Domain-Specific Multilingual Strategies for Medical NLP: A Cross-Lingual Analysis of Orthographic and Phonemic Representations

Kyungjin Kim[1,5,6,7], Jinju Kim[2,5], Haeji Jung[3,5], David R. Mortensen[4] and Jongmo Seo*[1,6,7,8]

***Abstract*— Medical natural language processing (NLP) has greatly benefited from the increasing availability of electronic health records (EHRs) across multiple languages. However, most available resources are heavily biased toward English, restricting the development of non-English medical NLP models and limiting their broader adoption in healthcare AI.**

In general domain NLP, monolingual models trained on high-resource European languages often outperform multilingual models. However, this paper takes a domain-specific perspective, highlighting the unique linguistic characteristics of medical terminology. Using a diverse corpus of medical texts in English, Italian, Spanish, and French, we pretrain RoBERTa-based models in both monolingual and multilingual settings. Contrary to conventional NLP trends, our results suggest that multilingual training enhances cross-lingual knowledge transfer and can even outperform monolingual models in medical domains. Through a comparative analysis of orthographic and phonemic representations, we find evidence suggesting that the strong Latin roots of medical terminology may facilitate effective knowledge sharing among languages with similar scripts.

These findings indicate that orthographically focused multilingual strategies could provide a more robust framework for modeling specialized medical terminology. In particular, multilingual training that capitalizes on script similarities appears to enable richer information utilization, which may contribute to improved medical NLP performance across languages.

***Clinical relevance*— Advancing multilingual medical language models can enhance clinical decision support and expand access to critical healthcare information across linguistically diverse communities.**

[1]Kyungjin Kim and Jongmo Seo are with the Department of Electrical and Computer Engineering, Seoul National University, 08826 Seoul, Republic of Korea. {kkj_james, callme}@snu.ac.kr

[2]Jinju Kim is with Department of Electrical and Computer Engineering, Sungkyunkwan University, 16419 Suwon, Republic of Korea. perla0328@g.skku.edu

[3]Haeji Jung is with Korea University, 02841 Seoul, Republic of Korea. gpwl0709@korea.ac.kr

[4]David R. Mortensen is with the Language Technologies Institute, School of Computer Science, Carnegie Mellon University, 15213 Pittsburgh, PA, USA. dmortens@cs.cmu.edu

[5]This research was conducted while Kyungjin Kim, Jinju Kim, Haeji Jung were visiting researcher at Carnegie Mellon University, 15213 Pittsburgh, PA, USA.

[6]Kyungjin Kim and Jongmo Seo are also affiliated with the Institute of Engineering Research and the Inter-university Semiconductor Research Center (ISRC), Seoul National University, 08826 Seoul, Republic of Korea.

[7]Kyungjin Kim and Jongmo Seo are also affiliated with the Biomedical Research Institute, Seoul National University Hospital, 03080 Seoul, Republic of Korea.

[8]Jongmo Seo is also affiliated with the Interdisciplinary Program of Medical Informatics, Seoul National University Hospital, 03080 Seoul, Republic of Korea.

I. INTRODUCTION

Electronic health records (EHR), including patient notes, radiology reports, and clinical conversation transcripts, are being collected in ever increasing volumes. These resources, along with the emergence of publicly available medical datasets (e.g., PubMedQA [1]), provide a rich foundation for the development of advanced language processing tools in the medical domain. This abundance of data has fueled significant progress in medical natural language processing (NLP), allowing a more comprehensive automated analysis of clinical documentation and physician diagnostic notes [2]. Various research efforts have already produced specialized language and speech models to support clinical decision-making, such as Med-Flamingo (a vision–language model for healthcare) [3] and automated medical report generation systems.

Despite these advances, a key challenge persists: medical text and speech data are overwhelmingly in English, creating a major barrier for multilingual medical NLP research [4]. Although multilingual data sets have recently become available, they still reflect an existing language imbalance [1]. This imbalance poses challenges for models trained solely on languages other than English, which often suffer from insufficient data. Although dataset imbalance is common in NLP, it is particularly problematic in the medical domain due to disparities in access to technology and digitized records across regions. As a result, multilingual models capable of effective cross-lingual knowledge transfer remain underexplored in medical NLP [5]. Addressing this gap is essential to expand access to AI tools for healthcare in linguistic communities.

In standard NLP tasks, researchers often focus on improving monolingual strategies, as multilingual models can under-perform in high-resource settings [6]. For example, a bilingual English-French model can be less effective than two separate monolingual models. To reduce this gap, some studies introduce phonemic representations that align words across languages by sound, helping highlight latent similarities [7], [8].

However, the medical domain presents distinct linguistic characteristics. Medical terms in many languages are rooted in Latin or Greek, resulting in shared spellings across languages [9]. Phonemic transcription may obscure these orthographic similarities, as pronunciation varies across languages even when spelling does not. We hypothesize that preserving orthographic forms (e.g., original spelling) is more advantageous for cross-lingual learning in this domain.

By maintaining Latin-derived spellings, models can more easily recognize and transfer knowledge of medical concepts across languages. Our approach leverages a multilingual model using English as a high-resource bridge language and emphasizes orthographic consistency. This strategy leads to improved performance across all metrics evaluated on PubMedQA [1].

This work makes three main contributions:

- We develop and evaluate a multilingual model that uses English as a mediating language to enhance performance in European languages. This approach suggests that a multilingual training strategy with cross-lingual transfer can effectively support medical NLP tasks in various linguistic settings.
- We provide an in-depth examination of linguistic features in the medical domain, focusing on the prevalence of Roman scripts and Latin (or Greek) roots. This analysis provides insights into why certain cross-lingual strategies, particularly those leveraging orthographic similarity, may be especially effective in healthcare contexts.
- We offer a comparative study of orthographic versus phonemic representations in multilingual medical language models. Our findings indicate that the preservation of orthographic information (e.g. original written text) tends to yield better cross-lingual performance over phonemic transcription in essential medical NLP tasks, providing methodological insights for future speech model development in healthcare.

II. RELATED WORKS

A. Medical Language Models

Recent advances in Large Language Models (LLMs) specialized for the medical domain have demonstrated remarkable performance on tasks such as clinical text classification, entity extraction, and medical question answering. Early examples include BioBERT [10] and PubMedBERT [11], which adapted BERT [12] to biomedical corpora, and more recently BioGPT [13] and BioMistral [14], which leveraged generative pre-training on large-scale biomedical datasets. Although these domain-focused models achieve state-of-the-art results on English medical QA benchmarks (e.g., PubMedQA), performance often declines for non-English medical data. Multilingual or cross-lingual evaluation in medical QA remains underexplored, underscoring the need for multilingual approaches that effectively capture clinical knowledge across different linguistic settings.

B. Phonemic Representations

As shown in Fig. 1, phonemic transcription captures pronunciation-based similarities across languages with divergent spellings. Transformer-based multilingual LLMs such as mBERT [12] and XLM-R [15] can generalize across languages but often struggle in low-resource or novel script scenarios [16]. To address these challenges, various approaches normalize orthographic differences using transliteration or subword modeling. For instance, [8] employed IPA-based phonemic representations to extend named entity recognition (NER) to unseen languages in a zero-shot setting. Other work has used morphological or phonological subword units for embedding adaptation [17], or phonetic representations in multimodal training [18]. These efforts suggest that phoneme-level input may enhance model robustness in low-resource scenarios, though its effectiveness varies depending on linguistic and domain-specific factors.

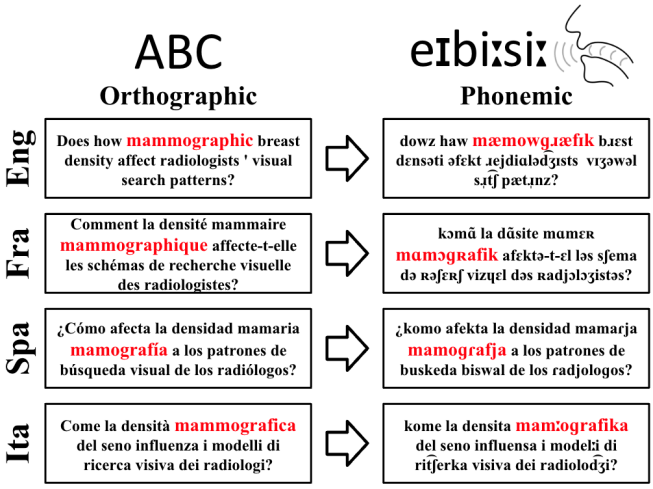

Fig. 1. **Comparison of Orthographic and Phonemic Representations Across Languages** This figure illustrates how phonemic transcription provides a distinct representation compared to orthographic text. The example illustrates how phonemic transcription differs from orthographic text in multilingual settings, while still preserving key medical terminology such as 'mammographic' across languages.

III. METHODS

A. Models and Data Preprocessing

For pretraining, we used the HitZ Multilingual Medical Corpus from Hugging Face, which comprises medical texts in Italian, Spanish, and French [19]. The corpus is divided into four subsets: separate monolingual corpora for Italian, Spanish, and French (each containing 500K samples) and a combined multilingual corpus totaling 1.5M samples.

To explore phoneme-based representations, we generated International Phonetic Alphabet (IPA) versions of each corpus using the Epitran library [20], which provides rule-based grapheme-to-phoneme conversion for a wide range of languages. Inputs were lowercased and symbols were cleaned to ensure consistent phoneme transcriptions with minimal noise. As illustrated in Fig. 1, phonemic transcription captures pronunciation-based similarities that orthographic representations may overlook, particularly in multilingual contexts.

For tokenization, we trained separate SentencePiece BPE tokenizers [21] for each of the 22 datasets (11 orthographic and 11 phonemic). This tokenizer setup allowed for subword segmentation optimized for each language and script type.

To evaluate the impact of orthographic and phonemic representations, we designed 11 configurations of monolingual and multilingual corpora, as summarized in Table I. For fine-tuning and evaluation, we used the pqa_artificial version of

the PubMedQA dataset, which contains English biomedical question-answering examples. Only samples labeled as "yes" or "no" were retained. The dataset was balanced by randomly sampling 13,500 examples per class for training (27,000 total) and 1,500 examples per class for testing (3,000 total). No separate validation split was used.

TABLE I
COMPLETE LANGUAGE REPERTOIRE COMBINATIONS

Language Repertoire	Eng	Fra	Ita	Spa
Monolingual (Fra)		✓		
Monolingual (Ita)			✓	
Monolingual (Spa)				✓
Bilingual (Eng-Fra)	✓	✓		
Bilingual (Eng-Ita)	✓		✓	
Bilingual (Eng-Spa)	✓			✓
Bilingual (Fra-Ita)		✓	✓	
Bilingual (Fra-Spa)		✓		✓
Bilingual (Ita-Spa)			✓	✓
Multilingual (Fra-Ita-Spa)		✓	✓	✓
Multilingual (Eng-Fra-Spa-Ita)	✓	✓	✓	✓

Each configuration was paired with its corresponding IPA-transformed variant, resulting in a total of 22 datasets (11 orthographic and 11 phonemic) for pretraining.

Before training, all datasets underwent standardized preprocessing steps:

- **Tokenization** We trained SentencePiece BPE tokenizers for each corpus variant to ensure optimal subword segmentation adapted to the respective language and representation type (orthographic or phonemic) [21], [22].
- **Masked Language Modeling (MLM)** Fill-mask corpora are prepared by masking tokens in the input sequences following established protocols for the MLM objective.

For fine-tuning, we used a translated version of the PubMedQA dataset[1], originally in English. The dataset, which includes context, question, and final decision (yes/no), is translated into Italian, Spanish, and French for the context and question fields using MarianMT translation models [23]:

- **French:** Helsinki-NLP/opus-mt-en-fr
- **Italian:** Helsinki-NLP/opus-mt-en-it
- **Spanish:** Helsinki-NLP/opus-mt-en-es

B. Model Analysis

To understand how models processed linguistic information, we examined internal representations and token-level interactions, focusing on the effects of language combinations and script types (orthographic vs. phonemic).

*1) **Word-wise Attention Visualization**:* To better understand the model's decision-making process, we analyzed word-wise attention distributions during inference. These distributions were visualized as heat maps overlayed on the input text, highlighting how different models assigned attention to each token within the given question and context.

By comparing these heatmaps between models trained in different datasets, we identified patterns in token-level attention and examined how linguistic and script variations influenced the model's internal mechanisms. This analysis complemented the CKA-based findings by offering a more fine-grained view of representational differences at the word level.

For visualization, we extracted the self-attention weights from the final transformer layer and averaged across all attention heads. This provided a unified view of the model's token-level focus throughout the input sequence.

*2) **CKA-Based Cross-lingual Similarity Calculation**:* To better understand how multilingual models encode linguistic information, we used the Centered Kernel Alignment (CKA) method to measure the similarity between the hidden states of different models. CKA quantifies how closely two model representations align, allowing us to compare the internal feature spaces learned by models trained with different language combinations and script types (orthographic versus phonemic). We computed the CKA scores as follows:

- **Hidden State Extraction:** For each model, we extracted hidden states by processing the entire test set from PubMedQA, consisting of 3,000 examples. The input sequences were constructed by concatenating the question and context fields. We then mean-pooled the final hidden states to create fixed-size vector representations, providing a concise summary of each model's learned encoding of the combined input.
- **CKA Score Calculation:** Using a custom CUDA-enabled CKA implementation, we computed pairwise CKA scores between the extracted hidden states of models trained on different datasets. These scores provided a direct measure of representational similarity.
- **Visualization:** We presented the computed CKA scores as triangular heat maps, using different color schemes to differentiate orthographic and phonemic models.

This analysis offered insights into how different language combinations and script types (orthographic and phonemic) influenced the models' ability to capture cross-lingual similarity. The similarity matrices presented in Fig. 3 are computed using the entire test set (3,000 samples), ensuring that they reflect overall representational alignment rather than single-instance behavior.

IV. RESULTS

Table II presents the evaluation results that compare monolingual and multilingual models across different metrics.

A. Pretraining Setup

Table III summarizes the pretraining hyperparameters and model architecture used in our experiments for reproducibility. The model follows a RoBERTa-based masked language modeling (MLM) objective, utilizing different multilingual and phonemic corpora configurations, as outlined in Table I. The vocabulary size is dynamically adjusted on the basis of the number of languages included in the training data, ensuring efficient tokenization across different script types.

TABLE II

PERFORMANCE COMPARISON ACROSS DIFFERENT MODELS

Languages	Accuracy		F1		Precision		Recall	
	Ortho	Phoneme	Ortho	Phoneme	Ortho	Phoneme	Ortho	Phoneme
Test on French Dataset								
Fra	0.743	0.765	0.740	0.760	0.750	0.777	0.730	0.745
Fra-Spa	0.762	0.771	0.766	0.755	0.754	**0.813**	0.779	0.705
Fra-Ita	0.770	0.765	0.774	0.770	0.760	0.753	0.788	0.789
Eng-Fra	**0.783**	**0.780**	**0.788**	**0.784**	0.770	0.771	**0.808**	**0.797**
Fra-Spa-Ita	0.757	0.753	0.758	0.756	0.755	0.746	0.760	0.766
Eng-Fra-Spa-Ita	0.771	0.774	0.758	0.760	**0.804**	0.807	0.717	0.719
Test on Spanish Dataset								
Spa	0.747	0.751	0.750	0.756	0.741	0.740	0.759	0.773
Fra-Spa	0.782	0.784	0.781	0.771	0.782	**0.821**	0.781	0.727
Spa-Ita	0.791	0.776	0.777	0.773	**0.830**	0.786	0.731	0.760
Eng-Spa	**0.799**	0.790	**0.801**	**0.787**	0.791	0.800	**0.813**	0.775
Fra-Spa-Ita	0.767	0.783	0.764	0.786	0.775	0.776	0.753	**0.795**
Eng-Fra-Spa-Ita	0.789	**0.792**	0.781	0.786	0.814	0.808	0.751	0.766
Test on Italian Dataset								
Ita	0.766	0.744	0.763	0.744	0.772	0.744	0.754	0.745
Fra-Ita	0.777	0.762	0.782	0.770	0.764	0.745	0.801	0.796
Spa-Ita	0.779	0.780	0.771	0.780	0.801	0.781	0.743	0.778
Eng-Ita	0.789	0.778	**0.792**	0.777	0.780	0.781	**0.805**	0.772
Fra-Spa-Ita	0.765	0.778	0.754	0.783	0.790	0.765	0.720	**0.802**
Eng-Fra-Spa-Ita	**0.795**	**0.795**	0.781	**0.788**	**0.840**	**0.813**	0.729	0.765

Bold values indicate the highest score among all models for each metric (Accuracy, F1, Precision, and Recall) in both orthographic and phonemic evaluations, depicted as Ortho and Phoneme, respectively. Red highlights denote the overall best performance.

TABLE III

TRAINING HYPERPARAMETERS

Pretraining			
Number of Layers	6	Max Steps	100k
Hidden Size	768	LR Decay	linear
FFN Inner Size	3072	Learning Rate	2.5e-4
Attention Heads	12	Warmup Steps	5k
MLM Probability	0.15	Weight Decay	0.01
Fine Tuning			
Epochs	10	LR Decay	linear
Metric	Accuracy	Learning Rate	2e-5
Warmup Steps	500	Weight Decay	0.01

B. Model Performance

*1) **Multilingual Models vs. Monolingual Models**:* Multilingual models consistently outperformed their monolingual counterparts, especially in configurations that included English. For example, in French evaluations, the monolingual Fra model achieved an accuracy of 0.743, whereas the Eng-Fra setup improved to 0.783. A similar trend appeared in Spanish (0.799 vs. 0.747) and Italian tasks (0.795 vs. 0.766). These results contrast with general-domain NLP, where monolingual training often prevails, and underscore English's value as a bridge language for medical QA.

*2) **Effect of IPA Conversion**:* Phonemic (IPA) transformations exhibited limited and inconsistent benefits in the medical domain. While phonemic variants occasionally outperformed orthographic ones (e.g., Fra-Spa on Spanish), such improvements were modest and irregular. Most IPA-based models failed to surpass orthographic counterparts, likely due to the high orthographic consistency of Latin-derived medical terms. Thus, although phonemic alignment aids general-domain cross-lingual transfer [7], [8], its impact appears diminished in this domain where spelling-based similarity dominates.

C. Model Analysis

*1) **Word-wise Attention Visualization**:* Fig. 2 visualizes the attention distributions for three models, illustrating differences in how they align the question and context tokens.

Ita Monolingual: The monolingual model (Fig. 2-(a)) exhibits scattered attention, failing to focus on key information in the context, and prioritizing less relevant terms in the question. This misalignment suggests that the Ita model struggled to establish meaningful connections between the question and the context.

Fra-Ita-Spa Multilingual: The multilingual model trained on French, Italian, and Spanish (Fig. 2-(b)) demonstrates better differentiation between the question and the context compared to the Ita monolingual model. However, it fails to consistently identify the critical terms needed for an accurate understanding. Similarly to Ita, its attention is diffused, resulting in suboptimal performance and ineffective attention generation.

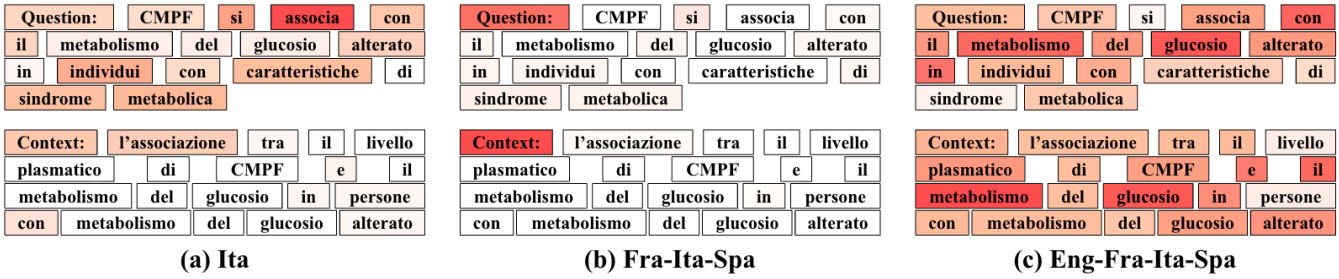

Fig. 2. **Visualization of Attention Matrices of Monolingual and Multilingual Models.** (a) Ita monolingual model, (b) Fra-Ita-Spa multilingual model, and (c) Eng-Fra-Spa-Ita multilingual model. This visualizes where the network "looks" while trying to extract the answer from the context. Key alignments and token focus differences are highlighted.

Eng-Fra-Spa-Ita Multilingual: The Eng-Fra-Spa-Ita multilingual model (Fig. 2-(c)) consistently attends to key medical terms such as metabolismo, glucosio, and CMPF, demonstrating a stronger alignment between the question and context, which likely contributes to its superior classification accuracy.

*2) **Cross-lingual Similarities of Orthographic and Phonemic Representations**:* While attention heatmaps reveal how models focus on important tokens, a broader view of cross-lingual representation similarity is shown in Fig. 3.

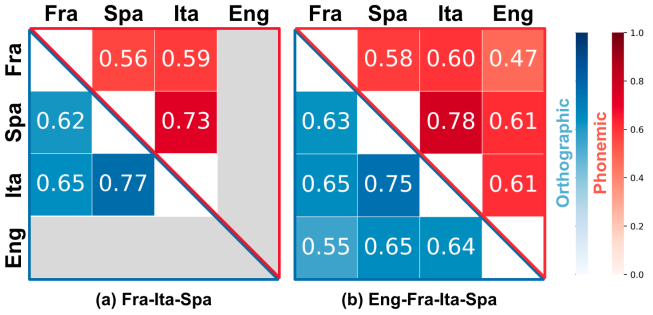

Fig. 3. **Cross-lingual Representation Similarity in Multilingual Models.** (a) Fra-Ita-Spa multilingual model excluding English. (b) Eng-Fra-Spa-Ita multilingual model including English. The lower triangle (blue) represents similarities in orthographic models, while the upper triangle (red) shows similarities in phonemic models.

Representation Similarity Including English in the model (Eng-Fra-Spa-Ita) increased the overall ability to capture representation similarities across languages. For example, the phonemic similarity between Italian and Spanish (Ita-Spa) increased from 0.73 (Fra-Spa-Ita) to 0.78, surpassing their orthographic similarity of 0.75. This suggests that incorporating English contributed to stronger cross-lingual representation alignment in both orthographic and phonemic spaces.

Classification Performance Comparison Despite the gains in phonemic similarity, orthographic representations continued to produce higher or comparable classification performance. On the Italian dataset, Eng-Fra-Spa-Ita achieved identical accuracy (0.795) for orthographic and phonemic settings. On the French dataset, phonemic accuracy showed a slight improvement (0.774 vs. 0.771), but this trend was not consistent across other datasets. These results indicate that, while phonemic representations benefit from language diversity, their utility in classification tasks within the medical domain remains limited.

V. DISCUSSION

Our experiments highlight how linguistic diversity, phonemic representations, and orthographic consistency influence multilingual medical NLP. Specifically, we show that while phonemic alignment improves cross-lingual representation similarity, orthographic consistency remains crucial for classification accuracy.

First, incorporating English improved the model's ability to capture cross-lingual similarities, particularly in the phonemic space. The phonemic similarity between Italian and Spanish increased from 0.73 (Fra-Spa-Ita) to 0.78 (Eng-Fra-Spa-Ita), surpassing their orthographic similarity of 0.75. This suggests that the inclusion of English, with its typological contrast, enhances phonemic representations, enabling the model to better capture linguistic relationships across multiple languages.

However, despite this representational gain, the classification performance remained stronger with orthographic representations. As seen in the results, orthographic models consistently outperformed or matched phonemic models in medical QA tasks. This is likely due to the inherent structure of medical terminology, which relies on orthographic consistency across languages due to shared Latin roots. In contrast, phonemic transformations introduce pronunciation variations that obscure these similarities, reducing their effectiveness for domain-specific tasks.

These findings highlight that the effectiveness of multilingual strategies is highly domain-dependent, particularly in medical NLP. Although phonemic representations are often valuable for bridging typological gaps in general NLP, medical NLP benefits more from preserving orthographic forms. Future research could explore hybrid approaches that selectively apply phonemic encoding to distant languages with different scripts (e.g., Roman and Brahmic) while preserving orthographic consistency for languages with shared etymological roots. Additionally, integrating script-aware tokenization methods could further refine cross-lingual knowledge transfer. Expanding this framework to include non-Latin scripts such as Devanagari, Arabic, or Cyrillic

may offer additional insight into the role of phonemic representations when orthographic overlap is minimal.

ACKNOWLEDGMENT

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant, funded by the Korea government (MSIT) (RS-2022-00143911, AI Excellence Global Innovative Leader Education Program).

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant, funded by the Korea government (MSIT) (IITP-2024-00441407, Global Data X Leader HRD Program).

REFERENCES

[1] Q. Jin, B. Dhingra, Z. Liu, W. Cohen, and X. Lu, "PubMedQA: A dataset for biomedical research question answering," Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 2567–2577, 2019.

[2] Singhal, Karan and Azizi, Shekoofeh and Tu, Tao and Mahdavi, S Sara and Wei, Jason and Chung, Hyung Won and Scales, Nathan and Tanwani, Ajay and Cole-Lewis, Heather and Pfohl, Stephen, "Large language models encode clinical knowledge," Nature, vol. 620, no. 7972, pp. 172–180, 2023.

[3] Moor, Michael and Huang, Qian and Wu, Shirley and Yasunaga, Michihiro and Dalmia, Yash and Leskovec, "Med-Flamingo: A multimodal medical few-shot learner," in Proc. Mach. Learn. Health (ML4H), 2023, pp. 353–367.

[4] Wu, Chaoyi and Lin, Weixiong and Zhang, Xiaoman and Zhang, Ya and Xie, Weidi and Wang, Yanfeng, "PMC-LLaMA: Toward building open-source language models for medicine," J. Amer. Med. Inform. Assoc., 2024, p. ocae045.

[5] Qiu, Pengcheng and Wu, Chaoyi and Zhang, Xiaoman and Lin, Weixiong and Wang, Haicheng and Zhang, Ya and Wang, Yanfeng and Xie, Weidi, "Towards building multilingual language model for medicine," Nat. Commun., vol. 15, no. 1, p. 8384, 2024.

[6] Chang, Tyler A and Arnett, Catherine and Tu, Zhuowen and Bergen, and Benjamin K, "When is multilinguality a curse? Language modeling for 250 high- and low-resource languages," arXiv preprint arXiv:2311.09205, 2023.

[7] B. Muller, A. Anastasopoulos, B. Sagot, and D. Seddah, "When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models," Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 448–462, 2021.

[8] J. Sohn, H. Jung, A. Cheng, J. Kang, Y. Du, and D. R. Mortensen, "Zero-shot cross-lingual NER using phonemic representations for low-resource languages," *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 13595–13602, 2024.

[9] S. S. Krishna and M. Hans, "Understanding medical free text: A terminology driven approach," in Proc. 5th Int. Workshop Comput. Terminol. (Computerm2016), Osaka, Japan, Dec. 2016, pp. 121–125. The COLING 2016 Organizing Committee.

[10] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234–1240, 2020.

[11] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," ACM Transactions on Computing for Healthcare, vol. 3, no. 1, pp. 1–23, 2021.

[12] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186, 2019.

[13] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu, "BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining" Briefings in Bioinformatics, vol. 23, no. 6, 2022.

[14] Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier, and R. Dufour, "BioMistral: A collection of open-source pretrained large language models for medical domains," Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), 2024.

[15] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 8440–8451, 2020.

[16] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual BERT?" Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 4996–5001, 2019.

[17] A. Chaudhary, C. Zhou, L. Levin, G. Neubig, D. R. Mortensen, and J. Carbonell, "Adapting word embeddings to new languages with morphological and phonological subword representations," Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 3285–3295, 2018.

[18] C. Leong and D. Whitenack, "Phone-ing it in: Towards flexible multi-modal language model training by phonetic representations of data," Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 5306–5315, 2022.

[19] I. García-Ferrero, R. Agerri, A. Atutxa Salazar, E. Cabrio, I. de la Iglesia, A. Lavelli, B. Magnini, B. Molinet, J. Ramirez-Romero, G. Rigau, J. M. Villa-Gonzalez, S. Villata, and A. Zaninello, "Medical mT5: An open-source multilingual text-to-text LLM for the medical domain," Proceedings of LREC-COLING 2024, 2024.

[20] D. R. Mortensen, S. Dalmia, and P. Littell, "Epitran: Precision G2P for many languages," Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[21] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 66–71, 2018.

[22] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 1715–1725, 2016.

[23] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. F. Aji, N. Bogoychev, A. F. T. Martins, A. Birch, "Marian: Fast neural machine translation in C++," Proceedings of ACL 2018, System Demonstrations, pp. 116–121, 2018.