MedPodGPT: A multilingual audio-augmented large language model for medical research and education

Shuyue Jia[1,*], Subhrangshu Bit[2,*], Edward Searls[3,*], Lindsey A. Claus[3,4,†], Pengrui Fan[2,†], Varuna H. Jasodanand[3,†], Meagan V. Lauber[3,†], Divya Veerapaneni[3,5,†], William M. Wang[2,†], Rhoda Au[3,6,7,8,9,10] & Vijaya B. Kolachalama[2,3,11,‡]

[1]*Department of Electrical & Computer Engineering, Boston University, MA, USA*
[2]*Department of Computer Science, Boston University, MA, USA*
[3]*Department of Medicine, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA*
[4]*Department of Surgery, Hospital of the University of Pennsylvania, Philadelphia, PA, USA*
[5]*Department of Neurology, The University of Texas Southwestern Medical Center, Dallas, TX, USA*
[6]*Department of Anatomy and Neurobiology, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA*
[7]*The Framingham Heart Study, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA*
[8]*Department of Neurology, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA*
[9]*Boston University Alzheimer's Disease Research Center, Boston, MA, USA*
[10]*Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA*
[11]*Faculty of Computing & Data Sciences, Boston University, MA, USA*

* These authors contributed equally to this work
† Listed in alphabetical order
‡Corresponding author: Vijaya B. Kolachalama, PhD; Email: vkola@bu.edu; ORCID: https://orcid.org/0000-0002-5312-8644

Abstract

The proliferation of medical podcasts has generated an extensive repository of audio content, rich in specialized terminology, diverse medical topics, and expert dialogues. Here we introduce a computational framework designed to enhance large language models (LLMs) by leveraging the informational content of publicly accessible medical podcast data. This dataset, comprising over $4,300$ hours of audio content, was transcribed to generate over 39 million text tokens. Our model, MedPodGPT, integrates the varied dialogue found in medical podcasts to improve understanding of natural language nuances, cultural contexts, and medical knowledge. Evaluated across multiple benchmarks, MedPodGPT demonstrated an average improvement of 2.31% over standard open-source benchmarks and showcased an improvement of 2.58% in its zero-shot multilingual transfer ability, effectively generalizing to different linguistic contexts. By harnessing the untapped potential of podcast content, MedPodGPT advances natural language processing, offering enhanced capabilities for various applications in medical research and education.

The emergence of generative artificial intelligence (AI), particularly through the development of large language models (LLMs), has marked a significant progression in data analysis and interpretation. Trained on extensive text corpora, these models have demonstrated their ability to generate contextually rich and accurate content, showcasing advanced analytical prowess. Notable achievements, such as GPT-4's performance in medical examinations, underscore the potential of LLMs to revolutionize various disciplines, including medicine [1]. Amid these advancements, the proliferation of medical podcasts has introduced a vast collection of audio content, rich in medical terminology, topic diversity, and expert dialogues. This burgeoning trend not only serves as a medium for disseminating the latest medical knowledge but also provides a unique opportunity to mine linguistic patterns and domain-specific knowledge, enhancing the capabilities of language models within the healthcare and clinical sectors.

Multimodal foundation models in medicine represent a significant leap forward, merging textual information with intricate data forms like medical imaging. These models excel in synthesizing and producing content that seamlessly integrates text and visuals, enhancing the comprehension of radiological and pathological imagery [2–6]. By training on comprehensive datasets that combine medical narratives with related visual elements, these models facilitate a deeper understanding of complex medical phenomena, thus improving diagnostic accuracy and the quality of medical education [4]. The integration of various data modalities, such as audio content from medical podcasts and lectures, pioneers a promising direction to refine language models' precision and relevance to the medical domain. Given the substantial progress in audio transcription technologies, there is a ripe opportunity to develop audio-linguistic foundation models that advance existing large language model capabilities. Such advancements could significantly enhance medical research and education.

In this work, we developed MedPodGPT, a computational framework designed to enhance language models by leveraging the depth of linguistic and informational content inherent in medical podcasts. It integrates the diverse dialogues from these podcasts to enhance its capability to interpret complex medical information and generate contextually informed content. MedPodGPT is pretrained on a vast corpus of podcast transcripts, encompassing specialized academic discussions. This database allows MedPodGPT to capture a wide range of linguistic styles and terminologies in the medical field, thus refining its ability to process and generate relevant texts. The development of MedPodGPT not only aligns with the complex nature of medical communication but also signifies a major step toward more informed research and educational purposes. By leveraging the untapped potential of medical podcasts, MedPodGPT can foster significant advancements in medical language understanding, ultimately enhancing the quality and popularity of medical and clinical knowledge.

Methods

Dataset description We curated a diverse collection of multilingual medical podcasts encompassing various types of medical knowledge. These podcasts are primarily created by biomedical and clinical journals, medical exam preparation organizations, and clinician educators, aiming to teach medical students, resident physicians, and other learners about different aspects of medical practice. To ensure a broad linguistic perspective, we collected publicly available podcasts in English, Spanish, and French, which rank among the top 5 most widely spoken languages in the world. This collection offers high-quality content that covers cultural ethics and inclusion, natural language nuances, medical knowledge, and clinical practices, thus potentially advancing the understanding and translation of medical science. Specifically, the medical podcast data was selected and filtered based on the following criteria: (1) Podcasts hosted by reputable scientific, medical, and clinical journals. (2) Podcasts aimed at preparing medical students for standardized medical examinations that are hosted by physicians, medical professionals, or organizations. (3) Podcasts produced by individuals with medical expertise, including Medical Doctors (M.D.) or Doctors of Philosophy (Ph.D.), discussing clinical and medical knowledge to educate medical students and residents. Finally, medical experts on our team reviewed each podcast to ensure the topics were limited to didactic medicine and clinical practice. The complete set of podcast episodes, along with relevant metadata, is detailed in Table 1.

Dataset processing The pretraining corpus for MedPodGPT consisted of thousands of hours of medical podcasts, encompassing academic discussions, clinical case studies, and expert interviews. We transcribed these audio files using the state-of-the-art automatic speech recognition model, OpenAI Whisper [7]. Built upon an encoder-decoder Transformer architecture, the Whisper model resampled the input audio to $16,000$ Hz and performed temporal chunking. Then, these chunks of audio data were represented by 80-channel log-magnitude Mel spectrograms with a 25-millisecond window and 10-millisecond stride. Before being processed by the Transformer modules, the input underwent a convolutional layer and was augmented with the sinusoidal position embeddings to incorporate positional information. Finally, the Transformer decoder module further interpreted the hidden representation of the audio data and generated textual output through a language head [8]. We utilized the latest Whisper series model, the Whisper large-v3, with $1,550$M parameters, to specify the spoken language for improved speech recognition.

All the transcripts were cleaned and preprocessed to remove unnecessary information and ensure consistency. Initially, we automatically removed sentences with duplicated content. Additionally, sentences containing words or characters in other languages were cleaned to avoid transcription errors. Finally, all the podcast transcripts were carefully and manually reviewed to maintain content quality. Consequently, $2,836,327$, and 34 sentences were filtered from English, Spanish, and French data, accounting for 0.16%, 0.14%, and 0.04% of the total content, respectively. This diverse and high-quality dataset ensured that MedPodGPT was well-equipped to handle a wide range of medical queries with high precision and contextual relevance.

Model architecture The Transformer model [9], renowned for its multi-head self-attention mechanism, has become the backbone of many state-of-the-art AI models. Unlike traditional methods, the self-attention mechanism of the Transformer model captures long-range dependencies with efficient parallelization and scalability. Additionally, its deep feedforward neural networks enhance the model's capacity to learn complex patterns in data. Our proposed MedPodGPT leverages this advanced architecture and is designed for medical and educational purposes. Built upon state-of-the-art general large language models (LLMs) such

as Gemma [10], LLaMA [11], and Mistral [12], MedPodGPT is pre-trained on a diverse corpus of textual data extracted from medical podcasts. By utilizing instruction-tuned variants of these LLMs, we aimed to improve instruction-following capabilities and conversational structure. To evaluate the effectiveness of our framework, we applied models of varying scales, ranging from 2 to 70 billion parameters.

Gemma is a series of lightweight open models developed by Google DeepMind. These are text-to-text auto-regressive language models, which have pre-trained versions as well as fine-tuned variants. These models were trained on the textual datasets on a context length of 8192 from a wide variety of sources. The primary sources include web documents, codes, and mathematical content. Several recent advancements have been made to improve the performance and training efficiency of the Transformer model. These include multi-query attention [13], rotary positional embeddings [14], and GeGLU activations [15]. We utilized the Gemma models 2B and 7B to validate our framework across different model sizes. LLaMA is a family of advanced general-purpose LLMs released by Meta Research, with a publicly accessible 70B model for outstanding capability. These models are constructed using a decoder-only Transformer architecture as well as the grouped-query attention for improved efficiency [16]. Both model variants 8B and 70B were pretrained using a context length of 8192 on publicly available sources, with over 5% of the pretraining dataset consisting of non-English data covering over 30 languages. We encoded medical knowledge into the 8B and 70B weights to enhance their understanding of medicine. The Mistral series has been open-sourced by the Mistral AI team for open and portable generative AI. The models employ the standard decoder-only architecture with improved efficiency using grouped-query and sliding window attention, rolling buffer cache, and chunking. Furthermore, the initial generative Sparse Mixture of Experts (MoE) model was designed to balance computational load and model capability. In this work, we implemented our framework to the 7B and $7\times$8B MoE models to assess the MoE-based architecture.

Pre-training is a crucial step in the development of LLMs, during which the model learns from a vast corpus of text data in an auto-regressive manner. This phase generally leverages self-supervised learning, employing methods like masked language modeling, *e.g.*, BERT [17], or autoregressive modeling, *e.g.*, GPT [18]. The self-supervised learning framework allows the model to gain a broad understanding of knowledge, thereby improving its performance in subsequent tasks. In this work, we utilized an auto-regressive objective to perform continual pretraining through an iterative gradient solver. The above-mentioned LLMs have been pre-trained on trillions of tokens. Thus, one cost-effective and efficient way to encode domain-specific knowledge is through continuous pre-training and evolving the pre-trained models with expertise corpora, instead of retraining them from scratch. The podcast transcripts were represented by a sequence of tokens, *i.e.*, $\mathbf{x} = \{x_1, x_2, \ldots x_N\}$, where x_i is a subword token and N denotes the length of the sequence. We trained our podcast data on pre-trained models in an auto-regressive manner, optimizing the models by minimizing the negative log likelihood. The training objective is as follows,

$$\mathcal{L}_{\pi_\theta} = -\sum \log\left(\pi_\theta(x_i|\mathbf{x}_{<i})\right),$$

where π_θ is the language model, parameterized by θ.

Experimental settings To comprehensively analyze the capabilities of MedPodGPT, we employed a wide range of model sizes and conducted extensive experiments on multilingual medical knowledge. In current literature, benchmarks for multiple-choice question-answering (QA) were commonly utilized to evaluate the performance of large medical language models. Thus, in this work, we utilized the multilingual multiple-choice QA benchmarks to evaluate the model's performance. In addition, we conducted experi-

ments and documented the performance of all the models that were used in this study on multilingual medical benchmarks. This potentially advances the field with an open-source and unified multilingual benchmarking library covering training, inference, answer extraction, performance evaluation, and real-world deployment. Furthermore, to guarantee scientific reproducible research, we implemented all our experiments with a set of unified hyperparameters. Thus, our work was out of the box without any specific hyperparameter tuning and further optimization for different models.

Evaluation benchmarks To evaluate the performance of MedPodGPT, we utilized a comprehensive set of medical benchmarks from the most spoken languages in the world, including English, Mandarin, French, Spanish, and Hindi. For intra-language experiments, we performed performance evaluations on datasets where the language aligned with the podcast content. Furthermore, for cross-language experiments, the model was evaluated on benchmarks in different languages compared to the podcasts. This evaluation was crucial for validating the effectiveness of the zero-shot multilingual transfer capability of medical LLMs. The detailed descriptions of multilingual benchmarks are as follows.

Medical benchmarks in English The benchmarks for medical natural language understanding in English have been significantly advanced over the past decade. In this study, we selected five well-known publicly accessible benchmarks, which include MedQA [19], PubMedQA [20], MedMCQA [21], MedExpQA [22], and MMLU clinical topics [23]. For the MMLU benchmark, we followed the Google PaLM work and chose six clinical subcategories, *i.e.*, anatomy, clinical knowledge, college biology, college medicine, medical genetics, and professional medicine [24]. These benchmarks cover a wide range of medical topics and question formats, providing a robust evaluation framework to assess the model's capabilities.

Medical benchmarks in Chinese The benchmarking of medical and clinical knowledge in Chinese has become increasingly popular recently. A range of databases have been successively proposed to assess the performance of Chinese language models on medical data [25]. In this study, we adopted the popular MedQA-MCMLE [19] and CMMU medical topics [25]. For the CMMLU benchmark, the medical and clinically related subsets were utilized, containing anatomy, clinical knowledge, medical school, genetics, nutrition, traditional Chinese medicine, and virology [26]. These eight datasets provide comprehensive and in-depth Chinese medical contexts for evaluating the knowledge and reasoning capabilities of multilingual language models.

Medical benchmarks in Spanish The Spanish medical testbed encourages the NLP community to develop new approaches for understanding and reasoning medical and clinical knowledge in Spanish. The HEAD-QA benchmark was utilized in our research. It is a multiple-choice healthcare dataset obtained from examinations in the Spanish healthcare system [27]. Additionally, we also employed the MedExpQA Spanish subset [22] and Spanish MMLU clinical topics [26]. The selected benchmarks cover various medical content, including medical, clinical, and healthcare knowledge, providing an adequate platform to evaluate the model's performance in Spanish.

Medical benchmarks in French We primarily selected the popular FrenchMedMCQA dataset, which consists of $3,105$ questions taken from the French pharmacy diploma examinations [28]. Following Wang et al., we only performed performance evaluations on questions with a single answer [26]. As a result, the total number of questions in the testing set was 321. Furthermore, the MedExpQA French subset [22] and French MMLU clinical topics [26] were also included in this work. The databases mentioned above played a significant role in interpreting French medical knowledge and assessing the performance of models in French.

Medical benchmarks in Hindi To encode medical and clinical content in Hindi, we included the Hindi MMLU clinical topics [26] in our benchmarking. Thus, we can evaluate the model's ability to understand medical language in Hindi and cover one of the most widely spoken languages in the world. It also sets the standards for evaluating multilingual LLMs in the medical and clinical field.

Implementation details We began transcribing podcast data using the OpenAI Whisper large-v3 model for an automatic speech recognition task. The chunk length was set to 30 seconds with a 5-second stride on both sides to improve the continuity and coherence of the transcriptions. The batch size was 96, and 384 tokens were generated per chunk to parallelly process audio chunks.

We encoded medical knowledge and clinical practice across a wide range of model sizes, from 2B to 70B. We have implemented publicly available language models, which include the 2B and 7B versions of the Gemma series, the most recent fine-tuned version (v0.3) of the Mistral 7B family, the instruction-tuned variant of the LLaMA 3 8B collections, the first open-sourced MoE model, which is the Mixtral 8×7B sparse MoE, and the instruction-tuned generative text LLaMA models in 70B. During model training, we utilized Brain float 16 data type with the AdamW optimizer to prevent overflow issues [29], and the context window was set to $2,048$ [26]. We trained all models for 5 epochs with an initial learning rate of 5×10^{-6} with a 0.03 warm-up ratio and a cosine schedule. The weight decay rate was 0.01, and the gradient was accumulated during each training step. Due to the computational limit, we have employed the 8-bit quantized AdamW optimizer and implemented the low-rank adaptation (LoRA) in the Mixtral 8×7B sparse MoE and LLaMA 3 70B models. The rank and alpha were set to 16 and 32, respectively, and the dropout rate was 0.1. All the models were optimized based on the unified hyper-parameter settings without specific tuning for superior performance.

Software and database infrastructure We created a custom graphical user interface (GUI) and platform infrastructure to allow users to interact with MedPodGPT, providing public access to our model. Our goal was to deliver our model with a user-friendly and responsive conversational interface. For hardware, we utilized a custom-built method to deploy our LLMs at scale using entirely self-hosted and open-source tools without relying on software as a service (SaaS) or proprietary software. MedPodGPT's hardware included 4 NVIDIA RTX 3080 Ti GPUs and 3 production servers, each with 4 CPU cores and 8GB RAM. This setup is modest, supporting only hundreds of individual users per day, but the architecture can be quickly scaled to match the load.

We employed a microservice architecture using Kubernetes as a container orchestration tool. Kubernetes manages clusters of nodes hosting microservices wrapped inside Docker containers. It facilitates the creation of highly available distributed systems that automatically scale to meet needs and ensure secure inter-cluster communication, IP address allocation, load balancing, and reverse proxy services. We utilized ReactJS and NextJS for the front end. ReactJS furnishes a collection of APIs and libraries to construct reusable web components, while NextJS provides scaffolding for ReactJS applications, encompassing an HTTP server, server-side rendering, and a "back end for a front end" design pattern. For LLM deployment, we employed the vLLM library, which offers a fast and portable inference server that batches inference tasks efficiently [30]. It requires a minimum of NVIDIA 11 and a 7.5 compute-capable NVIDIA GPU, supporting several GPUs on different host machines simultaneously.

Authentication and user management are crucial components of our architecture. In order to distribute resources equitably among potential researchers, we have implemented OAuth 2.0 compliant authorization and user management in addition to a per-token rate limiting system based on user scopes and total system load. MedPodGPT implements features which users familiar with widely-used chatting services expect, such as multiturn conversations and the ability to open many conversations. Furthermore, MedPodGPT utilized Apache Cassandra, a distributed NoSQL database designed for high availability and query optimization. The backend API router, which was built with Flask, stores new chats and conversations in Cassandra and sends text inference requests to a queue. For queuing and message processing, we utilized RabbitMQ and Redis, which are a message broker and key-value databases, respectively. Each fine-tuned model can be assigned its own queue in RabbitMQ to receive messages. When a user requests a message, the conversation is processed by a vLLM gateway module. This module asynchronously generates text completions from vLLM, acknowledges the message to the broker, and stores the message in Redis. The API then serves the completed text inference via another text completion endpoint, referenced by a unique text completion ID.

Data and model availability A multilingual LLMs benchmarking library along with the source codes are made available at https://github.com/vkola-lab/MedPodGPT.

Results

We conducted comprehensive experiments to assess MedPodGPT's performance on various multilingual medical QA benchmark datasets. Our results demonstrate that incorporating medical audio podcast data enhances the model's ability to understand and generate medically relevant information. In addition, the models across a wide range of scales outperformed their respective baselines in both in-domain benchmarks and zero-shot domain generalization across multilingual medical datasets.

Performance on in-domain benchmarks The evaluation of MedPodGPT across diverse medical question-answering benchmarks demonstrated enhanced model efficacy following pre-training with multilingual medical podcast datasets (Table 2). Specifically, on the MedExpQA benchmark, MedPodGPT achieved significant performance gains, *i.e.*, a 10.80% increase with the Gemma 7B model, 8.40% with the Mixtral 8×7B MoE, and 8.20% with the Gemma 2B model. In MedMCQA, improvements were notable, with the Gemma 7B model increasing by 4.20% and the Mixtral MoE by 3.34%. Additionally, the Gemma 7B model showed enhancements of 6.30% and the 2B model 3.69% on the MedQA database. Evaluation on French benchmarks revealed substantial improvements, with MedPodGPT achieving 10.67% and 9.81% enhancements on FrenchMedMCQA with Gemma 7B and LLaMA 3 70B models, respectively. Moreover, on French MedExpQA, the Gemma 7B model outperformed the baseline by a remarkable 12.80%. In Spanish benchmarks, the Gemma 7B model of MedPodGPT demonstrated improvements of 6.26% on HeadQA and 5.60% on MedExpQA. Lastly, across multilingual MMLU benchmarks, MedPodGPT consistently surpassed baseline models, achieving improvements up to 13.50% and averaging 7.23%. Overall, MedPodGPT showed a cumulative 2.31% enhancement across in-domain benchmarks, highlighting the advantage of leveraging open-source multilingual podcast datasets to enhance model efficacy.

As shown in Table S1, we further evaluated MedPodGPT across various English medical QA benchmarks after pre-training with English medical podcast data. On the MedExpQA dataset, MedPodGPT demonstrated a notable increase of 6.60% in the Gemma 2B model, 7.80% in the Gemma 7B model, and 7.00% in the Mixtral MoE model. Similarly, on the MedMCQA dataset, there were improvements of 3.89% in the Gemma 7B model and 2.59% in the Mistral 7B model. For the MedQA dataset, the performance enhancements included a 6.87% increase in the Gemma 7B model and a 3.85% increase in the Gemma 2B model. In the PubMedQA dataset, the Gemma 2B model saw an improvement of 9.40%. In the MMLU anatomy dataset, the Mixtral MoE and Gemma 7B improved by 2.97% and 2.59%, respectively. Additionally, for the college biology dataset, there were increases of 4.34% in the Gemma 2B model, 8.16% in the Gemma 7B model, and 4.68% in the Mistral 7B model. For the college medicine dataset, the Gemma 7B and Mixtral MoE models showed increases of 4.19% and 4.05%, respectively. Lastly, in the clinical knowledge dataset, the Gemma 7B model showed a 7.07% improvement, while the Mixtral MoE model had an increase of 7.84%. These results underscore the effectiveness of integrating podcast data into the training process, resulting in performance gains across most instances, with an average improvement of 2.16%.

Zero-shot cross-lingual performance In Table 3, we validated MedPodGPT's zero-shot cross-lingual performance using multilingual benchmarks. These benchmarks encompass a wide array of medical subjects, including traditional Chinese medicine, medical nutrition, and Hindi MMLU. The Gemma 7B model of MedPodGPT showcased a significant 5.47% improvement on the MedQA-MCMLE benchmark. Moreover, it exhibited superior performance on CMMLU benchmarks, achieving average increases up to 5.19%.

Remarkably, the Gemma 7B model achieved significant performance improvements of 8.65%, 8.29%, and 6.59% on CMMLU benchmarks focusing on clinical knowledge, anatomy, and virology topics. Lastly, across Hindi benchmarks, particularly clinical knowledge, medical genetics, and professional medicine, MedPodGPT demonstrated notable performance gains, with improvements reaching up to 10.94% across various models. Overall, MedPodGPT demonstrated its superiority by enhancing its zero-shot multilingual transfer capability, achieving an average improvement of 2.58% across models and effectively generalizing to diverse linguistic contexts.

In addition, MedPodGPT was trained on English podcast data, and its zero-shot transfer capability was assessed as well in Table S2. These benchmarks encompass a wide range of medical subjects, including traditional Chinese medicine, French pharmaceutical examinations, and specialized assessments in the Spanish healthcare system. MedPodGPT showed improved performance on multilingual MMLU and CMMLU benchmarks. In Mandarin benchmarks, such as MedQA-MCMLE and clinical knowledge, MedPodGPT outperformed the baseline models, showing an average improvement of 1.87%. It also achieved enhancements of up to 7.28% on Mandarin benchmarks. Second, for the French benchmarks, including FrenchMedMCQA and MedExpQA, MedPodGPT demonstrated notable performance gains, with improvements ranging from 1.72% to 3.87% across different categories. Lastly, in the Hindi and Spanish benchmarks, the model also exhibited enhanced performance, particularly in categories such as anatomy and clinical knowledge, where it showed increases of up to 11.67%. Overall, MedPodGPT exhibited a 2.28% enhancement in zero-shot multilingual transfer, further propelling AI advancements in medicine.

Discussion

We present MedPodGPT, a large language model that leverages the rich and diverse linguistic content of medical podcasts, capturing a wide array of medical terminologies and conversational contexts. Extensive pre-training on podcast data has endowed MedPodGPT with the capability to generate relevant medical information. When benchmarked against existing datasets such as MedQA, PubMedQA, MedMCQA, and various MMLU categories, MedPodGPT demonstrated superior performance, particularly in areas requiring detailed medical knowledge and contextual understanding. These results highlight its potential to serve as a valuable tool for medical education and research.

Our results indicate that our audio-augmented LLM framework improves the accuracy and relevance of medical information generated by the model. This enhancement is particularly evident when compared to a series of baseline models, such as Google Gemma, Meta LLaMA, and Mistral models, where MedPodGPT consistently outperformed these models across multiple benchmarks. This demonstrates that incorporating audio data provides a richer understanding of medical conversations, which is crucial for accurate medical language processing.

Our study has a few limitations. First, we focused on publicly available medical podcasts based on content feasibility and availability. While we incorporated content from popular medical podcasts, there are certainly more medically relevant contexts available, such as textbooks and even video tutorials. Extending the language medium beyond English, we downloaded multilingual medical podcast data, specifically Spanish and French. We sought to include podcasts in Hindi and Mandarin, but we found relevant content to be limited. Despite these constraints, our model successfully learned from the multilingual podcast content, performing well on respective language benchmarks and even showing zero-shot performance on Hindi and Mandarin benchmarks. In the future, we aim to acquire richer and more relevant podcast data in numerous languages to further enhance model training and performance. Future work on MedPodGPT should also include a comprehensive ethical evaluation to ensure the model consistently adheres to high standards in diverse settings. Also, we observed that pre-training using podcast data did not improve performance on a few benchmarks. This finding can be attributed to the nature and structure of podcasts, which contrasts with the demands of these benchmarks. Podcast data, while rich in narrative and contextual content, lacks the precision, structure, and specific terminologies found in traditional medical texts and scientific literature. The informal and conversational style of podcasts may not align well with the formal, structured, and detail-oriented requirements of benchmarks such as PubMedQA, clinical knowledge, and professional medicine. To address this limitation and enhance performance, it is crucial to complement podcast training data with more structured and detailed medical texts, ensuring a balanced and comprehensive training dataset.

The findings from this study indicate that MedPodGPT represents an important advancement in the application of language models for medical applications. Its ability to process and generate medically relevant text holds promise for enhancing medical education and research. However, the deployment of such advanced models must be accompanied by rigorous considerations, particularly concerning patient confidentiality and data integrity. By continuing to advance the intersection of AI and medicine, we can ultimately improve the accessibility and quality of medical education and research, ensuring that such technologies benefit trainees and researchers alike. MedPodGPT highlights the value of integrating podcast data to enhance language models, which can be extended to applications beyond health and medicine by incorporating diverse audio podcasts.

Acknowledgements

This project was supported by grants from the Karen Toffler Charitable Trust (V.B.K.), the National Institute on Aging's Artificial Intelligence and Technology Collaboratories (P30-AG073014, V.B.K.), the American Heart Association (20SFRN35460031, V.B.K. & R.A.), Gates Ventures (R.A. & V.B.K.), and the National Institutes of Health (R01-HL159620 [V.B.K.], R21-CA253498 [V.B.K.], R43-DK134273 [V.B.K.], RF1-AG062109 [R.A. & V.B.K.], and U19-AG068753 [R.A.]).

Author contributions

S.J., S.G., and E.S. contributed equally to this work. S.J., S.G., L.A.C., P.F., V.H.J., M.V.L., and D.V. curated and processed the data. S.J. and S.G. performed model training. E.S. and W.M.W. worked on software development. S.J., S.G., L.A.C., P.F., V.H.J., M.V.L., E.S., D.V. and W.M.W. generated the results. R.A. provided clinical context. V.B.K. wrote the manuscript. All authors reviewed, edited, and approved the manuscript. V.B.K. conceived, designed, and directed the study.

Ethics declarations

V.B.K. is on the scientific advisory board for Altoida Inc. and serves as a consultant to AstraZeneca. R.A. is a scientific advisor to Signant Health and NovoNordisk. The remaining authors declare no competing interests.

References

1. Thirunavukarasu, A. J. *et al.* Large language models in medicine. *Nature Medicine* **29**, 1930–1940 (2023).

2. Liu, G. *et al.* Medical-VLBERT: Medical visual language BERT for COVID-19 CT report generation with alternate learning. *IEEE Transactions on Neural Networks and Learning Systems* **32**, 3786–3797 (2021).

3. Zhang, X., Wu, C., Zhang, Y., Xie, W. & Wang, Y. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications* **14**, 4542 (2023).

4. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. & Zou, J. A visual-language foundation model for pathology image analysis using medical twitter. *Nature Medicine* **29**, 2307–2316 (2023).

5. Lu, M. *et al.* A visual-language foundation model for computational pathology. *Nature Medicine* **30**, 863–874 (2024).

6. Huemann, Z., Tie, X., Hu, J. & Bradshaw, T. ConTEXTual Net: A multimodal vision-language model for segmentation of pneumothorax. *Journal of Imaging Informatics in Medicine* (2024).

7. Radford, A. *et al.* Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, 28492–28518 (PMLR, 2023).

8. Press, O. & Wolf, L. Using the output embedding to improve language models. In Lapata, M., Blunsom, P. & Koller, A. (eds.) *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 157–163 (Association for Computational Linguistics, Valencia, Spain, 2017).

9. Vaswani, A. *et al.* Attention is all you need. vol. 30 (2017).

10. Team, G. *et al.* Gemma: Open models based on Gemini research and technology (2024).

11. Touvron, H. *et al.* LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

12. Jiang, A. Q. *et al.* Mistral 7b (2023).

13. Shazeer, N. Fast Transformer decoding: One write-head is all you need (2019).

14. Su, J. *et al.* RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing* **568**, 127063 (2024).

15. Shazeer, N. GLU variants improve Transformer (2020).

16. Ainslie, J. *et al.* GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 4895–4901 (2023).

17. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional Transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186 (2019).

18. Brown, T. *et al.* Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)* **33**, 1877–1901 (2020).

19. Jin, D. *et al.* What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences* **11**, 6421 (2021).

20. Jin, Q., Dhingra, B., Liu, Z., Cohen, W. & Lu, X. PubMedQA: A dataset for biomedical research question answering. In Inui, K., Jiang, J., Ng, V. & Wan, X. (eds.) *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2567–2577 (Association for Computational Linguistics, Hong Kong, China, 2019).

21. Pal, A., Umapathi, L. K. & Sankarasubbu, M. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Flores, G., Chen, G. H., Pollard, T., Ho, J. C. & Naumann, T. (eds.) *Proceedings of the Conference on Health, Inference, and Learning*, vol. 174 of *Proceedings of Machine Learning Research*, 248–260 (PMLR, 2022).

22. Alonso, I., Oronoz, M. & Agerri, R. MedExpQA: Multilingual benchmarking of large language models for medical question answering. *arXiv preprint arXiv:2404.05590* (2024).

23. Hendrycks, D. *et al.* Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)* (2021).

24. Singhal, K. *et al.* Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).

25. Li, H. *et al.* CMMLU: Measuring massive multitask language understanding in Chinese. *arXiv preprint arXiv:2306.09212* (2023).

26. Wang, X. *et al.* Apollo: An lightweight multilingual medical LLM towards democratizing medical AI to 6b people. *arXiv preprint arXiv:2403.03640* (2024).

27. Vilares, D. & Gómez-Rodríguez, C. HEAD-QA: A healthcare dataset for complex reasoning. In Korhonen, A., Traum, D. & Màrquez, L. (eds.) *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 960–966 (Association for Computational Linguistics, Florence, Italy, 2019).

28. Labrak, Y. *et al.* FrenchMedMCQA: A French multiple-choice question answering dataset for medical domain. In Lavelli, A. *et al.* (eds.) *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, 41–46 (Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022).

29. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)* (2019).

30. Kwon, W. *et al.* Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, 611–626 (2023).

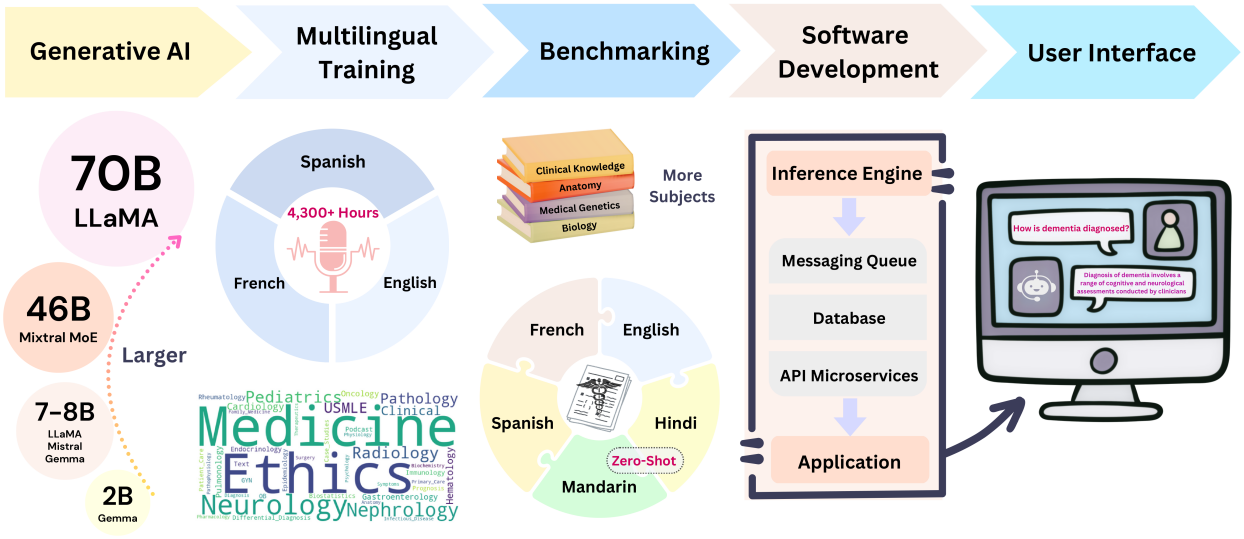

Figure 1: **MedPodGPT framework.** This figure illustrates the workflow and components involved in developing MedPodGPT, a multilingual audio-augmented large language model designed for medical research and education. The process began by utilizing publicly available generative AI auto-regressive language models across various scales, including the Gemma series, LLaMA collections, and the Mistral family. These models underwent multilingual pre-training on podcast content from journals, exam preparation materials, and clinical practice in English, Spanish, and French, totaling over $4,300$ hours of context covering diverse medical topics indicated in the word cloud. Following pre-training, the models were evaluated using multilingual medical question-answering benchmarks, spanning various subjects, including clinical knowledge, anatomy, medical genetics, and biology, in the most commonly spoken languages worldwide. Additional benchmarks in Hindi and Mandarin were also employed to assess MedPodGPT's zero-shot transfer capability. The next phase involved software development, encompassing the inference engine for model deployment, messaging queue, database, API microservices, and responsive human-machine interface. This infrastructure enables users to engage through a chat interface supported by an adaptive chatbot, facilitating multi-turn conversations.

Journal Podcasts						
Podcast	**Language**	**Episodes**	**Audio Time (min)**	**Mean Length Episode $\pm\sigma$ (min)**	**Number of Text Tokens**	**Mean Text Tokens per Episode $\pm\sigma$**
NEJM	English	1974	39256.0	19.89 ± 9.74	4,760,783	1928.22 ± 14.87
JAMA	English	2235	32163.0	14.39 ± 8.66	3,454,191	1928.64 ± 15.54
The Lancet	English	2029	28279.0	13.94 ± 7.62	3,300,982	1925.89 ± 20.88
The BMJ	English	300	13264.2	44.21 ± 10.35	2,235,458	1897.67 ± 75.07
Annals Latest Highlights	English	396	6427.0	16.23 ± 8.09	803,958	1927.96 ± 14.60
Annals On Call	English	142	3440.0	24.22 ± 3.65	522,547	1928.22 ± 15.64
Pediatrics on Call	English	98	3299.0	33.66 ± 6.09	565,781	1930.99 ± 15.00
Procedure Ready: Ob/Gyn	English	20	383.7	19.19 ± 5.00	63,667	1929.30 ± 13.41
Revista Médica AFP Podcast	Spanish	40	1055.0	26.38 ± 3.70	190,518	1924.42 ± 16.34
Test Preparation Podcasts						
Podcast	**Language**	**Episodes**	**Audio Time (min)**	**Mean Length of Episode $\pm\sigma$ (min)**	**Number of Text Tokens**	**Mean Text Tokens per Episode $\pm\sigma$**
Divine Intervention Podcasts	English	480	18363.8	38.26 ± 24.07	2,269,153	1931.19 ± 13.53
The Radiology Review Podcast	English	127	2517.7	19.82 ± 10.10	292,949	1927.29 ± 26.81
Crush Step 1: The Ultimate USMLE Step 1 Review	English	49	2176.2	44.41 ± 15.31	328,194	1930.55 ± 13.03
The USMLE Guys Podcast	English	31	1464.3	47.24 ± 43.47	156,923	1937.32 ± 6.12
Harrison's PodClass: Internal Medicine Cases and Board Prep	Spanish	95	905.2	9.53 ± 2.24	101,574	1916.49 ± 22.77
El Interno Desvelado	Spanish	4	99.13	24.78 ± 11.91	17,121	1902.33 ± 25.54
Curso MIR Asturias	Spanish	3	17.7	5.89 ± 4.53	3,872	1936.00 ± 9.00
Clinical Podcasts						
Podcast	**Language**	**Episodes**	**Audio Time (min)**	**Mean Length of Episode $\pm\sigma$ (min)**	**Number of Text Tokens**	**Mean Text Tokens per Episode $\pm\sigma$**
The Curbsiders Internal Medicine Podcast	English	485	28749.2	59.39 ± 16.08	5,772,083	1929.82 ± 17.06
This Podcast Will Kill You	English	168	18363.8	38.26 ± 24.07	2,269,153	1931.19 ± 13.53
The Clinical Problem Solvers	English	315	13500.1	42.86 ± 14.51	2,493,777	1927.18 ± 23.32
PsychEd: educational psychiatry podcast	English	62	3556.3	57.36 ± 17.52	607,237	1927.74 ± 16.36
Run the List	English	97	1973.0	20.34 ± 6.44	352,977	1928.84 ± 15.17
Goljan Pathology Lectures	English	37	1886.0	50.97 ± 4.58	412,086	1934.68 ± 13.45
Core IM: 5 Pearls	English	54	1847.1	34.21 ± 10.19	361,213	1931.62 ± 10.16
Neurology Clinical Pearls	English	27	333.2	12.34 ± 3.19	42,494	1931.54 ± 10.78
Tutorías Medicina Interna	Spanish	570	19834.9	34.80 ± 25.01	4,311,263	1898.39 ± 64.31
Leucocitos isotópicos	Spanish	68	2537.8	37.32 ± 9.55	481,676	1797.29 ± 154.42
Medicina Con Cabeza	Spanish	246	2457.8	9.99 ± 3.44	462,383	1902.81 ± 57.55
Medicina de impacto	Spanish	57	2406.5	42.22 ± 9.13	492,363	1915.81 ± 29.28
Ronda, El Podcast de Medicina Interna	Spanish	20	1084.4	54.22 ± 25.01	206,218	1891.91 ± 71.90
Medicina De Bolsillo | Hablando de Medicina	Spanish	45	958.3	21.30 ± 10.79	186,268	1844.24 ± 124.82
La Tertulia de Cajal	Spanish	27	876.3	32.46 ± 18.28	186,001	1897.97 ± 57.71
PedCast: Dos Pediatras y un Podcast	Spanish	14	458.5	32.75 ± 10.62	89,127	1896.32 ± 58.05
Neurobiologie et Immunite	French	21	1882.8	89.66 ± 14.77	383,189	1896.97 ± 40.12
Incubateur Néonat	French	25	1579.3	63.17 ± 21.28	391,475	1918.99 ± 24.15
Guideline.care	French	68	1369.1	20.13 ± 6.30	293,301	1917.0 ± 29.29
La Minute Rhumato	French	119	921.0	7.74 ± 2.19	132,354	1918.17 ± 23.59
Oncologie cellulaire et moléculaire - Hugues de Thé	French	11	852.9	77.53 ± 19.81	186,693	1905.03 ± 44.42
Le podcast des Conférenciers (UFR3S) by Université de Lille	French	65	768.4	11.82 ± 19.58	86,105	1913.44 ± 42.78
Super Docteur	French	47	676.3	14.39 ± 6.50	139,824	1915.40 ± 33.26
Médecine, Sciences et Recherche clinique	French	24	332.2	13.84 ± 4.58	63,314	1918.61 ± 26.60
NéphrOdio	French	40	318.6	7.96 ± 2.58	55,716	1921.24 ± 19.59
La Minute Néonat	French	37	307.6	8.31 ± 1.93	57,435	1914.50 ± 31.58
Le Med G Eclairé	French	11	249.2	22.66 ± 16.76	51,988	1925.48 ± 12.57
La Minute du Pancréas	French	22	209.4	9.52 ± 2.34	38,376	1918.80 ± 23.34
L'essentiel des principales pathologiesaâ	French	14	151.3	10.81 ± 13.10	23,098	1924.83 ± 11.25
AR-Pod le Podcast de lanesthésie-réanimation	French	12	139.0	11.59 ± 4.52	22,998	1916.50 ± 26.48

Table 1: **Podcasts used for model development.** This table presents an overview of 46 journal, test preparation, and clinical podcasts used for the continual pre-training of MedPodGPT. It includes information on podcast names, languages, number of episodes, total audio time, mean length of episodes with standard deviation, number of text tokens, and mean text tokens per episode with standard deviation. For journal podcasts, NEJM, JAMA, The Lancet, and the BMJ have extensive episode counts with significant audio durations and token counts, showcasing their depth and breadth in medical discussions. Test preparation podcasts like "Crush Step 1" and "Divine Intervention" highlight detailed USMLE preparation with varying episode lengths and comprehensive content coverage. Clinical podcasts such as "The Clinical Problem Solvers" and "The Curbsiders Internal Medicine Podcast" emphasize educational content for medical professionals, with substantial episode counts and detailed discussions. The data from these podcasts, transcribed using OpenAI Whisper, demonstrates the diverse and robust dataset used for enhancing MedPodGPT's medical knowledge and comprehension.

Language	Benchmark Datasets	Model											
		Gemma 2B		Gemma 7B		Mistral 7B		LLaMA 3 8B		Mixtral MoE		LLaMA 3 70B	
		Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*
English	MedExpQA	15.20	**23.40**	34.40	**45.20**	**47.20**	46.20	57.60	**61.40**	52.80	**61.20**	**78.40**	77.60
	MedMCQA	34.81	**35.24**	40.66	**44.86**	42.65	**45.50**	58.64	**58.82**	50.20	**53.54**	**71.12**	70.58
	MedQA	29.69	**33.38**	38.26	**44.56**	46.27	**47.80**	**61.12**	59.21	**54.05**	53.22	**77.85**	77.51
	PubMedQA	47.80	**55.50**	**63.40**	55.30	**51.60**	41.75	**59.40**	49.20	**42.80**	32.20	73.00	**75.75**
	Anatomy	**43.70**	42.04	49.63	**52.96**	56.30	**56.67**	68.89	**69.82**	64.44	**68.15**	77.04	**77.78**
	Clinical Knowledge	**41.51**	38.78	55.47	**62.17**	61.89	**62.26**	72.08	**73.68**	67.92	**74.90**	82.26	**83.40**
	College Biology	44.44	**47.05**	61.11	**68.06**	61.81	**64.93**	74.31	**77.43**	72.92	**77.95**	91.67	**92.36**
	College Medicine	36.99	**37.14**	50.29	**55.06**	57.80	**59.97**	67.05	**68.06**	63.58	**69.07**	**78.61**	78.18
	Medical Genetics	43.00	**44.75**	54.00	**66.00**	64.00	**65.50**	**80.00**	77.25	70.00	**78.00**	91.00	**91.00**
	Professional Medicine	29.78	**34.10**	50.37	**60.02**	56.99	**63.33**	**76.84**	75.64	72.06	**73.07**	**90.44**	90.26
	Average	36.69	**39.14**	49.76	**55.42**	54.65	**55.39**	**67.59**	67.05	61.08	**64.13**	81.14	**81.22**
French	FrenchMedMCQA	**29.91**	28.43	29.60	**40.27**	**45.48**	44.32	41.74	**44.63**	55.14	**58.02**	63.24	**73.05**
	MedExpQA	19.20	**20.60**	26.40	**39.20**	40.80	**41.20**	**48.00**	43.60	50.40	**56.00**	**76.80**	74.00
	Anatomy	**35.56**	35.18	48.15	**49.63**	33.33	**39.45**	45.19	**47.41**	55.56	**59.63**	67.41	**68.52**
	Clinical Knowledge	32.45	**36.51**	50.94	**57.92**	**55.47**	53.02	**61.89**	61.13	65.66	**71.51**	78.87	**80.56**
	College Biology	33.33	**38.02**	46.53	**52.78**	**53.47**	49.65	57.64	**62.50**	67.36	**72.92**	86.81	**87.67**
	College Medicine	32.95	**35.84**	43.93	**47.98**	**51.45**	48.56	57.80	**59.40**	57.80	**63.44**	69.94	**74.71**
	Medical Genetics	35.00	**40.00**	50.00	**57.25**	47.00	**59.00**	66.00	**67.00**	71.00	**72.00**	**90.00**	89.50
	Professional Medicine	24.26	**28.95**	33.09	**42.00**	43.38	**43.84**	51.47	**55.51**	59.56	**64.15**	72.79	**73.34**
	Average	30.33	**32.94**	41.08	**48.38**	46.30	**47.38**	53.72	**55.15**	60.31	**64.71**	75.73	**77.67**
Spanish	HeadQA	33.77	**34.32**	48.21	**54.47**	53.79	**55.54**	59.66	**61.24**	64.77	**68.00**	81.44	**82.44**
	MedExpQA	21.60	**23.00**	32.80	**38.40**	**46.40**	40.40	40.00	**43.00**	**52.80**	52.40	73.60	**76.60**
	Anatomy	37.78	**39.08**	42.22	**51.11**	45.93	**49.63**	48.15	**52.96**	60.74	**62.22**	71.11	**74.44**
	Clinical Knowledge	37.74	**38.78**	53.96	**55.47**	54.34	**56.13**	58.49	**62.08**	**68.68**	68.40	78.49	**80.00**
	College Biology	29.17	**35.94**	48.61	**50.35**	55.56	**56.25**	54.86	**55.04**	66.67	**69.10**	**85.42**	84.20
	College Medicine	32.37	**34.39**	43.93	**48.84**	**54.34**	48.99	49.71	**54.05**	**59.54**	58.24	69.94	**72.97**
	Medical Genetics	32.00	**34.75**	46.00	**59.50**	53.00	**57.25**	**72.00**	68.00	**67.00**	66.75	86.00	**86.75**
	Professional Medicine	26.47	**30.06**	38.24	**43.56**	**47.06**	45.68	**51.84**	50.74	53.68	**56.90**	**69.49**	68.94
	Average	31.36	**33.79**	44.25	**50.21**	**51.30**	51.23	54.34	**55.89**	61.74	**62.75**	76.94	**78.29**

Table 2: **MedPodGPT's performance on multilingual medical QA benchmarks.** All the models were fine-tuned with English, French, and Spanish medical podcast data and evaluated on various medical QA benchmarks in three in-domain languages. Benchmarks included MedExpQA, MedMCQA, MedQA, PubMedQA, HeadQA, and MMLU medical and clinical topics (covering anatomy, clinical knowledge, college biology, college medicine, medical genetics, and professional medicine). The baseline model's performance was compared with our MedPodGPT (indicated as *Ours*). The superior performances of MedPodGPT highlight the effectiveness of incorporating podcast data into the training process. The numbers in bold font indicate the best-performing model in each category.

Language	Benchmark Datasets	Model											
		Gemma 2B		Gemma 7B		Mistral 7B		LLaMA 3 8B		Mixtral MoE		LLaMA 3 70B	
		Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*	Baseline	*Ours*
Chinese	MedQA-MCMLE	33.39	**33.43**	40.51	**45.98**	39.67	**39.25**	63.63	**66.32**	45.80	**47.14**	**84.68**	83.73
	Anatomy	**28.38**	23.98	25.00	**31.59**	25.00	**30.41**	33.78	**35.98**	**33.11**	26.35	63.51	**64.02**
	Clinical Knowledge	**29.11**	28.27	31.22	**39.87**	33.33	**32.60**	49.37	**50.95**	**39.24**	38.61	71.73	**71.94**
	College Medicine	28.94	**32.23**	33.70	**36.08**	**30.77**	30.68	52.01	**56.50**	38.46	**40.94**	75.82	**80.49**
	Medical Genetics	32.39	**32.39**	43.75	**45.17**	38.64	**42.33**	43.18	**44.60**	45.45	**45.88**	**61.36**	57.53
	Medical Nutrition	33.79	**35.69**	40.69	**44.66**	**42.07**	37.24	**53.10**	50.00	49.66	**51.90**	66.21	**68.28**
	Traditional Chinese Medicine	27.57	**28.52**	31.35	**36.35**	24.86	**28.52**	**43.24**	39.46	30.27	**30.94**	66.49	**67.98**
	Virology	**37.28**	36.98	46.15	**54.44**	43.79	**48.22**	**59.76**	58.88	**53.25**	50.15	76.33	**77.51**
	Average	31.36	**31.44**	36.55	**41.77**	34.77	**36.16**	49.76	**50.34**	**41.91**	41.49	70.77	**71.44**
Hindi	Anatomy	25.93	**32.22**	34.07	**36.86**	23.70	**30.00**	**40.00**	35.18	31.11	**34.44**	52.59	**57.78**
	Clinical Knowledge	26.42	**28.96**	**41.89**	41.04	24.91	**35.85**	**48.30**	46.70	**38.11**	36.70	63.40	**69.06**
	College Biology	26.39	**33.16**	26.39	**34.03**	19.44	**28.47**	32.65	**37.16**	**30.56**	32.81	58.33	**68.06**
	College Medicine	24.86	**27.60**	42.20	**43.35**	23.12	**33.09**	41.04	**43.64**	27.17	**33.24**	60.69	**64.74**
	Medical Genetics	**31.00**	30.50	36.00	**41.75**	28.00	**29.25**	**46.00**	45.75	40.00	**43.25**	71.00	**77.00**
	Professional Medicine	25.37	**26.19**	30.88	**41.08**	22.06	**28.67**	36.40	**39.34**	29.41	**29.50**	45.59	**64.70**
	Average	26.66	**29.77**	35.24	**39.69**	23.54	**30.89**	40.73	**41.29**	32.73	**34.99**	58.60	**66.89**

Table 3: **MedPodGPT's zero-shot performance on non-English medical QA benchmarks.** All models were fine-tuned using English, French, and Spanish medical podcast data and assessed on cross-lingual medical QA benchmarks, including Mandarin and Hindi. Benchmarks included MedQA-MCMLE and multiple categories within MMLU and CMMLU medical and clinical topics, covering anatomy, clinical knowledge, college medicine, medical genetics, medical nutrition, traditional Chinese medicine, virology, and professional medicine. The baseline model's performance was compared with the performance of our model, MedPodGPT (indicated as *Ours*). Model performances are displayed to demonstrate the effectiveness of integrating podcast data into the training process. The numbers in bold font indicate the better-performing model in each category.