🩺MedMT5: An Open-Source Multilingual Text-to-Text LLM for The Medical Domain

Anonymous submission

Abstract

Research on language technology for the development of medical applications is currently a hot topic in Natural Language Understanding and Generation. Thus, a number of large language models (LLMs) have recently been adapted to the medical domain, so that they can be used as a tool for mediating in human-AI interaction. While these LLMs display competitive performances on automated medical texts benchmarks, they have been pre-trained and evaluated with a focus on a single language (English mostly). This is particularly true of text-to-text models, which typically require large amounts of domain-specific pre-training data, often not easily accessible for many languages. In this paper, we address these shortcomings by compiling, to the best of our knowledge, the largest multilingual corpus for the medical domain in four languages, namely English, French, Italian and Spanish. This new corpus has been used to train MedMT5, the first open-source text-to-text multilingual model for the medical domain. Additionally, we present two new evaluation benchmarks for all four languages with the aim of facilitating multilingual research in this domain. A comprehensive evaluation shows that MedMT5 outperforms both encoders and similarly sized text-to-text models for the Spanish, French, and Italian benchmarks while being competitive with current state-of-the-art LLMs in English. Data, code and models will be made publicly available upon publication.

Keywords: Natural Language Processing in Medicine, Multilingualism, Large Language Models, Deep Learning

1. Introduction

As it is the case for many application domains, there is an increasing interest in applying Artificial Intelligence (AI) and Natural Language Processing (NLP) techniques to assist medical experts in their everyday activities. With this aim in mind, in the last few years a number of large language models (LLMs) have been trained or adapted to the medical domain. These include encoder models such as SciBERT (Beltagy et al., 2019), BioBERT (Lee et al., 2020) or PubmedBERT (Gu et al., 2022). While these models have obtained state-of-the-art results in discriminative tasks, they are typically smaller in scale and scope with respect to medical text-to-text models such as SciFive (Phan et al., 2021), BioGPT (Luo et al., 2022) Med-PaLM (Singhal et al., 2022), PMC-LLaMA (Wu et al., 2023) or ClinicalGPT (Wang et al., 2023).

However, the development of all the aforementioned text-to-text LLMs has been focused on a single language, usually English. As a consequence, there is a lack of high-quality multilingual evaluation benchmarks for the medical domain. Thus, although there have been efforts to generate evaluation data in languages other than English (Wang et al., 2023; Carrino et al., 2022), they have consisted largely in monolingual approaches.

In order to address these issues, we have compiled, to the best of our knowledge, the largest multilingual corpus for training LLMs adapted to the medical domain. Our corpus includes 3B words in four languages, namely, English, Spanish, French, and Italian. While relatively small when compared to English existing datasets (Wu et al., 2023), it allowed us to build MedMT5 (Medical mT5), the first open-source text-to-text multilingual model for the medical domain.

MedMT5 is an encoder-decoder model developed by continuing the training of publicly available mT5 (Xue et al., 2021) checkpoints on medical domain data for English, Spanish, French, and Italian. Additionally, we have also created two new multilingual sequence labeling (argument component detection) and generative question answering datasets for the evaluation of multilingual LLMs in the medical domain.

A comprehensive experimental evaluation shows that MedMT5 outperforms similarly-sized text-to-text models for the Spanish, French, and Italian benchmarks while being competitive in English with respect to current state-of-the-art text-to-text (Xue et al., 2021; Chung et al., 2022) and encoder-only models (Lee et al., 2020; He et al., 2023). The results show that continuing pre-training of a multilingual text-to-text model such as mT5 allows to successfully adapt it to the medical domain, even when the amount of domain-specific data is relatively modest (ranging between 1B words for English and Spanish to 150M in Italian). Summarizing, the contributions of our work are the following: (i) the collection of the largest publicly available in-domain medical multilingual corpus for Spanish, French, and Italian languages. Together with the already existing English data, we release a corpus of 3 billion tokens. (ii) two new datasets for Spanish, French, and Italian on sequence labeling and generative Question Answering tasks, generated taking their original English versions as a starting point. (iii) the public release of two MedMT5 versions: a 770M and 3B parameter text-to-text open-source models which obtain state-of-the-art results in multilingual sequence labelling for the medical domain, most notably in multi-task and zero-shot crosslingual settings.

Other benefits of our MedMT5 models include the comparatively low hardware requirements needed for both fine-tuning on downstream tasks (the large 770M version easily fits in a 24GB V100 GPU) and for inference (a 12GB GPU should be enough). As an example, a LLaMA 7B model (Wu et al., 2023) requires at least a 80GB A100 GPU using LoRA (Hu et al., 2021) or a more demanding 4 80GB A100 GPUs without it. Code, data, models, and benchmarks will be made publicly available upon publication to facilitate reproducibility of results and encourage future multilingual research on the medical domain.

2. Related Work

As it has been the case in most application domains, Large Language Models (LLMs) have facilitated huge improvements in the state-of-the-art for medical NLP tasks (Singhal et al., 2022; Wu et al., 2023; Mayer et al., 2021). The most popular approaches are those that use models pre-trained on medical corpora such as SciBERT (Beltagy et al., 2019), BioBERT (Lee et al., 2020), PubmedBERT (Gu et al., 2022), BSC-BIO (Carrino et al., 2022) or BioLinkBERT (Yasunaga et al., 2022).

While the previous encoder-only models focused on discriminative tasks, the emergence of generative models such as LLaMa (Touvron et al., 2023), PaLM (Singhal et al., 2022) or GPT-3 (Brown et al., 2020) has resulted in a huge interest in adapting such LLMs to the medical domain. These models, to name but a few, include SciFive (Phan et al., 2021), and English T5 encoder-decoder model adapted to the scientific domain, and decoder models such as BioGPT (Luo et al., 2022), Med-PaLM (Singhal et al., 2022), PMC-LLaMA (Wu et al., 2023) and ClinicalGPT (Wang et al., 2023).

Additionally, a range of Abstractive Question Answering tasks have been proposed as evaluation benchmarks on which the larger models (Wu et al., 2023; Singhal et al., 2022; Wang et al., 2023) obtain best results. While interesting, both these LLMs and benchmarks have been developed with a focus on a single language, usually English. Furthermore, these LLMs require hardware which is simply not affordable for the large majority of end-users and researchers. In order to address these issues, we propose MedMT5, a multilingual text-to-text model adapted to the medical domain which, despite its relatively modest size and cheap running requirements, obtains competitive results, most notably in multi-task and zero-shot cross-lingual settings.

3. Compiling a Multilingual Corpus for the Medical Domain

Obtaining good quality medical corpora is usually difficult due to the sensitive nature of the data. This is even more challenging for non-English languages, as the availability of data for other languages is in general more restricted. Despite these issues, we have successfully gathered and curated a diverse collection of public relevant corpora of medical texts in English, French, Italian and Spanish to generate the MedMT5 model.

3.1. English

As listed in Table 1, we collected around 1B words from three sources related to the medical domain: (i) **ClinicalTrials** is a set of documents of clinical studies from all over the world (National Library of Medicine, 2022a); (ii) **EMEA** is an English-Spanish parallel corpus with documents provided by the European Medicines Agency (Tiedemann, 2012) and, (iii) PubMed (National Library of Medicine, 2022b), which contains data from various sources such as MEDLINE, life science journals and online books, provides the bulk of the English data.

Source	Words
ClinicalTrials	127.4M
EMEA	12M
PubMed	968.4M

Table 1: English data sources and word counts.

3.2. Spanish

Apart from **EMEA** and **PubMed**, which we also used for Spanish, the biggest portion of the data came from the **Medical Crawler**, a biomedical corpus compiled by Carrino et al. (2022). Additionally, we also included **SPACC** (Ander Itxaurrondo, 2018), **UFAL** (Institute of Formal and Applied Linguistics, 2017) and **WikiMed**, a corpus built ad-hoc from Wikipedia entries. Table 2 provides the details of the collected data, which amounts to ≈1B words.

Source	Words
EMEA	13.6M
PubMed	8.4M
Medical Crawler	918M
SPACC	350K
UFAL	10.5M
WikiMed	5.2M

Table 2: Spanish data sources and word counts.

3.3. French

A total of 7,192,779 sentences and 670,972,717 words were compiled using the data sources listed in Table 3.

Source	Words
PubMed	1.4M
Science Direct	15.2M
Wikipedia - Médecine	5M
EDP	48K
Google Patents	654M

Table 3: French data sources and word counts.

PubMed data was extracted using the `Bio.Entrez` package[1]. **Science Direct** offers a collection of scientific and medical publications which can be extracted via their the official API[2]. We filtered relevant articles with the keyword "Médecine", and the obtained XML documents were parsed to extract the `<dc:description>` tag. As for Spanish, we took advantage of **Wikipedia** as a source of medical knowledge to obtain HTML formatted data from the category "Category:Médecine". **The EDP French/English Parallel Medical Corpus** (Jimeno-Yepes et al., 2017) provides bilingual content from journals that address domains such as dentistry and life sciences. From this source we downloaded the dataset labeled "EDP French corpus, text format". Finally, **Google Patents** is a comprehensive repository of patent data from around the world. Google Patents data were retrieved by filtering using the IPC code and abstract language.

A final French language verification step was undertaken by applying the `langdetect` package (version 1.0.9).

3.4. Italian

The crawling and pre-processing of the Italian split of the corpus followed the methodology described by Carrino et al. (2022). First, a list of 504 medical terms, which we use as seeds to scrape the Italian split of the **MC4 Common Crawl Corpus** (Common Crawl, 2022) by only selecting the pages which contained at least one of the keywords in their URL domain. To create the list, we extracted 600 keyword terms related to medicine from the *Dizionario analogico della Lingua Italiana* (Zanichelli). We excluded some sectors and discarded terms that may lead to ambiguous queries (e.g., actions, which contained mainly verbs, proverbs, general terms like "assistente", etc.). We normalized rare variants ("bacteriologia" to "batteriologia") and stemmed all terms without lemmatizing, as most terms are already lemmatized in the dictionary; we performed univerbation of multiword units (e.g., "esamedelleutine", "follow-up"), and removed the duplicates. This resulted in a corpus of 67 million tokens, which we joined with other sources of text such as **Medical dissertations**, **Drug use instructions**, **PubMed abstracts**, etc. as detailed in Table 4, resulting in a ≈145M word corpus.

Source	Words
Medical Commoncrawl - IT	67M
Drug instructions	30.5M
Wikipedia - Medicina	13.3M
E3C Corpus - IT	11.6M
Medicine descriptions	6.3M
Medical theses	5.8M
Medical websites	4M
PubMed	2.3M
Supplement description	1.3M
Medical notes	975K
Pathologies	157K
Medical test simulations	26K
Clinical cases	20K

Table 4: Italian data sources and word counts.

4. 🩺MedMT5

Multilingual T5 (mT5) (Xue et al., 2021) is an extension of the original T5 (Raffel et al., 2020) framework, which is optimized for multilingual tasks. The T5 model is grounded in the transformer encoder-decoder architecture (Vaswani et al., 2017). With its decoder block, T5 is capable of generating sequences of tokens in an auto-regressive fashion. T5 was designed to convert every NLP problem into a text-to-text task, and mT5 extends this strategy to a multitude of languages, leveraging a shared vocabulary for diverse scripts. mT5 was trained using mC4, a 1 Trillion token Common Crawl-based dataset covering 101 languages. The pre-training is based on a masked language modeling "span-corruption" objective, where consecutive spans of input tokens are replaced with a mask and the model is trained to reconstruct the masked-out tokens.

4.1. Pre-training MedMT5

MedMT5 is built upon the same architecture as mT5 (Xue et al., 2021). We release two diffent models: MedMT5-Large (738M parameters) and MedMT5-XL (3 billion parameters). Both models were initialized using the pre-trained weights of

---

[1] https://biopython.org/docs/1.75/api/Bio.Entrez.html

[2] https://dev.elsevier.com/

their corresponding mT5 checkpoints and continued their pre-training using the 3B word medical domain dataset described in Section 3 (with x2 oversampling for the Italian split). To prevent over-fitting, we run the training for only one epoch, as preliminary experiments showed that performance degraded with more epochs. We adhered to the self-supervised parameter settings recommended by Xue et al. (2021) and detailed in Table 5. It should be noted that MedMT5-Large was trained with a sequence length of 1024 tokens whereas MedMT5-XL was limited to a sequence length of 480 tokens due to GPU memory limitations. MedMT5 was trained using the Flax implementation of mT5 in the Hugging Face transformers library (Wolf et al., 2020). All experiments were conducted on our private servers, employing 4xA100 80GB GPUs. We made calculations for a carbon footprint estimation based on a 400W consumption per GPU and a carbon intensity of 0.171 kg/kWh[3].

	MedMT5-Large	MedMT5-XL
Param. no.	738M	3B
Sequence Lenght	1024	480
Token/step	65536	30720
Epochs	1	1
Total Tokens	4.5B	4.5B
Optimizer	Adafactor	Adafactor
LR	0.001	0.001
Scheduler	Constant	Constant
Hardware	4xA100	4xA100
Time (h)	10.5	20.5
CO_2eq (kg)	2.9	5.6

Table 5: Pre-Training settings for MedMT5.

5. Generating New Multilingual Benchmarks

The lack of multilingual evaluation benchmarks for the medical domain motivated us to generate new evaluation data for our languages of interest, as only the relatively small E3C (Magnini et al., 2021) was already available for all 4 languages. We focused on two different types of tasks: (i) a sequence labelling task, **Argument Mining**, consisting in detecting and classifying the argument component spans and their relations, (ii) **Abstractive Question Answering**, where the model is expected to generate an answer in response to an input question. In both cases we took existing English labelled data as a starting point.

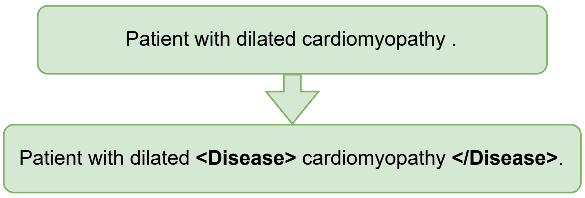


Figure 1: Text-to-Text representation of the Sequence Labeling task. Given an input sentence, the model is expected to generate the same sentence annotated with html-style tags.

5.1. Argument Mining

The AbstRCT dataset is composed by English medical and scientific texts collected from the MEDLINE database and manually annotated with two types of argument components: Claims and Premises (Mayer et al., 2021). We generated French and Italian parallel versions of the dataset using the same method as for Spanish, based on machine translation and semi-manual annotation projection (Yeginbergenova and Agerri, 2023). The AbstRCT dataset is divided in three splits, neoplasm, glaucoma and mixed. Following previous work, we fine-tune the models with the first one and then evaluate the in-domain performance on the neoplasm test split and the cross-domain performance on the glaucoma and mixed splits. Previous works using the AbstRCT datasets have employed different definitions of the F1 score metric, such as token-level F1 (Mayer et al., 2021; Yeginbergenova and Agerri, 2023). However, in this paper we report results using the standard sequence level F1 score (Tjong Kim Sang and De Meulder, 2003), a much more strict metric, which explains the lower results for all the models.

5.2. Question Answering

We use the BioASQ-6B English Question Answering dataset (Tsatsaronis et al., 2015) to generate parallel French, Italian and Spanish versions. Given a biomedical question and a set of snippets of text with relevant information about the question, the model must generate the *ideal* answer. A set of ideal gold answers are provided to assess the performance of the models. We machine translated the questions and ideal answers into French, Italian and Spanish using the NLLB200 3B parameter model (Costa-jussà et al., 2022).

6. Experimental Setup

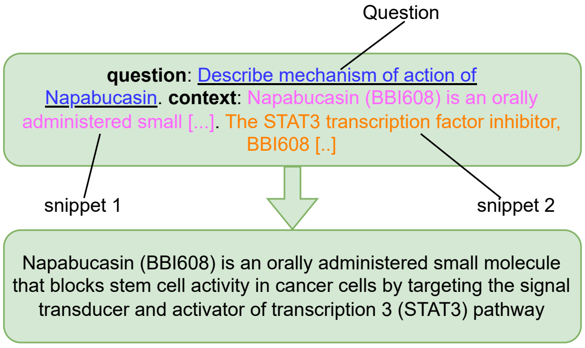
MedMT5 is a text-to-text model. This means that, given a text input, it learns to generate a text as output. Therefore, every evaluation task must be converted into a text-to-text format (Xue et al., 2021). In

---

[3] Sourced from `https://app.electricitymaps.com/map`

Representation	Task	Dataset	Languages	Entity Type
Sequence Labelling	Named Entity Recognition	NCBI-Disease (Dogan et al., 2014)	EN	Disease
		BC5CDR Disease (Li et al., 2016)	EN	Disease
		BC5CDR Chemical (Li et al., 2016)	EN	Chemical
		DIANN (Fabregat et al., 2018)	EN, ES	Disability
		E3C (Magnini et al., 2021)	EN, ES, FR, IT	Clinical Entity
		PharmaCoNER (Gonzalez-Agirre et al., 2019)	ES	Pharmacological
	Argument Mining	AbstRCT (Mayer et al., 2021)	EN, ES, FR, IT	Claims and Premises
Generative Question Answering	Question Answering	BioASQ 6B (Tsatsaronis et al., 2015)	EN, ES, FR, IT	Biomedical QA

Table 6: List of evaluation tasks used to measure the performance of MedMT5.



question: Describe mechanism of action of Napabucasin. **context**: Napabucasin (BBI608) is an orally administered small [...]. The STAT3 transcription factor inhibitor, BBI608 [..]

Napabucasin (BBI608) is an orally administered small molecule that blocks stem cell activity in cancer cells by targeting the signal transducer and activator of transcription 3 (STAT3) pathway

Figure 2: Text-to-Text representation of the BioASQ task. Given a question and a set of relevant snippets, the model generates an answer.

our experiments the output text is always generated using beam search with 4 beams.

The list of tasks used for evaluation is listed in Table 6. The **Sequence Labelling tasks** include medical NER, detecting and classifying named entities according to some pre-defined categories, and Argument Mining, described in Section 5. Performance for every sequence labelling task is evaluated using standard sequence level F1 score (Tjong Kim Sang and De Meulder, 2003).

In order to address sequence labelling tasks, text-to-text models such as MedMT5 are prompted with the sentence to label. As illustrated in Figure 1, the expected output is the same sentence annotated with HTML-style tags. The HTML tags for each task are added as special tokens to the model vocabulary. Furthermore, we use constrained decoding to ensure that the output contains the same words as the input and a valid HTML annotation.

With respect to the BioASQ **Abstractive Question Answering task**, the input prompt contains the question and a context. As shown in Figure 2, the context is generated by concatenating all the provided possible snippets. The expected output should be the generated answer to the question, which is then compared to the gold ideal answer.

6.1. Baselines

As we have developed MedMT5 by continuing the training of mT5 checkpoints, our primary point of comparison should be mT5 (Xue et al., 2021). Thus, our first objective is to assess whether training the model on our multilingual medical-domain corpus enhances its performance for tasks specific to this domain. Furthermore, we also benchmark our model against SciFive (Pubmed+PMC) a T5-based 738M parameter model (Phan et al., 2021) trained exclusively on a corpus of 78B words containing scientific and medical English data. Additionally, we compare the performance of MedMT5 with Flan-T5 (Chung et al., 2022), which also adopts the T5 architecture but has been finetuned on a huge instruction-following dataset for almost 2K tasks. Flan-T5 achieves state-of-the-art performance in numerous benchmarks, including some from the the medical domain (Singhal et al., 2022). We tested all three types of text-to-text models under identical settings and hyperparameters.

We also measure MedMT5 with the performance of encoder-only models in sequence labelling tasks. We report results with mDeBERTaV3 (He et al., 2023) which is widely used for sequence labeling and excels in multilingual tasks (Adelani et al., 2022; Agerri and Agirre, 2023). Although we also tested XLM-RoBERTa (Conneau et al., 2020) and GLOT500 (Imani et al., 2023), their results were worse than those obtained by mDeBERTaV3. Finally, we also compare with BioBERT v1.1 (Lee et al., 2020), which has been pretrained on a large English-only biomedical dataset. We do not evaluate the performance of encoder-only models in the question answering task, as their architecture is not designed for text generation.

The specific hyperparameter settings used to fine-tune the models will be available in the Appendix upon publication.

7. Experimental Results

In this section, we report on the performance of MedMT5 and of the baselines in the **sequence**

Lang	Dataset	mT5$_{large}$	mT5$_{XL}$	SciFive	FlanT5$_{large}$	FlanT5$_{XL}$	mDeBERTa$_{V3\ base}$	BioBERT	MedMT5$_{large}$	MedMT5$_{XL}$
EN	NCBI-Disease	85.1	87.7	**89.4**	88.6	89.3	85.7	87.4	89.1	87.2
EN	BC5CDR Disease	78.5	81.4	85.4	85.0	**85.8**	82.5	84.3	84.4	82.4
EN	BC5CDR Chemical	89.1	90.8	**93.3**	92.0	92.9	91.1	92.9	92.8	91.3
EN	DIANN	70.1	77.8	71.9	74.4	74.2	**80.3**	79.0	74.8	77.6
ES	DIANN	72.4	74.9	70.5	70.7	70.9	**78.3**	70.2	74.9	74.8
EN	E3C	54.3	60.1	62.8	**64.2**	63.1	58.2	58.6	59.4	57.9
ES	E3C	61.6	71.7	62.7	64.4	67.1	65.9	57.4	**72.2**	69.5
FR	E3C	55.6	64.9	61.7	65.2	64.3	62.0	53.3	65.2	**65.8**
IT	E3C	61.8	63.8	59.6	61.9	65.1	63.9	52.1	**67.5**	65.9
ES	PharmaCoNER	86.3	90.6	87.5	88.5	89.1	89.4	88.6	**90.8**	90.1
EN	Neoplasm	68.2	71.0	**74.8**	71.7	73.0	64.5	67.5	72.5	72.2
EN	Glaucoma	69.0	72.6	78.4	76.7	**78.5**	71.2	74.8	76.8	75.0
EN	Mixed	67.3	69.1	73.4	72.8	**74.4**	63.4	69.6	71.5	71.0
ES	Neoplasm	45.8	56.1	45.8	38.5	47.9	**63.0**	57.1	36.8	35.1
ES	Glaucoma	68.1	70.7	**77.0**	73.2	76.7	68.6	64.5	76.9	75.6
ES	Mixed	65.8	66.2	68.8	70.2	**72.9**	61.3	58.9	71.6	70.4
FR	Neoplasm	63.5	66.6	68.9	68.9	**69.7**	56.3	50.0	69.3	65.5
FR	Glaucoma	63.2	69.2	71.7	69.3	**72.8**	60.3	52.5	71.0	71.8
FR	Mixed	61.0	65.4	68.0	66.8	**68.8**	54.1	48.9	68.6	66.3
IT	Neoplasm	64.2	69.9	68.5	69.1	70.7	61.1	55.7	**71.7**	70.6
IT	Glaucoma	65.8	71.5	72.9	69.9	74.4	67.3	59.3	**76.1**	72.9
IT	Mixed	64.6	67.7	69.1	69.8	71.5	62.0	55.8	71.3	**71.8**
AVERAGE		67.3	71.8	71.9	71.5	**73.3**	68.7	65.4	73.0	71.9
AVERAGE ES, FR, IT		64.3	69.2	68.0	67.6	70.1	65.3	58.9	**70.3**	69.0

Table 7: Single-task supervised F1 scores for Sequence Labelling.

labelling tasks across different settings. Due to space constraints, we only report the best performing results.

Single Task Monolingual Supervised Results: The results when fine-tuning and evaluating the models for each dataset and language are shown in Table 7. The first observation is that MedMT5-large significantly outperforms both mT5-large and mT5-XL, demonstrating the benefits of further training these models with our multilingual medical domain corpus.

When comparing MedMT5 with FlanT5 and SciFive, the latter models are systematically superior on English. This was anticipated since both have been pre-trained with a much larger amount of English-only data. With respect to encoder-only models, they achieve in general worse results than text-to-text models across all tasks and languages (except for the DIANN dataset). It is also noteworthy that FlanT5-XL exhibits robust performance across all datasets and languages, even though it was fine-tuned with English-only data not specific to the medical domain. Nonetheless, MedMT5-large obtains in general better results for French, Spanish and Italian while being much smaller in size (738M parameters vs 3B parameters), showing the impact of training MedMT5 with domain-specific data for those languages.

Multi-Task Supervised Results: Text-to-text models have demonstrated improved performance when trained in multi-task settings (Chung et al., 2022). Following this, we also experimented with fine-tuning them across all the sequence labeling tasks simultaneously. To inform the model about which labels should classify for each input example, we add the list of predefined labels from the corresponding dataset to the beginning of the input sentence. For instance, the input depicted in Figure 1 is adjusted to *"<Disease> Patient with dilated cardiomyopathy"*. A comparison of the Single Task and Multi-Task settings is presented in Table 8. It can be seen that in this setting MedMT5 achieves the best overall results for Spanish, French and Italian. On average, MedMT5-XL also obtains the best performance, slightly improving over the results of FlanT5-XL and MedMT5-large.

Zero-shot Cross-Lingual Transfer Results: Manually annotated medical domain datasets for languages other than English are scarce. Therefore, developing models that can successfully generate predictions for languages different to those used for fine-tuning is crucial. We evaluate this ability to perform zero-shot cross-lingual transfer by fine-tuning MedMT5 and the baselines on the English AbsRCT Neoplasm dataset, and then evaluating them on the Neoplasm, Glaucoma, and Mixed datasets for Spanish, French, and Italian. The results are presented in Table 9. Results show that MedMT5 significantly outperforms any other model. Moreover, MedMT5-XL achieves significantly better results than MedMT5-large. The Spanish Neoplasm subset represents an anomaly in the performance of the models, with all text-to-text models underperforming. This issue aligns with previous results (García-Ferrero et al., 2022) and, upon man-

Lang	Dataset	Single Task			MultiTask		
		FlanT5$_{XL}$	MedMT5$_{large}$	MedMT5$_{XL}$	FlanT5$_{XL}$	MedMT5$_{large}$	MedMT5$_{XL}$
EN	NCBI-Disease	**89.3**	89.1	87.2	87.6	87.6	86.9
EN	BC5CDR Disease	**85.8**	84.4	82.4	85.1	83.4	83.0
EN	BC5CDR Chemical	**92.9**	92.8	91.3	92.7	92.5	91.6
EN	DIANN	74.2	74.8	77.6	**80.0**	75.4	75.3
ES	DIANN	70.9	74.9	74.8	**77.1**	72.6	73.6
EN	E3C	**63.1**	59.4	57.9	62.1	60.9	62.0
ES	E3C	67.1	72.2	69.5	66.5	**74.9**	73.3
FR	E3C	64.3	65.2	**65.8**	62.9	65.4	65.1
IT	E3C	65.1	**67.5**	65.9	60.7	66.9	65.1
ES	PharmaCoNER	89.1	**90.8**	90.1	89.9	90.3	89.5
EN	Neoplasm	73.0	72.5	72.2	**74.1**	72.7	73.1
EN	Glaucoma	**78.5**	76.8	75.0	76.9	76.0	76.1
EN	Mixed	74.4	71.5	71.0	**75.1**	71.0	73.4
ES	Neoplasm	47.9	36.8	35.1	**72.7**	43.3	72.3
ES	Glaucoma	76.7	76.9	75.6	76.6	76.1	**77.6**
ES	Mixed	**72.9**	71.6	70.4	70.9	69.9	70.1
FR	Neoplasm	69.7	69.3	65.5	**69.7**	68.3	68.3
FR	Glaucoma	72.8	71.0	71.8	71.0	**73.9**	72.1
FR	Mixed	**68.8**	68.6	66.3	68.7	65.7	67.0
IT	Neoplasm	70.7	71.7	70.6	69.2	72.5	**72.5**
IT	Glaucoma	74.4	76.1	72.9	73.7	76.8	**77.4**
IT	Mixed	71.5	71.3	71.8	72.1	70.7	**73.6**
AVERAGE		73.3	73.0	71.9	74.3	73.0	**74.5**
AVERAGE ES, FR, IT		70.1	70.3	69.0	71.6	70.5	**72.7**

Table 8: Multi-task supervised F1 scores for Sequence Labelling.

Lang	Dataset	mT5$_{XL}$	SciFive	FlanT5$_{XL}$	mDeBERTa$_{V3\ base}$	MedMT5$_{large}$	MedMT5$_{XL}$
ES	Neoplasm	56.9	39.2	29.4	**62.9**	33.5	33.8
ES	Glaucoma	63.0	68.0	67.0	68.3	68.4	**72.2**
ES	Mixed	61.4	65.9	64.4	60.8	64.6	**69.1**
FR	Neoplasm	61.0	58.7	61.0	58.1	65.0	**66.6**
FR	Glaucoma	60.5	64.8	61.7	61.7	64.6	**68.1**
FR	Mixed	60.1	60.1	59.9	57.8	62.4	**68.4**
IT	Neoplasm	56.0	55.6	62.2	54.4	64.3	**64.8**
IT	Glaucoma	66.7	58.5	63.1	63.1	69.3	**71.9**
IT	Mixed	59.7	55.5	60.6	56.0	64.0	**68.2**
AVERAGE		60.6	58.5	58.8	60.3	61.8	**64.8**

Table 9: Zero-shot F1 scores for Argument Mining. Models have been trained in English and evaluated in Spanish, French and Italian.

ual inspection, it seems to be caused by errors in the annotation of argument component spans.

To summarize, MedMT5 stands out for its superior performance in the evaluation for Spanish, French, and Italian languages, especially for the multitask and the zero-shot transfer settings. These capabilities can help mitigate the scarcity of manually annotated medical data for other target languages. In contrast, SciFive and FlanT5, having been trained on extensive English-only datasets, emerge as the top choices when the primary focus is on English-only tasks.

Finally, despite MedMT5-XL being larger than MedMT5-large (3B vs 738M), its performance is worse in the single-task evaluation setting. This behaviour is not observed in the multi-task and zero-shot experiments, leading us to hypothesize that the larger MedMT5-XL model is more prone to overfit in the single-task supervised setting.

7.1. Abstractive Question Answering

In this section, we explore the text generation capabilities of MedMT5 and other baseline text-to-text models on the BioASQ question answering dataset described in Section 5.2. Previous work typically evaluate the performance on this task using the ROUGE score (Tsatsaronis et al., 2015) to compare the gold standard answer with the answer generated by the model. However, we find this metric inadequate for medical domain tasks as it does not address crucial aspects of the generation such as factuality, potential harm, and bias (Singhal et al., 2022). Consequently, we enlisted medical professionals to analyze the answers produced by the models.

Ideal Answer	Which type of urinary incontinence is diagnosed with the Q tip test? Stress urinary incontinence is diagnosed with the Q tip test. The test evaluates urethral mobility.
mT5-large	Q-tip test is used for urodynamic assessment of patients with severe urinary incontinence.
FlanT5-large	The Q tip test is used for assessing urethral mobility.
SciFive	Q tip test is used for evaluation of stress urinary incontinence.
MedMT5-large	Q-tip test is used for evaluation of stress urinary incontinence.
Ideal Answer	Which are the main manifestations of Ohdo syndrome? Severe ID, absent or deficient language, skeletal manifestations including bilateral patella dislocations.
mT5-large	Skeletal manifestations in Ohdo syndrome are a case with bilateral patella dislocations where surgical intervention has been indicated.
FlanT5-large	The main manifestations of Ohdo syndrome are: 1) severe ID, 2) absent or deficient language and 3) milder, clinical manifestation in heterozygotes.
SciFive	Ohdo syndrome is characterized by severe ID, absent or deficient language and, milder, clinical manifestation in heterozygotes.
MedMT5-large	The main manifestations of Ohdo syndrome are: 1) absent or deficient language and 2) mildder clinical manifestation in heterozygotes.

Table 10: Examples of answers generated by each model for two different BioASQ questions together with the rank assigned by medics.

During annotation, medical doctors were displayed the question, the ideal gold answers and the answers generated by each model. If required, they could also inspect the snippets that provide context to answer each of the questions. We narrowed the evaluation to MedMT5-large, mT5-large, FlanT5-large and SciFive. The evaluation was conducted by medical doctors proficient/native speakers of English, French and Spanish. For each question, doctors were asked to rank the answers generated by the models as the best, second-best, third-best, and worst answer.

Two Spanish medical doctors proficient or native in English and Spanish analyzed 50 English examples and 252 Spanish. With respect to French, 3 French clinicians analyzed 186 answers, of which 47 were done by 2 doctors to calculate IAA (Cohen's Kappa Score: 0.28 and Average Spearman's Rank Correlation: 0.48), which indicates a low level of agreement. This exercise provided interesting insights with respect to the performance of the models in text generation tasks in the medical domain. First, medical doctors could not in general establish significant differences between the answers generated by each of the models; predictions were far too similar, and all tended to fail on the same questions. As an example, Table 10 shows the answers to two different questions. As it can be observed, the answers generated by each model are very similar, and the doctors ended up ranking them primarily based on style.

The final result of the manual analysis is that all the models were chosen a similar number of times as the best. We believe that this demonstrates the difficulty of performing and obtaining meaningful evaluation results for this kind of tasks on this specific domain. This is in fact supported by the low IAA agreement obtained in the French annotation.

However, there could other underlying reasons for this behaviour. First, perhaps the T5 architecture is not ideally suited for text generation as formulated in the BioASQ task as these models are trained on a masking reconstruction objective rather than on direct text generation tasks. Consequently, the knowledge acquired during pre-training might not generalize well when the models are subsequently trained for text generation purposes. Second, perhaps using much larger models such as MedPaLM (Singhal et al., 2022) may generate better answer generation, but models of 540B parameters are currently unusable for the large majority of the NLP research labs, including ours. Nonetheless, it should be stressed that research on appropriate evaluation metrics for these tasks is still a difficult challenge which requires further investigation. In any case, our results demonstrate the potential of a text-to-text model such as MedMT5 for multilingual sequence labelling in the medical domain, establishing new state-of-the-art results in the multi-task and zero-shot cross-lingual settings.

8. Conclusion

In this paper we have presented MedMT5, the first open source multilingual text-to-text LLM for the medical domain. Its development has required the compilation of a new 3B word corpus in English, French, Italian and Spanish specific to the medical domain. Furthermore, motivated by the lack of multilingual benchmarks, we have generated evaluation benchmarks for French, Italian and Spanish for Argument Mining and Abstractive Question Answering.

A comprehensive experimentation on sequence labelling tasks shows that MedMT5 outperforms strong text-to-text baselines of similarly-sized models in the multi-task and zero-shot cross-lingual evaluation settings. This is particularly interesting as these settings fully exploit the multilingual nature of a text-to-text model such as MedMT5.

9. Ethical Statement

Our research in developing MedMT5, a multilingual text-to-text model for the medical domain, has ethical implications that we acknowledge. Firstly, the broader impact of this work lies in its potential to improve medical communication and understanding across languages, which can enhance healthcare access and quality for diverse linguistic communities. However, it also raises ethical considerations related to privacy and data security. To create our multilingual corpus, we have taken measures to anonymize and protect sensitive patient information, adhering to data protection regulations in each language's jurisdiction or deriving our data from sources that explicitly address this issue in line with privacy and safety regulations and guidelines. Furthermore, we are committed to transparency and fairness in our model's development and evaluation. We have worked to ensure that our benchmarks are representative and unbiased, and we will continue to monitor and address any potential biases in the future. Finally, we emphasize our commitment to open source by making our data, code, and models publicly available, with the aim of promoting collaboration within the research community.

10. Bibliographical References

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rodrigo Agerri and Eneko Agirre. 2023. Lessons learned from the evaluation of Spanish Language Models. *Proces. del Leng. Natural*, 70:157–170.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. 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 3615–3620.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Casimiro Pio Carrino, Joan Llop, Marc Pàmies, Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Joaquín Silveira-Ocampo, Alfonso Valencia, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. Pretrained biomedical language models for clinical NLP in Spanish. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 193–199, Dublin, Ireland. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

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, Loïc 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. *CoRR*, abs/2207.04672.

Rezarta Islamaj Dogan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: A resource for disease name recognition and concept normalization. *J. Biomed. Informatics*, 47:1–10.

Hermenegildo Fabregat, Juan Martínez-Romo, and Lourdes Araujo. 2018. Overview of the DIANN task: Disability annotation task. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 1–14. CEUR-WS.org.

Iker García-Ferrero, Rodrigo Agerri, and German Rigau. 2022. T-projection: High quality annotation projection for sequence labeling tasks. *CoRR*, abs/2212.10548.

Aitor Gonzalez-Agirre, Montserrat Marimon, Ander Intxaurrondo, Obdulia Rabal, Marta Villegas, and Martin Krallinger. 2019. PharmaCoNER: Pharmacological substances, compounds and proteins named entity recognition track. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, BioNLP-OST@EMNLP-IJNCLP 2019, Hong Kong, China, November 4, 2019*, pages 1–10. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Heal.*, 3(1):2:1–2:23.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *arXiv preprint*, 2106.09685.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André F. T. Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1082–1117. Association for Computational Linguistics.

Antonio Jimeno-Yepes, Aurélie Névéol, Mariana L. Neves, Karin Verspoor, Ondrej Bojar, Arthur Boyer, Cristian Grozea, Barry Haddow, Madeleine Kittner, Yvonne Lichtblau, Pavel Pecina, Roland Roller, Rudolf Rosa, Amy Siu, Philippe Thomas, and Saskia Trescher. 2017. Findings of the WMT 2017 biomedical translation shared task. In *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 234–247. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. Biocreative V CDR task corpus: a resource for chemical disease relation extraction. *Database J. Biol. Databases Curation*, 2016.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Bernardo Magnini, Begoña Altuna, Alberto Lavelli, Manuela Speranza, and Roberto Zanoli. 2021. The E3C project: European clinical case corpus. In *Proceedings of the Annual Conference of the Spanish Association for Natural Language Processing: Projects and Demonstrations (SEPLN-PD 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), Málaga, Spain, September, 2021*, volume 2968 of *CEUR Workshop Proceedings*, pages 17–20. CEUR-WS.org.

Tobias Mayer, Elena Cabrio, and Serena Villata. 2020. Transformer-based argument mining for healthcare applications. In *ECAI 2020 - 24th*

European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020), volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2108–2115. IOS Press.

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.

Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. SciFive: a text-to-text transformer model for biomedical literature. *CoRR*, abs/2106.03598.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Schärli, Aakanksha Chowdhery, Philip Andrew Mansfield, Blaise Agüera y Arcas, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle K. Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large language models encode clinical knowledge. *arXiv preprint*, abs/2212.13138.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint*, 2302.13971.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artières, Axel-Cyrille Ngonga Ngomo, Norman Heino, Éric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinform.*, 16:138:1–138:28.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Guangyu Wang, Guoxing Yang, Zongxin Du, Longjun Fan, and Xiaohu Li. 2023. ClinicalGPT: Large Language Models Finetuned with Diverse Medical Data and Comprehensive Evaluation. *ArXiv preprint*, abs/2306.09968.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies. Association for Computational Linguistics.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8003–8016. Association for Computational Linguistics.

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

11. Language Resource References

Ander Itxaurrondo. 2018. *SPACCC: Spanish Clinical Case Corpus*. Barcelona Supercomputing Center. PID https://github.com/PlanTL-GOB-ES/SPACCC.

Common Crawl. 2022. *Common Crawl*. Common Crawl. PID https://commoncrawl.org/.

Institute of Formal and Applied Linguistics. 2017. *UFAL Medical Corpus v. 1.0*. Charles University, Czech Republic. PID https://ufal.mff.cuni.cz/ufal_medical_corpus.

National Library of Medicine. 2022a. *Clinical Trials*. National Library of Medicine. PID https://clinicaltrials.gov/.

National Library of Medicine. 2022b. *PubMed*. National Library of Medicine. PID https://pubmed.ncbi.nlm.nih.gov.