Cross-Lingual Medical Knowledge Transfer with Large Language Models

WashU CSE 527A Final Project Report

Vincent Siu
Washington University in St. Louis
vincent.siu@wustl.edu

Boshen Wang
Washington University in St. Louis
boshen@wustl.edu

Ziwei Wang
Washington University in St. Louis
ziwei.wang@wustl.edu

Abstract

In the field of Natural Language Processing (NLP), Large Language Models (LLMs) such as OpenAI's GPT series and Google's PaLM family have significantly advanced multilingual LLM capabilities. However, datasets in non-English languages are still relatively scarce compared to English resources, and the ability of multilingual LLMs to specialize in specific domains, particularly those with complex terminologies like the medical field, remains untested. This project addresses this issue by fine-tuning the BigScience Large Open-science Open-access Multilingual (BLOOM) model on the MedMCQA dataset using LoRA and evaluating the extent to which subsequent knowledge carries over to non-English language tasks. The evaluation of the performance of the fine-tuned BLOOM model and an untrained PaLM-2 model in a zero-shot context in Spanish and Vietnamese shows that fine-tuning in English creates performance increases in non-English medical question answering tasks.

1 Introduction

The field of Natural Language Processing (NLP) has seen a transformative shift with the advent of Large Language Models (LLMs) such as OpenAI's GPT series and Google's PaLM family. These models, trained on extensive text datasets and equipped with billions of parameters, are capable of outputting text in many different languages. Additionally, datasets for specialized tasks such as question answering are scarce in non-English languages as opposed to English.

This presents a problem given the need for non-English text generation in fields with specialized vocabulary, such as the medical field. Due to language barriers between healthcare providers and patients, translators and interpreters are both vital and necessary in delivering healthcare information to patients with limited English proficiency. Research indicates that despite translation and interpreting resources, the health outcomes of limited English proficiency patients is still generally worse than that of English speaking patients (Fernández et al., 2017; Kim et al., 2017; Schenker et al., 2010).

Despite this issue, multilingual language models and LLMs in general have not been evaluated for their ability to respond to medical questions in non-English fields, likely due to scarcity of datasets in non-English languages (Hershcovich et al., 2022).

As a result, our project aims to address the problem of domain adaptation of multilingual models to the medical domain. The challenge is twofold: to have the model learn the complexity of medical terminology and concepts, and to then apply that information to languages other than English. Therefore, the purpose of our project aims to establish and explore methods for utilizing English medical datasets to guide the tuning and optimization of medical question answering in non-English datasets.

To tackle this, we fine-tune the BigScience Large Open-science Open-access Multilingual (BLOOM) model, a collaborative LLM funded by the French Government, on the MedMCQA dataset. Due to time and computational resource constraints, we loaded, fine-tuned, and evaluated using the 3 billion parameter version of BLOOM. Our approach involves using Low-Rank Adaptation (LoRA)(Hu et al., 2021) to freeze the pre-trained model weights and instead train injected low-rank matrices in each layer of the transformers architecture. Then, we compare the base model to the trained model and evaluated the dataset in a zero-shot context in Spanish and Vietnamese. Additionally, we assessed the performance of an untrained PaLM-2 (Anil et al., 2023)model on these languages to establish a state of the art multilingual LLM benchmark.

2 Related Work

Medical text analysis has seen significant advances in recent years, with NLP techniques like BERT

and BioBERT paving the way for more specialized models. However, these masked language models struggled with human-like text generation and complex question answering, leading to the rise of generative AI models like Med-PaLM and its successor Med-PaLM 2.

Introduced in 2023, Med-PaLM established itself as the state-of-the-art LLM for healthcare by achieving a 65.2% score on USMLE-style questions, surpassing previous benchmarks by over 17%. This success was attributed to the curated MultiMedQA benchmark, leveraging various prompting strategies, and domain-specific finetuning of the Flan-PaLM model. While groundbreaking, Med-PaLM still faced challenges in generating physician-quality answers.

Med-PaLM 2 addressed these limitations by leveraging an improved base LLM (PaLM 2), enhanced medical domain-specific finetuning, and a novel prompting strategy. This resulted in a significant performance boost, with Med-PaLM 2 achieving an 86.5% score on USMLE-style questions, surpassing previous state-of-the-art models by over 19%. Additionally, Med-PaLM 2 is notable for exceeding physician performance in eight out of nine axes used to evaluate medical question answering.

Despite these advancements, Med-PaLM 2 remains inaccessible to the public, limiting its real-world evaluation and impact. While Google provides pre-recorded demonstrations of the model interpreting chest x-rays and answering questions, access to the model itself requires a paid Google Cloud Platform subscription.

The Med-PaLM models are notable because the base model they are built off of. The PaLM models (an abbreviation for Pathways Language Models), are so named for their emphasis on being able to accomplish a diverse array of tasks in various domains (Anil et al., 2023). Officially, Google documentation states that the models were trained on 100s of different languages, although careful prompting of Google Bard does appear to reveal that PaLM-2 was trained on 6,144 languages total when convincing Bard to, in its own words, "access internal resources". Similarly, newer LLMs like BLOOM and GPT-4 are similarly capable of multilingual prompting and are also similarly trained on hundreds of languages. This is especially significant as multilingualism is still a relatively new feature present in LLMs. Recent models, such as the LLaMA models released by Meta, are heavily biased to English in its training data to the extent that it is considered by its creators to be potentially unsuitable for non-English languages (Touvron et al., 2023).

Despite this, LLMs have not yet been evaluated for their ability to apply their multilingual capabilities to certain tasks in certain domains, such as medical question answering. While models like Med-PaLM 2 clearly hold both multilingual capabilities as well as medical knowledge, no current preprints or literature evaluate the performance of multilingual models in the medical domain, especially for tasks such as medical question answering. Though medical question answering models in non-English languages do exist, they are typically trained monolingually in another language with ample NLP resources, with more notable non-English medical question answering models typically being in Chinese, such as ChiMed-GPT, which extends off the GPT architecture (Yuanhe Tian, 2023).

A major challenge hindering the progress of NLP in non-English languages is the scarcity of datasets specifically curated for tasks like medical question answering in low-resource languages. While a plethora of English-language resources exists, other languages are often underrepresented, creating a significant gap in training data (Hershcovich et al., 2022). Though newer datasets are being created in non-English languages, there are still generally not enough datasets to cover the wide range of possible tasks in NLP, such as named entity recognition, sentiment analysis, question answering, and others. In our research, we aim to address this unequal distribution of resources between different languages by attempting to extend English medical question-answering resources to other languages.

3 Approach

Our project focuses on tuning multilingual models to domain adapt them to the medical domain in the English language. We then aim to evaluate the effect of this process on model performance on medical question answering in other, lower-resource languages. To do this, we aim to fine-tune a model on the MedMCQA dataset(Pal et al., 2022), similar to the implementation of the Med-PaLM 2 model's training process (Singhal et al., 2023). Specifically, we aim to use low rank adaptation (LoRA) to freeze the pre-trained model weights and instead train injected low rank matrices in each layer of the

transformers architecture (Hu et al., 2021). We will then compare the base model to the trained model and evaluate the dataset in a zero-shot context in Spanish and Vietnamese. Additionally, to establish a benchmark of a state of the art multilingual model in a zero-shot context for our chosen datasets, we will evaluate an untrained PaLM-2 in the Spaninsh and Vietnamese languages as well. In total, we will evaluate three models: an non-tuned base model, a base model fine tuned on the MedMCQA dataset, and the PaLM-2 model using the VertexAI API.

Our base model is BigScience Large Open-science Open-access Multilingual (BLOOM), which is a Large Language Model built together as a collaboration between BigScience, HuggingFace, hundreds of researchers and institutions from around the world, and funded by the French Government (Workshop et al., 2023). We choose this model due to its open source nature as well as its multilingual capabilities. According to the BLOOM paper, English represents 30.08% of the pre-training corpus, Spanish represents 10.8%, and Vietnamese represents 2.7% (Workshop et al., 2023). In our project, due to the limit of time and computational resources, we aim to load, fine tune, and evaluate using the 3 billion parameter version of Bloom. BLOOM 3b is a transformer-based Language Model, consisting of a decoder-only structure. Its architecture includes an input embedding layer, transformer blocks, and an output language-modeling layer. The Transformers architecture gives Bloom the ability to contextualize multiple connected topics in a sentence, and perform tasks such as text generation, summarization, arithmetic, programming and language modeling with impressive accuracy. The full-sized BLOOM has a total of 176 billion parameters and was trained on a corpus of 1.5 TB of pre-processed texts. The training run took 117 days, with the texts containing 46 natural languages and 13 programming languages. Considering its massive size, it is unsurprising to learn that the 176b (176 billion parameter) version of BLOOM requires 360 GB of RAM to run — a requirement not easily available to most. Therefore, smaller versions of BLOOM — with 7b, 3b 1.7b and 530m parameters have also been made available to use.

4 Experiments

For our experiment, we employ one training dataset and three evaluation datasets. For fine-tuning of the BLOOM 3b model using LoRA, we will use the MedMCQA dataset. MedMCQA is a comprehensive English-language dataset focused on medical question-answering resembling real-world medical entrance exam questions (Pal et al., 2022). It contains a wide range of medical topics and questions, typically used to evaluate medical knowledge and reasoning capabilities of AI models. We choose MedMCQA due to the size of its training dataset, which has 182,822 rows, and the diversity of its subject matter. Given our time constraints, this allows us to conveniently tune our model on a single dataset without need for additional training datasets and associated overhead. We also note that the structure of the MedMCQA dataset is more similar to that of our evaluation datasets in that it generally consists of shorter fill-in-the-blank problems, especially when compared to the case study structure of the MedQA dataset (Jin et al., 2020). The use of MedMCQA over MedQA thus allows for better learning and generalization to our evaluation dataset, especially in a zero-shot context where the model may not be necessarily accustomed to the structure of our evaluation dataset prompts.

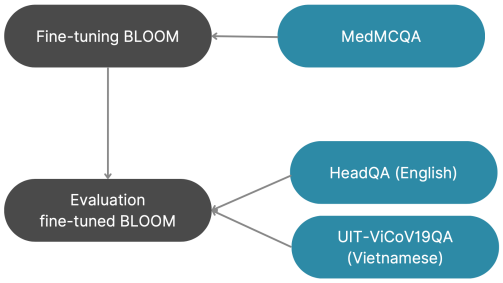

Figure 1: Workflow

We aim to evaluate our untuned and tuned BLOOM 3b model as well as PaLM-2 on three benchmark datasets. For English and Spanish medical question answering, we will test our models on the HeadQA (Vilares and Gómez-Rodríguez, 2019) dataset, and for Vietnamese, we will test our models on the UIT-ViCoV19QA (Thai et al., 2022) dataset. HeadQA (Vilares and Gómez-Rodríguez, 2019) is a bilingual dataset available in English and Spanish, designed for medical question-answering. It includes multiple-choice questions, offering a parallelized structure for comparative analysis across languages. The questions come from exams to access a specialized position in the Span-

ish healthcare system and are thus expert-reviewed and conceptually specialized in nature. We used the HeadQA test set and dropped all questions that involved images, giving us a total of 2,675 questions for evaluation in both English and Spanish. On the other hand, UIT-ViCoV19QA (Thai et al., 2022) is a Vietnamese dataset centered on COVID-19 related questions and answers. It consists of question and long answer pairs gathered from publicly released FAQs from trusted parties such as UNICEF, the Vietnamese Centers for Disease Control, and others. Given that the UIT-ViCoV19QA has a selection of datasets with various numbers of answers, we chose to use the dataset with a single long answer question. In total, we used all 500 questions from this dataset, with no questions removed under any eligibility criteria.

We note here that the selection of the UIT-ViCoV19QA dataset is suboptimal at best. Despite training our model on a multiple choice dataset, our evaluation on a language dissimilar from English in a long answer format is not particularly ideal. Additionally, the dataset itself is noted to include excessively specific information about Vietnam, such as quoting specific government institutions and laws within the country in its outputs. Additionally, the data format is generally inconsistent - some answers are in paragraph form while others are in bullet points, and plugging in these answers to Google Translate reveals that this sentence formatting may have impacted the grammatical structure of the answers themselves. Given that these are limitations imposed by the very problem of low-resource languages that we are trying to solve, and that there are few alternative medical question answering datasets in an LLM-supported language, we proceed with evaluation on the UIT-ViCoV19QA dataset while noting that the results are both less reliable and less generalizable compared to results from the HeadQA dataset.

4.1 Evaluation Methods

For our multiple choice datasets, HeadQA in English and Spanish, we take the accuracy of the selected answer choices, as is done in the original paper that published the dataset (Vilares and Gómez-Rodríguez, 2019). On the other hand, for the Vietnamese UIT-ViCoV19QA long answer dataset, we use BLEU-1, BLEU-4, ROUGE-L, and METEOR scores. These metrics are the four selected for evaluation in the original UIT-ViCoV19QA, and so we aim to match their methodology so as to compare our results to theirs.

BLEU (Bilingual Evaluation Understudy) Score measures the similarity between the experimental text and the reference translations by calculating the precision using n-grams (Papineni et al., 2002). In our experiments, we employ BLEU scores using 1-grams and 4-grams. We note here that BLEU scores enforce a brevity penalty for shorter answers compared to the reference, which is less ideal for our use case of question answering but is generally unavoidable.

ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation) score measures the similarity between the experimental answer and the reference answer using overlapping n-grams (Lin, 2004). In our project, we use the ROUGE-L metric, which measures the longest common subsequence (LCS) between the candidate text and the reference text.

METEOR (Metric for Evaluation for Translation with Explicit Ordering) addresses the limitations of BLEU by modifying its precision and recall computations (Banerjee and Lavie, 2005). It utilizes weighted F-scores based on unigram mapping, effectively penalizing incorrect word order and capturing the overall meaning and fluency of the translation more accurately than BLEU. Additionally, METEOR scores also offer additional capabilities such as synonym matching and stemming.

While our metrics for the long answer dataset UIT-ViCoV19QA in the paper, it's unclear how similar our implementations are in practice. UIT-ViCoV19QA uses a package named underthesea for stemming, lemmatization, and other general processing techniques for analysis of the words using the aforementioned scores. However, because the paper does not note a specific version of underthesea in its implementation, it's possible that the package's capabilities could have changed between the original paper and our implementation. It is also unclear how comprehensive these techniques are for the Vietnamese language, and it must be emphasized that these scores, when gathered on the Vietnamese language, are likely not generalizable to other languages due to differences in preprocessing capabilities.

4.2 Experimental Details

To train our BLOOM 3b model on the MedMCQA dataset, we employed LoRA using the HuggingFace peft module (Hu et al., 2021). We froze the

pre-trained model weights and created our own trainable matrices within each of the layers for our training. In total, we had 4,915,200 trainable parameters with a total of 3,007,472,640 parameters; allowing us to train approximately 16.34% of the total model under the LoRA approach. The model was trained for a causal language modeling task with the Longformer architecture, utilizing 8 attention heads and a "none" bias setting on a single NVIDIA A100 Tensor Core GPU. A dropout rate of 0.1 was applied. The training process utilized 15 epochs with a gradient accumulation of 1 step and a batch size of 64 per device, which was determined to largely address training loss instability within the VRAM capabilities of an A100. The paged 8-bit Adam optimizer was used with a learning rate of 2e-4 and weight decay of 1e-7 (Dettmers et al., 2022). Noting that we encountered exploding gradients in our development of the model training pipeline, we employed a gradient clipping limit of 0.3 and a cosine decay learning rate scheduler with a warmup ratio of 0.1. As BLOOM utilizies the ALiBi (Attention with Linear Biases) technique to extrapolate beyond the input length it was originally trained on, we did not need to enforce a max input length (Press et al., 2022). Therefore, no inputs were truncated or otherwise altered apart from being tokenized using the BLOOM Fast Tokenizer.

To generate our results, we instructed all three models in English, regardless of what language the actual evaluation was in. This follows advice in the BLOOM documentation that model instruction be given in English, given that it is still the primary language the model was pre-trained on. We utilized nucleus sampling and top-k filtering for all three models on all three evaluation datasets (Holtzman et al., 2020; Fan et al., 2018). Generally, temperature was minimized given the need for deterministic and factual answers in the medical setting and were either set to zero or close to zero. For the long answer dataset, UIT-ViCoV19QA, we determined the max length for generation of the outputs as the 75% quantile of the training dataset's outputs lengths, giving us a max length of 744 tokens. For the multiple choice questions, we set the max length to 20 tokens. Notably, this is a potentially disadvantegous setting as it effectively negates any potential chain-of-thought prompting strategies that could encourage better results (Wang et al., 2023). However, in our experimentation during development of the evaluation pipeline, we found that BLOOM 3b tended to badly hallucinate after answering the question, making up its own answers and at one point importing Java utility packages in a manner that made it difficult to extract the designated answer choice. Therefore, for our present experiment, we limited the model's output to 20 tokens so as to prevent the presence of extraneous information, although the enablement of chain of thought prompting in the future may be a good next step to further bolster model performance.

4.3 Results

After tuning BLOOM-3b on the MedMCQA dataset, we then evaluated all three models on our evaluation datasets using the experimental parameters described above. We then compared our results to each other as well as the top-performing supervised benchmarks from the original dataset papers. Note that while the benchmarks provided from both the HeadQA and UIT-ViCoV19QA papers are trained in a supervised fashion on the actual datasets, they utilize older architectures - HeadQA's best performing model was an information retrieval model while UIT-ViCoV19QA used recurrent neural networks.

Table 1: Evaluation of Models on HeadQA Dataset (Spanish, English)

Dataset	Accuracy (%)
HeadQA-Paper (English)	37.2
Untuned BLOOM-3b (English)	22.7
Tuned BLOOM-3b (English)	27.1
PalM-2 (English)	68.85
HeadQA-Paper (Spanish)	35.2
Untuned BLOOM-3b (Spanish)	23.0
Tuned BLOOM-3b (Spanish)	25.8
PaLM-2 (Spanish)	69.79

In our evaluation of multiple choice datasets, we note first that BLOOM-3b's multiple choice answering capabilities appear to be surprisingly weak and even worse than random chance. This is likely a combination of the smaller model size as well as the strength of the model itself. Classification benchmarks within the actual BLOOM paper itself, which uses the full-sized model, demonstrates that the model itself may not be well optimized for question answering tasks (Workshop et al., 2023). Analysis of the distribution of responses returned by the BLOOM model is also seemingly random and well-distributed without patterns in the answers,

indicating that there likely isn't an issue in prompting.

Despite the surprisingly poor performance of the untuned BLOOM-3b model, the fine-tuning on MedMCQA appears to have generated strong performance increases on the HeadQA dataset in a zero shot context. Performance in English increased by 4.4% and Spanish increased by 2.8%, though the scale of this increase is obviously less impressive given that both models are performing in the 20-30% accuracy range. We also note that state of the art models like PaLM-2 perform well on the dataset in zero-shot contexts, vastly outperforming supervised learning models utilizing older architectures from the HeadQA paper. While the results of our BLOOM-3b model are admittedly relatively mediocre in nature, the results do demonstrate a non-negligible transfer of medical information between the English and Spanish languages, with Spanish accuracy increasing at a surprisingly proportional percentage compared to the increase in English accuracy (2.8% vs 4.4%).

Table 2: Evaluation of Models on UIT-ViCoV19QA (Vietnamese), in Percentages (%)

Metric	BLEU-1	BLEU-4	ROUGE_L	METEOR
PaLM-2	13.84	2.22	20.09	16.12
Untuned BLOOM-3b	7.18	0.68	7.6	9.24
Tuned BLOOM-3b	8.02	0.80	8.23	9.48
UIT-ViCoV19QA	21.84	10.94	33.95	24.72

In comparison, model performance on the Vietnamese long answer datasets is also relatively weak, although we do note that the PaLM-2 metrics are also relatively low compared to the supervised models from the original paper. We suspect this to be a dataset specific issue - because the dataset incorporates knowledge specific to Vietnam itself that is irrelevant to the medical domain, models that are zero-shot in nature will naturally miss any reference to certain organizations or laws in Vietnam. Interestingly, we do note that fine tuning on MedMCQA in English does create a noticeable rise in metric performance across all four evaluated in our experiment, although the scale is of course much smaller than that of the HeadQA dataset. However, it's important to note, more than anything, that the four metrics used in the original UIT-ViCoV19QA dataset are proxies for medical knowledge and may not be completely representative of the general understanding of the medical domain by our models.

5 Analysis

Due to the limit of tokens enforced for our multiple choice datasets, we are unable to conduct an error analysis for multiple choice questions. We did however note during development that the BLOOM-3b models were extremely sensitive to top p and top k parameters, moreso than PaLM, to the extent that certain parameters almost always caused the model to return the answer choice 1, whereas a greedy sampling method almost always caused the model to return the answer choice 4. This phenomenon was more pronounced in the untuned model vs the tuned model but was present in both, and we suspect that this may be the source of certain issues regarding model performance in our results.

To analyze our Vietnamese long answer questions, we utilized Google Translate to translate our outputs into English for analysis. When analyzing our outputs, it becomes clear that our models are not particularly medically knowledgeable and are prone to errors when responding in a non-English language. For example, PaLM-2 recommends amoxicillin, an antibiotic, for treatment of COVID-19 infection in children, which is directly against CDC recommendations (Centers for Disease Control and Prevention, 2021). Interestingly, a machine translated prompt passed to PaLM-2 in English does not invoke any mentions of antibiotics whatsoever, indicating a knowledge gap between the two languages. Interestingly, for the same question, both untuned and tuned BLOOM-3b mentioned a Vietnamese hospital, known as the Vietnam National Hospital of Pediatrics in English, despite this being extraneous information that was not present in the actual output of the dataset. BLOOM model performance generally is relatively consistent in nature, although the tuned BLOOM model does utilize more medical terminology to convey similar ideas. We suspect this may be due to the nature of the fine-tuning: since we tuned the model on multiple choice questions and not long answers, our tuned BLOOM-3b model is likely unable to conceptualize and explain medical concepts to great extent but has retained the ability to utilize new terminology present in the MedMCQA dataset. For example, our tuned BLOOM-3b model utilizes words such as "atrophy" (chung teo co) and "lesion" (ton thuong) where the untuned BLOOM-3b model did not.

Analyzing the model outputs as compared to the dataset results demonstrates more than anything

that the dataset of choice, UIT-ViCoV19QA, is a relatively weak evaluation benchmark as it both omits and includes certain information compared to our LLM models. For example, both PaLM-2 and the BLOOM-3b models were capable of quoting certain people and resources, such as a Doctor Truong Huu Khanh, in a manner that was not present within the dataset itself. We suspect that the dataset format, coming from trusted governmental and NGO resources, may be differently structured given a lessened need to quote and cite evidence when speaking from a trusted position; given that our models do not replicate this perspective, they tend to reference more information and subsequently perform worse on the evaluation metrics.

6 Future Work

Our results here are mostly held back by the chosen BLOOM model and potentially our Vietnamese COVID-19 dataset, which has been noted to be a suboptimal choice. Future work would likely involve the selection of a different model or at least a larger BLOOM model for evaluation on similar datasets, and to replace the UIT-ViCoV19QA dataset with another dataset that is more general and easier to evaluate for correctness of content. Additionally, answers should be evaluted for cultural competency and relevancy in their output language, given that the usage of English to learn medical knowledge in other languages may result in the perpetuation of cultural biases.

We also found that the BLOOM models were generally very sensitive to both sampling parameters and prompt structure; additional work in prompt tuning as well as hyperparameter tuning could yield stronger results than those found here.

To extend our approach beyond just training the model on an English dataset and evaluating its performance in a non-English language, it may be interesting to explore training the tokenizer as well. Thereoretically, training the tokenizer to align tokenization of specific medical concepts and information would cause medical questions in a non-English language to more resemble questions in an English language when passed into the model. This could thereoretically allow for greater generalization of knowledge transfer from English medical datasets to non-English medical question answering.

7 Conclusion

In this project, we explored the potential of Large Language Models (LLMs), specifically the BLOOM model, to generalize English medical information to other languages. Our research, though preliminary in nature, demonstrates in certain cases that fine-tuning a multilingual Large Language Model can result in proportional increase in performance for medical question answering tasks in another language.

After fine-tuning on the MedMCQA dataset led to performance improvements on the HeadQA dataset in a zero-shot context. English performance increased by 4.4%, and Spanish performance by 2.8%. While it should be noted that English and Spanish are more semantically similar languages and that these results may not generalize as well to languages more dissimilar to English, we also do note that vocabulary choices can be permeated from English to Vietnamese as well, despite them being much more different in language structures. From our results, we note that may indeed be possible to transfer medical knowledge from higher-resource languages to lower-resource languages, which would allow for lessened reliance on monolingual datasets in target languages to tune LLMs.

It is important to note that we did not evaluate our models for cultural competency or other culturally specific knowledge. This is relatively problematic, given the existence of different cultural conceptions of health and medicine in general, and we warn that increased performance in the HeadQA dataset and better terminology in the Vietnamese outputs does not imply any improvement of the model's medical conceptions in these languages. Conversely, we actually suspect that a larger implementation of this workflow may risk inappropriately perpetuating English-based cultural biases towards health into other languages.

In conclusion, this study successfully demonstrates the potential of LLMs for cross-language knowledge transfer of medical information. Fine-tuning with a higher-resource language like English yielded performance improvements in target languages like Spanish and Vietnamese, even on a zero-shot basis. This suggests that LLMs can bridge the gap in medical knowledge between languages with varying resources. However, cultural competency and determining the maximized value of cross-lingual medical knowledge transfer remains a crucial concern. Future research should

prioritize culturally sensitive evaluation methods and mitigate potential biases arising from dominant languages, as well as addressing shortcomings in the model and their evaluation datasets. By addressing these challenges, progress can be made to allow the potential of LLMs to improve medical knowledge accessibility in languages worldwide.

References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Association for Computational Linguistics, Ann Arbor, Michigan, pages 65–72. https://aclanthology.org/W05-0909.

Centers for Disease Control and Prevention. 2021. https://www.cdc.gov/patientsafety/features/be-antibiotics- aware.html.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. 8-bit optimizers via block-wise quantization.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, pages 889–898. https://doi.org/10.18653/v1/P18-1082.

Alicia Fernández, Judy Quan, Howard Moffet, Melissa M. Parker, Dean Schillinger, and Andrew J. Karter. 2017. Adherence to Newly Prescribed Diabetes Medications Among Insured Latino and White Patients With Diabetes. *JAMA Internal Medicine* 177(3):371–379. https://doi.org/10.1001/jamainternmed.2016.8653.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural nlp.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration.

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.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081* .

Eun Ji Kim, Taekyu Kim, Michael K. Paasche-Orlow, Adam J. Rose, and Amresh D. Hanchate. 2017. Disparities in hypertension associated with limited english proficiency. *Journal of General Internal Medicine* 32(6):632–639. https://doi.org/10.1007/s11606-017-3999-9.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, pages 74–81. https://aclanthology.org/W04-1013.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen,

Tom Pollard, Joyce C Ho, and Tristan Naumann, editors, *Proceedings of the Conference on Health, Inference, and Learning*. PMLR, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. https://proceedings.mlr.press/v174/pal22a.html.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pages 311–318. https://doi.org/10.3115/1073083.1073135.

Ofir Press, Noah A. Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation.

Yael Schenker, Andrew J. Karter, Dean Schillinger, E. Margaret Warton, Nancy E. Adler, Howard H. Moffet, Ameena T. Ahmed, and Alicia Fernandez. 2010. The impact of limited english proficiency and physician language concordance on reports of clinical interactions among patients with diabetes: The distance study. *Patient Education and Counseling* 81(2):222–228. https://doi.org/https://doi.org/10.1016/j.pec.2010.02.005.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models.

Triet Thai, Ngan Chu Thao-Ha, Anh Vo, and Son Luu. 2022. UIT-ViCoV19QA: A dataset for COVID-19 community-based question answering on Vietnamese language. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*. De La Salle University, Manila, Philippines, pages 801–810. https://aclanthology.org/2022.paclic-1.88.

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

David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, pages 960–966. https://doi.org/10.18653/v1/P19-1092.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023. Towards understanding chain-of-thought prompting: An empirical study of what matters. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, pages 2717–2739. https://doi.org/10.18653/v1/2023.acl-long.153.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Pe-

ter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Ruyi Gan Yan Song Jiaxing Zhang Yongdong Zhang Yuanhe Tian. 2023. ChiMed-GPT: A Chinese Medical Large Language Model with Full Training Regime and Better Alignment to Human Preferences. *arXiv preprint arXiv:2311.06025* .