results from the previous section, we base our next experiments on the PMP translation approach. For the *Medication* subset, we will

	BG		CS		PL	
	EM	F1	EM	F1	EM	F1
Awesome	**54.1**	77.4	**61.7**	81.4	53.0	76.8
FastAlign	50.4	**79.4**	57.5	**82.0**	**55.2**	**80.4**

Table 4: Comparison of mBERT performance on *Relations* translated to Slavic languages aligned by Awesome/FastAlign (paragraphs joined into full reports).

utilize FastAlign for Greek while adopting Awesome for all remaining languages. For the *Relations* subset, FastAlign will be employed for Greek, and Awesome for the Romance languages. To make a final decision on the most appropriate alignment method for Slavic languages in the *Relations* subset, this section evaluates the QA performance of the mBERT model using full clinical reports as context (rather than paragraphs, where we could not consider translated contexts that do not contain any question-answer pairs), considering both alignment models. Then, we compare alignment quality based on QA performance.

We follow the experiments of Yue et al. (2020). For this purpose, we focus on the Slavic languages within the *Relations* subset, Bulgarian, Czech, and Polish, and compare the QA results obtained using FastAlign and Awesome alignments, measured using the official SQuAD evaluation script. The results are presented in Table 4.

For Polish, we confirmed that FastAlign is the superior method. For Bulgarian and Czech, the choice is less clear, as the EM and F1 scores diverge. Although FastAlign shows a marginal F1 advantage, Awesome substantially outperforms in EM, so we proceeded with Awesome-based alignment for both languages in the following experiments on the *Relations* subset.

3.4 Filtering-out Low-Quality Alignments

Despite the alignment being mostly good, it is not always perfect. One reason might be flawed translations from the first stage. We also lack information about paragraphs that do not contain answers that need to be aligned to a new language. Therefore, paragraphs and answers with low alignment scores need to be filtered out, ignoring paragraphs without answers. This simplifies the task to Paragraph QA (similar to Oracle QA from Lanz and Pecina (2024)), resembling the SQuAD-like format (context is a paragraph rather than a document). Therefore, we examine which substring alignments we should discard and which ones we should keep (similarly as was done by Macková and Straka (2020)).

Low-quality answer evidence substring alignments negatively impact both the quality of the training and subsequent evaluation. Thanks to the forward-backward substring alignment evaluation, the quality of answer evidence projection can be estimated. This allows for filtering out those with low scores from the dataset, along with their corresponding paragraph context and question. Additionally, paragraphs can be removed if no question-answer pair is available, as there is no information about the quality of such paragraphs. As a result, in the remainder of this work, we focus on Paragraph QA instead of full report QA.

To determine how many answer evidences should be discarded, we conduct the following experiment. We sort the answer evidences from the training data based on their PM scores and sequentially remove $0, 5, 10, 15, 20, 30, 40, ...\%$ of the low-quality instances and for each resulting subset, we fine-tune the mBERT model (for each language separately) and compare the performance on the (silver) full test sets using Exact Match (EM) and F1 measures as in Yue et al. (2020). The results are averaged over three measurements with different random seeds and visualized in Figure 5 in Appendix E. Removing about 15% of lowest-quality instances improves the scores. Beyond this point, further removal risks losing complex data samples that may not have been perfectly aligned but remain essential for our task.

The pipeline described above is applied to the generated non-English training data and also to test data. Traditionally, such data is referred to as *silver data*, a term used to describe data that is automatically generated through processing of the original high-quality gold standard data. We experiment with two test sets: the full test set (which may contain alignment errors) and the intersection test set, formed by intersecting the translated and filtered test sets in each language, assuming higher reliability. The intersection test set contains identical instances across languages.

4 Multilingual Paragraph Question Answering Experiments

In this section, the performance of multilingual models is evaluated using the original English test set by assessing EM/F1 on the Paragraph QA task. The quality of the emrQA translations is also dis-

EM Score	Full Test Set							Intersection Test Set						
Models	**EN**	**BG**	**CS**	**EL**	**ES**	**PL**	**RO**	**EN**	**BG**	**CS**	**EL**	**ES**	**PL**	**RO**
distilmBERT (*mono*)	30.5	19.7	23.1	16.6	26.4	23.2	24.9	32.6	24.7	27.8	20.6	30.0	28.0	29.2
mBERT (*mono*)	32.7	21.4	25.0	17.8	28.7	24.3	27.8	34.6	26.5	29.7	22.0	32.4	29.0	32.5
XLM-R (*mono*)	33.4	22.1	26.0	18.3	29.1	25.5	28.0	35.4	27.3	30.9	22.3	32.8	30.5	32.6
XLM-R Large (*mono*)	**33.7**	23.0	26.5	19.1	**30.4**	26.0	28.5	35.4	28.2	31.5	23.3	**34.3**	30.6	33.1
distilmBERT (*multi*)	31.3	21.2	24.8	18.2	28.1	25.0	26.7	33.2	26.2	29.4	22.4	31.3	29.8	31.2
mBERT (*multi*)	33.0	22.6	26.6	19.4	29.9	26.6	28.5	35.1	27.6	31.3	23.9	33.5	31.7	33.2
XLM-R (*multi*)	33.5	22.8	26.8	19.5	30.0	27.1	28.6	35.4	27.7	31.5	24.2	33.3	31.9	33.1
XLM-R Large (*multi*)	33.6	**23.7**	**27.4**	**20.6**	30.3	**27.1**	**29.0**	**35.5**	**29.1**	**32.0**	**25.3**	33.6	**32.1**	**33.8**

Table 5: QA results on the *Medication* subset (EM scores) for monolingual (*mono*) and multilingual (*multi*) models.

F1 Score	Full Test Set							Intersection Test Set						
Models	**EN**	**BG**	**CS**	**EL**	**ES**	**PL**	**RO**	**EN**	**BG**	**CS**	**EL**	**ES**	**PL**	**RO**
distilmBERT (*mono*)	71.6	62.6	65.8	56.8	67.8	65.4	67.2	72.6	66.2	68.4	60.3	69.7	68.3	69.1
mBERT (*mono*)	75.3	66.0	69.7	60.1	71.0	67.9	70.7	76.0	69.8	72.1	63.6	72.5	71.0	72.8
XLM-R (*mono*)	75.9	67.4	71.1	61.8	72.3	69.9	72.2	76.6	71.0	73.8	65.5	74.0	72.8	74.5
XLM-R Large (*mono*)	**77.4**	69.3	72.7	63.7	74.1	70.9	73.6	**78.0**	72.8	75.2	67.5	75.7	73.6	75.8
distilmBERT (*multi*)	74.5	66.9	70.4	61.1	71.7	69.4	71.4	75.2	70.5	72.4	65.1	73.3	72.5	73.4
mBERT (*multi*)	76.7	68.6	72.3	63.5	74.0	71.5	73.3	77.3	72.2	74.2	67.3	75.4	74.4	75.2
XLM-R (*multi*)	77.0	69.6	72.8	64.5	74.1	72.0	73.5	77.6	73.0	75.0	68.4	75.5	74.6	75.7
XLM-R Large (*multi*)	77.3	**70.3**	**73.7**	**65.5**	**74.9**	**72.7**	**74.2**	77.8	**73.7**	**75.6**	**69.3**	**76.4**	**75.5**	**76.3**

Table 6: QA results on the *Medication* subset (F1 scores) for monolingual (*mono*) and multilingual (*multi*) models.

cussed by analyzing the performance of multilingual models on the translated data. In addition, the impact of including multilingual data during fine-tuning on model performance is investigated.

For these experiments, we selected four multilingual models mBERT, distilmBERT, XLM-R, and XLM-R Large (as Gaschi et al. (2023) did). In all experiments, we use filtered training data (discarding the 15% weakest alignments of the answer evidence substrings). Based on the analysis of Yue et al. (2020), we randomly sample the QA pairs to have the same number of training samples as 20% and 5% of the original unfiltered training data in the *Medication* and *Relations* subsets, respectively.

For the test set, we analyze two approaches. The first uses the entire unfiltered test sets. The second filters each translation by discarding the weakest 15% of alignments of the answer evidence substrings and then takes the intersection of filtered test sets across languages, allowing direct comparison. This filtering roughly retains 63% of the question-answer-paragraph triplets from the full unfiltered test sets. We split both *Medication* and *Relations* reports into train/dev/test according to a 7:1:2 ratio and perform experiments with three different random seeds for the splits. Finally, we examine multilingual training, where a single model is trained on the combined training data of all languages and evaluated separately on each. The results are shown in Tables 5, 6, 16 and 17.

4.1 QA Evaluation Across Languages

When the results of the full test set of other languages are compared with English, the results for Romance languages show a slight decrease, Slavic languages drop a bit more, and Greek displays a substantial difference. The results clearly reflect the quality already measured by the unsupervised forward-backward substring alignment evaluation method, which assesses the overall quality of the MT process, including substring alignment. This trend is seen not only across languages, but also in EM and F1 scores. Although F1 scores remain high under the alignment evaluation method, and therefore the Paragraph QA F1 score differences of new languages and English are not that large, EM scores in Paragraph QA show a much larger drop.

When trying to balance the quality of the test sets by filtering out poor-quality answer alignments and taking the intersection of languages, the scores across languages become more similar (except for Greek, which remains considerably lower).

Interestingly, we also observe that in the case of *Medication*, the English results improve on the intersection test set. This suggests that by removing poorly aligned answers during translations, we also excluded more complex answers regarding the QA prediction process. The remaining question

	Medication		Relations	
	EM	F1	EM	F1
BERTbase	31.0	72.9	91.1	96.2
BioBERT	31.1	74.4	91.7	96.9
ClinicalBERT	31.4	73.9	92.0	96.9
mBERT (*w/o tgt*)	31.0	75.9	90.0	96.0
mBERT (*mono*)	32.7	75.3	**92.8**	**97.3**
mBERT (*multi*)	**33.0**	**76.7**	92.6	**97.3**

Table 7: Performance comparison of clinical-domain monolingual and general-domain multilingual models.

is whether these are genuinely complex question-answer-paragraph triplets or if they represent annotation errors in the original emrQA dataset, which, due to its synthetic origin, contains numerous inaccuracies (Yue et al., 2020).

4.2 Impact of Multilingual Training

As we can see in Tables 5, 6, 16 and 17, multilingual training almost always slightly improves both EM and F1 scores, except in rare cases. As was already described, this training involves using all training sets from all languages to train a single model. In some cases, the improvement from multilingual training is even a few percentage points, particularly for smaller and faster models or for more problematic dataset translations.

When comparing multilingual training on the gold data in English, we arrive at a similar conclusion: augmenting the data with additional languages helps, particularly for the *Medication* subset, where Paragraph QA performance improves in all cases except with the XLM-R Large model. For the *Relations* subset, however, the differences are almost negligible, which may be due to the fact that the *Relations* task is approaching its oracle and has little room for further improvement (Yue et al., 2020).

5 Domain-Specific Models: Not Always Superior

In the previous section, we learned that multilingual models demonstrate strong performance, particularly on the *Relations* subset, despite never being specifically pretrained on clinical or medical data. To assess how much multilingual models are impacted by this, we measured the performance of BERTbase, ClinicalBERT, and BioBERT models fine-tuned only on the original English emrQA dataset on the same Paragraph QA task. In contrast, these models are not multilingual.

Table 7 compares these three models with their multilingual counterpart, mBERT. The evaluation includes three settings: monolingual fine-tuning (*mono*), fine-tuning with multilingual data augmentation (*multi*), as described earlier, and mBERT fine-tuned on train sets of all emrQA translations except the original English data (*w/o tgt*).

The results show that multilingual models perform as well as domain-specific models in our clinical QA task. Moreover, for the *Medication* subset, multilingual models outperform their domain-specific counterparts by a few percentage points. Additionally, while omitting the original English data during fine-tuning results in a performance drop, the decrease is not substantial, indicating a reasonable degree of cross-lingual transfer.

6 Conclusions

Our study focuses on the clinical QA task of finding answer evidence substrings within a given context for a specific question by multilingual models rather than domain-specific ones assessing their potential of medical support for various languages (since current clinical models are predominantly focused on English). This work investigated the effect of multilingual data augmentation in the clinical domain. Therefore, we described the MT pipeline including the process of answer evidence substring projection to translated paragraphs. Then, we compared different alignment and translation approaches. For our experiments, we used two subsets - *Medication* and *Relations* - from the emrQA dataset, translating them into six European languages: Bulgarian, Czech, Greek, Spanish, Polish, and Romanian.

During the data augmentation process, we observed that different languages pose distinct challenges for translation and subsequent QA evaluation. However, multilingual augmentation itself can be effective even in the clinical domain, as demonstrated by experiments on the *Medication* subset. However, it has a more limited effect on the *Relations* subset. However, we find that domain-specific models in our clinical QA task do not outperform multilingual models. In fact, general-domain multilingual models noticeably outperformed clinical domain-specific models on the *Medication* subset.

Limitations

This work is limited by the quality of the emrQA dataset, and our conclusions that clinical monolingual domain-specific models do not outperform multilingual general-domain models are based on a single specific clinical task evaluated in one specific language, rather than a broader range of tasks.

Acknowledgments

This research was partially supported by the SVV project number 260 698 and the Charles University GAUK grant No. 284125. It has also received support and funding from the European Union's Horizon Europe research and innovation programme project *RES-Q plus* (Grant Agreement No. 101057603). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

Mohamed Abdelghafour, Mohammed Mabrouk, and Zaki Taha. 2024. Hallucination mitigation techniques in large language models. *International Journal of Intelligent Computing and Information Sciences*, 24(4):73–81.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Mihaela Bornea, Lin Pan, Sara Rosenthal, Radu Florian, and Avirup Sil. 2021. Multilingual transfer learning for qa using translation as data augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12583–12591.

C. Carrino, Marta R. Costa-jussà, and José A. R. Fonollosa. 2020. *Automatic Spanish translation of SQuAD dataset for multi-lingual question answering*, page 5515–5523. European Language Resources Association (ELRA).

Oralie Cattan, Christophe Servan, and Sophie Rosset. 2021. On the usability of transformers-based models for a French question-answering task. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 244–255, Held Online. INCOMA Ltd.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. Meditron-70b: Scaling medical pretraining for large language models. *Preprint*, arXiv:2311.16079.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Dina Demner-Fushman, Wendy W. Chapman, and Clement J. McDonald. 2009. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772. Biomedical Natural Language Processing.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Ondřej Dušek, Jan Hajič, Jaroslava Hlaváčová, Jindřich Libovický, Pavel Pecina, Aleš Tamchyna, and Zdeňka Urešová. 2017. Khresmoi summary translation test data 2.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 conference of the North American chapter of the association for* *computational linguistics: human language technologies*, pages 644–648.

Félix Gaschi, Xavier Fontaine, Parisa Rastin, and Yannick Toussaint. 2023. Multilingual clinical ner: Translation or cross-lingual transfer? In *5th Clinical Natural Language Processing Workshop*, pages 289–311. Association for Computational Linguistics.

Aron Henriksson, Hans Moen, Maria Skeppstedt, Vidas Daudaravičius, and Martin Duneld. 2014. Synonym extraction and abbreviation expansion with ensembles of semantic spaces. *Journal of Biomedical Semantics*, 5(1):6.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2019. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

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, Hong Kong, China. Association for Computational Linguistics.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. *Preprint*, arXiv:2309.04662.

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.

Vojtech Lanz and Pavel Pecina. 2024. Paragraph retrieval for enhanced question answering in clinical documents. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 580–590, Bangkok, Thailand. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *CoRR*, abs/1901.08746.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5834–5846, Online. Association for Computational Linguistics.

Guillermo López-García, Francisco J. Moreno-Barea, Héctor Mesa, José M. Jerez, Nuria Ribelles, Emilio Alba, and Francisco J. Veredas. 2023. Named entity recognition for de-identifying real-world health records in spanish. In *Computational Science – ICCS 2023*, pages 228–242, Cham. Springer Nature Switzerland.

Zining Luo, Haowei Ma, Zhiwu Li, Yuquan Chen, Yixin Sun, Aimin Hu, Jiang Yu, Yang Qiao, Junxian Gu, Hongying Li, Xuxi Peng, Dunrui Wang, Ying Liu, Zhenglong Liu, Jiebin Xie, Zhen Jiang, and Gang Tian. 2024. Clinical large language models with misplaced focus. *Nature Machine Intelligence*, 6(12):1411–1412.

Kateřina Macková and Milan Straka. 2020. Reading comprehension in czech via machine translation and cross-lingual transfer. In *Text, Speech, and Dialogue*, pages 171–179, Cham. Springer International Publishing.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Emil Nuutinen, Iiro Rastas, and Filip Ginter. 2023. Finnish squad: A simple approach to machine translation of span annotations.

Anusri Pampari, Preethi Raghavan, Jennifer J. Liang, and Jian Peng. 2018. emrqa: A large corpus for question answering on electronic medical records. *CoRR*, abs/1809.00732.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtskỳ. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, 11(1):1–15.

Damith Premasiri, Tharindu Ranasinghe, and Ruslan Mitkov. 2023. Can model fusing help transformers in long document classification? an empirical study. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 871–878, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of* *the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Phillip Richter-Pechanski, Philipp Wiesenbach, Dominic M. Schwab, Christina Kiriakou, Nicolas Geis, Christoph Dieterich, and Anette Frank. 2024. Clinical information extraction for low-resource languages with few-shot learning using pre-trained language models and prompting. *Preprint*, arXiv:2403.13369.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Sarvesh Soni and Dina Demner-Fushman. 2025. Archehr-qa: Bionlp at acl 2025 shared task on grounded electronic health record question answering (version 1.1).

Ján Staš, Daniel Hládek, and Tomáš Koctúr. 2023. Slovak question answering dataset based on the machine translation of the squad v2.0. *Journal of Linguistics/Jazykovedný casopis*, 74(1):381–390.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael Alvers, Dirk Weißenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artieres, Axel-Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, and Georgios Paliouras. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138.

Xiang Yue, Bernal Jimenez Gutierrez, and Huan Sun. 2020. Clinical reading comprehension: A thorough analysis of the emrQA dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4474–4486, Online. Association for Computational Linguistics.

Jamil Zaghir, Mina Bjelogrlic, Jean-Philippe Goldman, Soukaïna Aananou, Christophe Gaudet-Blavignac, and Christian Lovis. 2024. FRASIMED: A clinical French annotated resource produced through crosslingual BERT-based annotation projection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7450–7460, Torino, Italia. ELRA and ICCL.

A Technical Details

This section provides additional details on fine-tuning, resource usage, and hyperparameters used in our experiments.

For alignment and translation models, default hyperparameters were used. QA models were trained with a learning rate of 3×10^{-5}, 3 epochs, weight decay of 0.01, batch size of 16, and a tokenizer processing 384-token blocks with a 128-token stride.

The experiments were carried out on nodes equipped with NVIDIA L40 GPUs (48GB per GPU).

The MT process took approximately 10 hours per language for the *Medication* subset and around 28 hours for the *Relations* subset. Alignment via Awesome required about 5 hours for the *Medication* subset and 8 hours for *Relations*. FastAlign training spanned several days, although the alignment step itself was completed in minutes.

For QA experiments, monolingual fine-tuning on the *Medication* subset took 1-4 hours (depending on model), while the *Relations* subset required 2-8 hours. Multilingual training ranged from 4–22 hours for the *Medication* subset and 8–40 hours for *Relations*.

B Clinical Performance of MT Models

Model	BLEU	METEOR	WER	CER
NLLB 600M	28.87	0.544	55.41	41.1
NLLB 1.3B dis	34.65	0.5911	50.35	37.7
NLLB 1.3B	33.02	0.5837	51.62	38.81
MadLad 3B	38.85	0.6367	45.91	34.71
NLLB 3.3B	35.04	0.6018	49.97	37.32
LINDAT	39.04	0.6337	**45.56**	34.55
MadLad 7B	38.77	0.6341	46.15	35.01
MadLad 10B	**39.28**	**0.6394**	45.61	**34.38**
NLLB 54B	38.23	0.623	47.28	35.36

Table 8: Translation from English into Czech.

Model	BLEU	METEOR	WER	CER
NLLB 600M	30.08	0.5732	52.18	38.48
NLLB 1.3B dis	31.3	0.585	51.14	37.6
NLLB 1.3B	31.4	0.5839	51.33	37.88
MadLad 3B	34.43	0.611	**49.03**	35.94
NLLB 3.3B	32.59	0.5949	50.95	37.44
LINDAT	30.77	0.5785	52.69	38.24
MadLad 7B	34.47	**0.613**	49.16	36.07
MadLad 10B	**34.7**	0.6101	**49.03**	**35.78**
NLLB 54B	33.46	0.5992	50.36	37.19

Table 9: Translation from English into German.

Model	BLEU	METEOR	WER	CER
NLLB 600M	46.67	0.713	41.43	27.82
NLLB 1.3B dis	47.65	0.7188	40.67	27.01
NLLB 1.3B	48.17	0.7224	39.93	26.94
MadLad 3B	49.21	0.7307	40.33	26.72
NLLB 3.3B	47.99	0.7218	40.68	27.17
LINDAT	47.28	0.7144	39.65	27.9
MadLad 7B	48.93	0.7305	41.03	26.87
MadLad 10B	**49.88**	**0.7364**	**39.46**	**26.4**
NLLB 54B	48.3	0.723	40.65	26.84

Table 10: Translation from English into French.

Model	BLEU	METEOR	WER	CER
NLLB 600M	13.04	0.3577	72.66	56.87
NLLB 1.3B dis	15.8	0.3948	69.78	55.27
NLLB 1.3B	15.29	0.3899	69.62	54.9
MadLad 3B	19.41	0.4403	65.37	52.33
NLLB 3.3B	16.96	0.4114	68.37	53.62
LINDAT	-	-	-	-
MadLad 7B	**20.48**	**0.4517**	64.89	51.33
MadLad 10B	19.94	0.448	**64.43**	**51.29**
NLLB 54B	18.91	0.4317	65.93	51.73

Table 11: Translation from English into Hungarian.

Model	BLEU	METEOR	WER	CER
NLLB 600M	14.97	0.3786	70.64	55.53
NLLB 1.3B dis	17.37	0.41	66.7	52.33
NLLB 1.3B	16.94	0.407	68.07	53.83
MadLad 3B	20.46	0.4545	62.33	48.11
NLLB 3.3B	18.41	0.4264	65.36	50.73
LINDAT	17.87	0.4163	65.1	50.24
MadLad 7B	**20.95**	**0.4598**	**61.8**	**47.67**
MadLad 10B	20.5	0.4546	62.1	47.9
NLLB 54B	19.24	0.4368	63.98	49.55

Table 12: Translation from English into Polish.

Model	BLEU	METEOR	WER	CER
NLLB 600M	46.09	0.7364	37.85	26.41
NLLB 1.3B dis	47.62	0.7462	37.12	26.3
NLLB 1.3B	47.19	0.7476	37.44	26.47
MadLad 3B	**49.05**	**0.7596**	**35.7**	**25.19**
NLLB 3.3B	48.05	0.7534	36.84	26.05
LINDAT	-	-	-	-
MadLad 7B	48.55	0.7555	36.27	25.72
MadLad 10B	48.27	0.7545	36.48	25.69
NLLB 54B	47.98	0.7505	36.7	26.12

Table 13: Translation from English into Spanish.

Model	BLEU	METEOR	WER	CER
NLLB 600M	41.93	0.6658	40.1	28.93
NLLB 1.3B dis	44.95	0.692	38.63	27.54
NLLB 1.3B	45.31	0.692	37.32	26.77
MadLad 3B	**52.34**	**0.748**	**31.4**	**23.07**
NLLB 3.3B	46.97	0.7059	36.55	26.17
LINDAT	-	-	-	-
MadLad 7B	51.42	0.7402	32.76	24.21
MadLad 10B	51.82	0.7437	31.78	23.14
NLLB 54B	47.26	0.7071	36.34	26.2

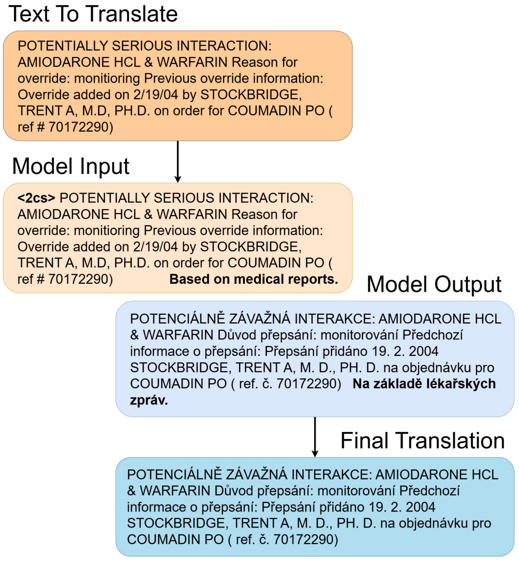
Table 14: Translation from English into Swedish.

C PMP Phrase Alternatives

Language	Translations
EN	Based on medical reports.
BG	Въз основа на медицинските доклади. Въз основа на медицински доклади. На базата на медицински доклади. Въз основа на медицински съобщения.
CS	Na základě lékařských zpráv.
EL	Βασισμένο σε ιατρικές εκθέσεις. Με βάση ιατρικές εκθέσεις. Βάσει ιατρικών εκθέσεων. Με βάση τις ιατρικές εκθέσεις. Βάσει των ιατρικών εκθέσεων. Σύμφωνα με τις ιατρικές εκθέσεις.
ES	Basado en informes médicos. Según los informes médicos. De acuerdo con los informes médicos. Con base en los informes médicos. Fundado en informes médicos.
RO	Pe baza rapoartelor medicale.
PL	Na podstawie raportów medycznych. Na podstawie sprawozdań lekarskich.

Table 15: Translations of the phrase "Based on medical reports." used as alternative phrases to look for in the translated paragraphs in the PMP MT approach.

D PMP Example



Text To Translate

POTENTIALLY SERIOUS INTERACTION: AMIODARONE HCL & WARFARIN Reason for override: monitioring Previous override information: Override added on 2/19/04 by STOCKBRIDGE, TRENT A, M.D, PH.D. on order for COUMADIN PO (ref # 70172290)

Model Input

<2cs> POTENTIALLY SERIOUS INTERACTION: AMIODARONE HCL & WARFARIN Reason for override: monitioring Previous override information: Override added on 2/19/04 by STOCKBRIDGE, TRENT A, M.D, PH.D. on order for COUMADIN PO (ref # 70172290) **Based on medical reports.**

Model Output

POTENCIÁLNĚ ZÁVAŽNÁ INTERAKCE: AMIODARONE HCL & WARFARIN Důvod přepsání: monitorování Předchozí informace o přepsání: Přepsání přidáno 19. 2. 2004 STOCKBRIDGE, TRENT A, M. D., PH. D. na objednávku pro COUMADIN PO (ref. č. 70172290) **Na základě lékařských zpráv.**

Final Translation

POTENCIÁLNĚ ZÁVAŽNÁ INTERAKCE: AMIODARONE HCL & WARFARIN Důvod přepsání: monitorování Předchozí informace o přepsání: Přepsání přidáno 19. 2. 2004 STOCKBRIDGE, TRENT A, M. D., PH. D. na objednávku pro COUMADIN PO (ref. č. 70172290)

Figure 4: Example of the MT process based on the PMP approach using the MadLad model.

E Filtration Experiments

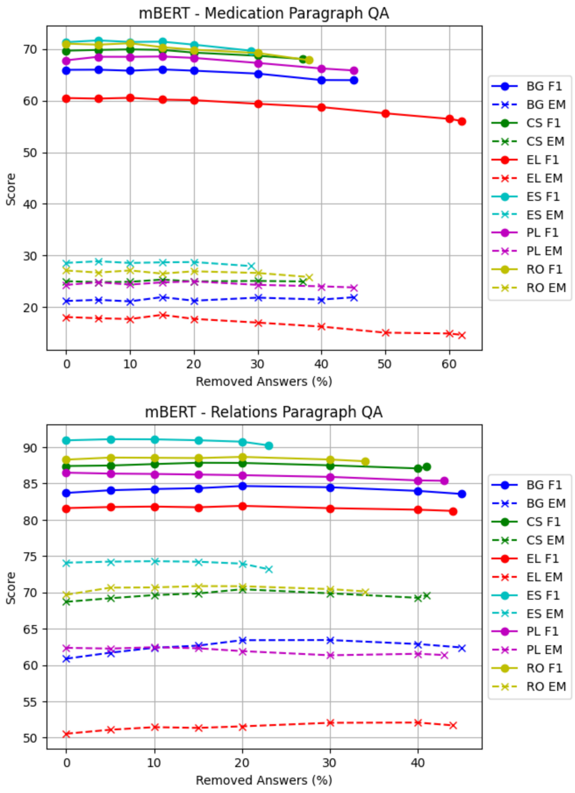


Figure 5: Filtration experiment for *Medication* and *Relations* subsets with mBERT. X-axis describes the percentage of the weakest answer evidence substrings that are removed from the training sets. Y-axis shows the F1 and EM scores of the Paragraph QA task for all translations.

F Multilingual Question Answering Results - Relations Subset

EM Score	**Full Test Set**							**Intersection Test Set**						
Models	**EN**	**BG**	**CS**	**EL**	**ES**	**PL**	**RO**	**EN**	**BG**	**CS**	**EL**	**ES**	**PL**	**RO**
distilmBERT (*mono*)	91.0	60.7	67.6	49.5	72.0	59.2	69.4	89.5	68.8	73.9	55.8	74.1	65.8	76.2
mBERT (*mono*)	92.8	63.2	70.0	51.5	74.3	61.8	70.8	90.7	71.3	76.6	57.6	76.3	68.5	77.2
XLM-R (*mono*)	93.2	63.3	71.1	52.3	75.3	62.9	72.2	91.1	70.9	77.4	58.7	77.1	69.6	79.0
XLM-R Large (*mono*)	**93.6**	64.7	72.4	**54.6**	76.2	**65.1**	73.1	**91.5**	72.8	**78.9**	**60.9**	78.1	**72.3**	80.0
distilmBERT (*multi*)	91.5	62.1	70.0	50.8	73.9	60.9	71.0	89.9	70.0	76.5	57.3	76.1	67.6	77.4
mBERT (*multi*)	92.6	63.3	70.6	52.3	75.1	62.8	72.1	90.3	71.2	77.3	58.6	76.5	70.0	78.5
XLM-R (*multi*)	93.0	64.1	72.4	53.1	75.8	63.8	72.7	91.0	72.2	78.9	59.3	77.8	70.7	79.6
XLM-R Large (*multi*)	93.2	**65.5**	**72.8**	54.1	**76.5**	64.8	**74.0**	91.0	**73.5**	**78.9**	60.8	**78.7**	71.6	**80.9**

Table 16: QA results on the *Relations* subset (EM scores) for monolingual (*mono*) and multilingual (*multi*) models.

F1 Score	**Full Test Set**							**Intersection Test Set**						
Models	**EN**	**BG**	**CS**	**EL**	**ES**	**PL**	**RO**	**EN**	**BG**	**CS**	**EL**	**ES**	**PL**	**RO**
distilmBERT (*mono*)	96.3	82.6	85.7	79.7	89.4	83.8	87.2	95.3	86.4	88.4	83.2	90.0	86.4	89.4
mBERT (*mono*)	97.3	84.5	87.7	81.9	91.0	86.2	88.6	96.1	90.4	88.2	85.2	91.5	88.8	90.8
XLM-R (*mono*)	97.4	85.2	88.6	82.5	91.7	87.2	89.5	96.2	88.7	91.0	85.7	92.1	89.6	91.7
XLM-R Large (*mono*)	**97.6**	86.1	89.5	84.3	92.2	88.7	90.3	**96.4**	89.8	92.0	87.3	92.7	91.0	92.5
distilmBERT (*multi*)	96.7	83.9	87.8	81.4	90.8	85.7	88.6	95.8	87.6	90.3	84.9	91.3	88.3	90.5
mBERT (*multi*)	97.3	85.2	88.7	83.0	91.8	87.3	89.6	96.1	88.9	91.2	86.2	92.1	89.8	91.6
XLM-R (*multi*)	97.4	85.9	89.3	83.7	92.5	88.4	90.3	96.3	89.6	91.7	86.7	93.0	90.6	92.4
XLM-R Large (*multi*)	97.5	**86.7**	**89.9**	**84.5**	**92.7**	**89.2**	**90.9**	**96.4**	**90.4**	**92.2**	**87.6**	**93.2**	**91.1**	**93.2**

Table 17: QA results on the *Relations* subset (F1 scores) for monolingual (*mono*) and multilingual (*multi*) models.