Traumatology and Orthopaedics of Kazakhstan
ISSN - e2789-9640

Original article

Cross-Lingual Adaptation of Medical LLM Models for Orthopedic Contexts: Comparative Results on Kazakhstani Clinical Data and the KazMMLU Benchmark

Zhanel Baigarayeva [1*], Assiya Boltaboyeva [2], Baglan Imanbek [3], Kassymbek Ozhikenov [4], Daulet Baiymbetov [5]

Received: February 11 2026
Revised: March 23 2026
Accepted: April 04 2026
Published: April 30 2026

Citation: Zhanel Baigarayeva, Assiya Boltaboyeva, Baglan Imanbek, Kassymbek Ozhikenov, Daulet Baiymbetov. Cross-Lingual Adaptation of Medical LLM Models for Orthopedic Contexts: Comparative Results on Kazakhstani Clinical Data and the KazMMLU Benchmark. Trauma & Ortho Kaz, 2026, 77 (2), jto046. https://doi.org/10.52889/1684-9280-2026-77-2-jto046

[1] Master of Engineering Sciences, K. I. Satbayev Kazakh National Technical Research University; Faculty of Information Technology, Al-Farabi Kazakh National University, Almaty, Kazakhstan. E-mail: zhanel.baigarayeva@gmail.com
[2] PhD student, K. I. Satbayev Kazakh National Technical Research University; Faculty of Information Technology, Al-Farabi Kazakh National University, Almaty, Kazakhstan. E-mail: boltaboyeva_assiya3@live.kaznu
[3] Associate Professor, Faculty of Information Technology, Al-Farabi Kazakh National University, Almaty, Kazakhstan. E-mail: imanbek.baglan18.06@gmail.com
[4] Head of the Robotics and Technical Means of Automation Department, Institute of Automation and Information Technologies, K. I. Satbayev Kazakh National Technical Research University, Almaty, Kazakhstan. E-mail: k.ozhikenov@satbayev.university
[5] Senior Lecturer, Department of Software Engineering, Institute of Automation and Information Technologies, K. I. Satbayev Kazakh National Technical Research University, Almaty, Kazakhstan. E-mail: d.baimbetov@satbayev.university

*** Corresponding author:** zhanel.baigarayeva@gmail.com

Abstract

Musculoskeletal disorders and injuries, including bone fractures, degenerative joint disease, ligament and meniscus injuries, and postoperative states after arthroplasty or osteosynthesis, often require prolonged treatment and staged rehabilitation. In trauma and orthopedic practice, clinical decision making depends heavily on narrative documentation, yet mixed Kazakh and Russian writing, highly variable terminology, and extensive free text complicate consistent clinical coding, outcome analytics, and the preparation of standardized discharge summaries and rehabilitation recommendations.
The aim of this study was to determine whether cross-lingual domain adaptation of pre-trained medical transformer models using Kazakhstan-specific trauma and orthopedic clinical narratives improves bilingual medical understanding.
Methods. We conducted a retrospective study using records of five hundred adult patients treated in the Orthopedic Surgery Department of Almaty City Clinical Hospital No.4, Kazakhstan. Multi-page electronic case histories stored in portable document format were de-identified, converted into text, and then transformed into bilingual instruction-style dialogue examples designed to reflect real clinical documentation patterns and musculoskeletal disease terminology. Two pre-trained medical transformer backbones were adapted using a parameter-efficient low-rank adaptation procedure: a compact healthcare-optimized model and a larger biomedical model. Performance was evaluated on the medicine subset “Professional and University, Russian language” of the Kazakh Massive Multitask Language Understanding benchmark, using accuracy as the primary outcome and the macro-averaged harmonic mean of precision and recall, balanced accuracy, and the Matthews correlation coefficient as secondary outcomes.

Ninety-five percent confidence intervals for accuracy and the macro-averaged harmonic mean of precision and recall were estimated using one thousand bootstrap resamples.
Results. After domain adaptation, the compact medical model achieved an accuracy of 33.00% (95% confidence interval - 27.95 to 38.72), compared with 20.88% (95% confidence interval - 16.50 to 25.59) before adaptation; the macro-averaged harmonic mean of precision and recall increased from 18.64% to 26.92%, balanced accuracy increased from 21.01% to 33.34%, and the Matthews correlation coefficient increased from 0.105 to 0.170. The larger biomedical model changed minimally, with accuracy increasing from 28.96% to 29.63%. A general-purpose multilingual baseline model achieved 30.64% accuracy without clinical domain adaptation.
Conclusions. These findings show that cross-lingual domain adaptation on a limited Kazakhstan-specific trauma and orthopedic corpus yields measurable gains, particularly for compact instruction-following medical models, and may support future tools for standardizing orthopedic documentation and accelerating rehabilitation planning. However, benchmark performance remains below levels required for high-responsibility clinical workflows, and further progress will require larger multi-center datasets, validation on practical documentation tasks such as structured extraction and discharge summary drafting, and dedicated evaluation of safety, privacy, and clinical risk.

Keywords: musculoskeletal diseases, fractures, bone, osteoarthritis, rehabilitation, medical records systems, computerized, natural language processing, machine learning, multilingualism.

1. Introduction

Musculoskeletal injuries and disorders are among the leading causes of disability and healthcare use in trauma and orthopedic practice. Fractures, degenerative disease of the hip and knee, spinal conditions, and postoperative care following arthroplasty or osteosynthesis often involve prolonged treatment pathways with staged rehabilitation and repeated follow-up. In these workflows, timely and high-quality decisions depend on the completeness and internal consistency of narrative documentation—history taking, physical examination findings, radiology reports, operative notes, discharge summaries, and longitudinal progress notes [1,2]. At the same time, large language models have shown promise for supporting clinical documentation and facilitating access to medical knowledge; however, clinically responsible use requires domain-specific adaptation, rigorous evaluation, and explicit attention to safety and privacy constraints [3–5].

Despite this progress, orthopedic documentation in routine care remains difficult to standardize, particularly in multilingual health systems. In Kazakhstan, clinical narratives commonly include heterogeneous terminology, frequent abbreviations, mixed Kazakh–Russian language use, and free-text diagnostic formulations. Such variability complicates standardized coding, registry development, outcome analysis, and multi-center data harmonization. Although general-purpose language models can capture broad meaning, they often struggle to reliably normalize orthopedic-specific concepts—such as fracture localization and displacement patterns, fixation techniques, implant nomenclature, and rehabilitation staging—when these are expressed in bilingual clinical text. In addition, most medical language models have been developed and validated primarily in English-dominant settings, leaving multilingual medical modeling and cross-lingual domain adaptation insufficiently studied for local clinical documentation needs [6,7]. Furthermore, benchmark-based evaluation may not fully reflect clinical readiness, and recent research has emphasized limitations of common assessment protocols for generative systems [8,9].

To address these gaps, this study investigates cross-lingual domain adaptation of pre-trained medical transformer models for orthopedic clinical narratives and evaluates their usefulness for information extraction and documentation standardization across core scenarios, including fractures, osteoarthritis, and postoperative follow-up after arthroplasty or osteosynthesis.

https://doi.org/10.52889/1684-9280-2026-77-2-jto046

2. Materials and methods

We constructed a dataset of 500 orthopedic surgery patients from City Clinical Hospital No. 4 (Almaty, Kazakhstan) by extracting text from long portable document format case records and converting them into structured bilingual, instruction-style dialogues. The models were adapted using a parameter-efficient low-rank fine-tuning approach [10,11], and performance was assessed on the medicine subset of the Kazakh Massive Multitask Language Understanding benchmark [12].

This study was designed as a retrospective observational analysis of clinical documentation from trauma and orthopedic care. The focus was on musculoskeletal conditions and injuries, including bone fractures, degenerative joint disease of the hip and knee, ligament and meniscus injuries, and postoperative follow-up after arthroplasty or osteosynthesis. The end-to-end workflow (Figure 1) includes data acquisition, preprocessing, model adaptation, and evaluation with deployment-oriented considerations. A dedicated pipeline for converting long portable document format case histories into structured training data is illustrated in Figure 2.

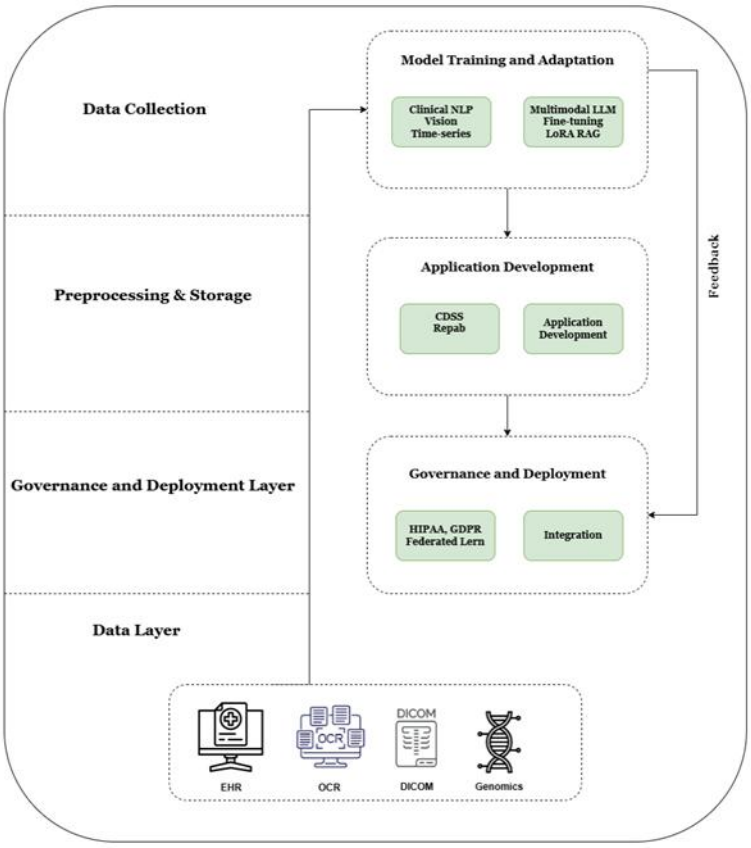

Figure 1 - End-to-end workflow for developing a clinical multimodal artificial intelligence system for musculoskeletal disorders (data integration, preprocessing, model adaptation, application prototyping, and governance/deployment)

Participants, eligibility criteria, and study material

The dataset consisted of records from 500 patients treated in the Orthopedic Surgery Department of City Clinical Hospital No. 4 (Almaty, Kazakhstan) (365 men and 135 women; mean age 42.3 ± 16.1 years; range 18–87 years). Source materials were stored as large portable document format files (120–170 pages per patient) and included admission notes, inpatient progress notes, operative reports, discharge summaries, and follow-up documentation.

Inclusion criteria were: (i) age 18 years or older; (ii) a primary musculoskeletal condition, such as bone fracture, osteoarthritis of the hip or knee, ligament or meniscus injury, or postoperative follow-up after arthroplasty or osteosynthesis; and (iii) sufficient narrative documentation describing diagnosis and clinical course.

Exclusion criteria were: (i) primary diagnosis not related to the musculoskeletal system; (ii) severely incomplete or internally inconsistent documentation,

including missing key diagnostic or procedural sections; (iii) poor-quality scans that remained unreadable after preprocessing; and (iv) records that could not be reliably de-identified according to the study protocol.

Data preprocessing and extraction of orthopedic clinical fields

Unstructured portable document format records were converted into analyzable text using Python and the PyMuPDF (fitz) library, with document layout preserved as much as possible. Rule-based parsing and regular expressions were applied to extract musculoskeletal-specific content, including: anatomical site of injury or disease; fracture descriptors (localization, displacement, comminution, and related features when available); osteoarthritis descriptors and joint-related complaints; procedure type (arthroplasty or osteosynthesis) and implant names when explicitly recorded; preoperative and postoperative diagnoses; longitudinal follow-up notes; and rehabilitation recommendations and functional restrictions when present. Post-processing steps included whitespace normalization, removal of conversion artifacts, and filtering of duplicated segments. Extracted information was first organized in spreadsheet format, with one row per patient representing standardized clinical attributes.

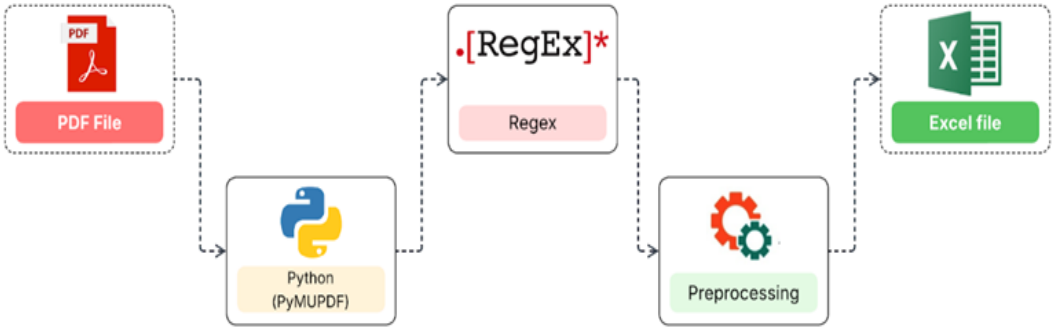

Figure 2 - Architecture for extracting and preprocessing clinical information from long portable document format case records into structured training examples

Dataset structuring for model adaptation

For supervised adaptation, the processed data were serialized into JSONL and formatted as instruction-style dialogues. The "user" message contained clinical context derived from the record, and the "assistant" message contained a structured clinical summary, diagnostic statement, and, where applicable, standardized terminology suggestions aligned with musculoskeletal conditions. The corpus was split into training, validation, and test sets using an 85:10:5 ratio while preserving the balance between Kazakh and Russian content. Tokenization was performed using an automatic tokenizer, and sequences were truncated to a maximum context length of 4096 tokens to ensure training compatibility.

Base models and domain adaptation strategy

Two medical transformer backbones were evaluated: Bio-Medical-LLaMA-3-8B, a biomedical-adapted derivative of a LLaMA-3 family model, and Med-Gemma-4B, a compact model optimized for healthcare tasks. Both models were adapted using low-rank adaptation, a parameter-efficient fine-tuning approach that introduces trainable low-rank updates into attention projections while keeping the original weights frozen [10,11]. Low-rank updates were applied to the attention projection modules (query, key, value, and output).

Low-Rank Adaptation fine-tuning and training configuration

Low-Rank Adaptation updates are defined as follows. The original weight matrix $W_0 \in R^{d\times r}$ remains frozen, and the trainable update is parameterized by two low-rank matrices $A \in R^{d\times r}$ and $B \in R^{d\times k}$ with $r \ll min\ (d,k)$.

$$\Delta W = \alpha AB,\ \mathrm{W} = \mathrm{W}_0 + \Delta\mathrm{W}$$

This approach substantially reduces the number of trainable parameters and lowers memory requirements, which is practical for clinical-domain adaptation under limited compute.

https://doi.org/10.52889/1684-9280-2026-77-2-jto046

Table 1 - Fine-tuning hyperparameters

Model	LoRA rank	Scaling factor	LoRA dropout	Learning rate	Training scheduler	Epochs	Mini-batch size	Global batch size	Precision format	Gradient clipping	Context length
Bio-Medical-LLaMA-3-8B	16	32	0.05	1×10^-4	warm-up 5%	10	2	16	bfloat16	1.0	4096
Med-Gemma-4B	16	32	0.05	2×10^-4	warm-up 5%	14	2	16	bfloat16	1.0	4096

The training pipeline was implemented in the Transformers ecosystem using parameter-efficient fine-tuning utilities and supervised fine-tuning tooling. Optimization employed an 8-bit paged AdamW optimizer. Training was performed on Runpod infrastructure using NVIDIA A100 SXM accelerators with eighty gigabytes of video memory, enabling full model loading and efficient completion of each fine-tuning run.

Evaluation protocol and statistical analysis

Model performance was assessed on KazMMLU, a large Kazakh–Russian bilingual extension of the Massive Multitask Language Understanding benchmark [12]. For clinical evaluation, we used the "Medicine (Professional and University, Russian)" subset. Predictions were generated in a zero-shot multiple-choice format: each prompt contained a question and five answer options (A through E), and the model was required to output exactly one option. Outputs with invalid formatting were counted as incorrect. The primary outcome was accuracy, defined as:

$$Accuracy = \frac{N_{correct}}{N_{total}} \times 100\%$$

Matthew's correlation coefficient computed from the confusion matrix. Uncertainty was quantified using ninety-five percent confidence intervals for accuracy and macro-averaged F1 score based on 1000 bootstrap iterations.

Ethics and data protection

The study was conducted in accordance with the core principles of biomedical ethics and the Declaration of Helsinki. All narrative records were de-identified prior to analysis, access to the dataset was restricted to authorized project members, and secure storage procedures were applied to minimize privacy risks. The study protocol was reviewed and approved by the local ethics committee of City Clinical Hospital No. 4 (Almaty, Kazakhstan).

3. Results

This study assessed which pre-trained medical large language model achieves the greatest performance improvement on the KazMMLU benchmark after fine-tuning on orthopedic department clinical documentation. Two openly available models were compared: Bio-Medical-LLaMA-3-8B and Med-Gemma-4B. Both models were adapted using a structured dataset extracted from patient medical records originally stored in portable document format. Records were preprocessed and converted into bilingual (Kazakh–Russian) dialogue-style data based on the ChatML format. Fine-tuning was performed using the Low-Rank Adaptation method on the RunPod.io platform with an NVIDIA A100 SXM (eighty gigabytes of video memory) accelerator.

Model performance was evaluated on the KazMMLU subset "Medicine (Professional and University, Russian)". Evaluation was conducted in LM Studio using quantized GGUF model variants. Each model was assigned a bilingual multiple-choice task, and predicted choices were compared against gold labels. The primary metric was accuracy. In addition, macro-averaged F1 score, balanced accuracy, and Matthew's correlation coefficient were computed. Accuracy represents the proportion of correct answers. Macro-averaged F1 score summarizes class-level stability. Balanced accuracy is the mean recall across classes. Matthew's correlation coefficient quantifies overall agreement between predictions and true labels. To quantify uncertainty, ninety-five percent confidence intervals for accuracy and macro-averaged F1 score were computed using one thousand bootstrap iterations.

https://doi.org/10.52889/1684-9280-2026-77-2-jto046

Table 2 - Performance of baseline and fine-tuned models on KazMMLU ("Medicine (Professional and University, Russian)")

Model	Type	Accuracy, 95% confidence interval	Macro-averaged F1, 95% confidence interval	Balanced accuracy	Matthews correlation coefficient
Med-Gemma 4B	Fine-Tuned	33.00% [27.95-38.72]	26.92% [22.84-35.45]	33.34%	0.170
Med-Gemma 4B	Baseline	20.88% [16.50-25.59]	18.64% [14.92-22.53]	21.01%	0.105
Bio-Medical-LLaMA 3 8B	Fine-Tuned	29.63% [23.90-34.01]	25.99% [21.09-30.76]	29.26%	0.119
Bio-Medical-LLaMA 3 8B	Baseline	28.96% [23.91-34.01]	23.49% [19.25-27.50]	29.25%	0.133
Meta LLaMA 3 8B	Baseline	30.64% [24.58-35.02]	23.20% [19.30-31.65]	30.24%	0.131
Gemma 3 4B	Baseline	28.28% [23.23-33.67]	23.65% [19.20-28.06]	28.12%	0.124

The results are summarized in Table 2 and Figures 3-5. For Med-Gemma-4B, accuracy increased from 20.88% [16.50-25.59] to 33.00% [27.95-38.72]. Macro-averaged F1 score increased from 18.64% to 26.92%, balanced accuracy increased from 21.01% to 33.34%, and Matthews correlation coefficient increased from 0.105 to 0.170. For Bio-Medical-LLaMA-3-8B, accuracy increased from 28.96% to 29.63%. A general-purpose baseline model (Meta LLaMA-3-8B) achieved 30.64% accuracy and 23.20% macro-averaged F1 score without domain adaptation.

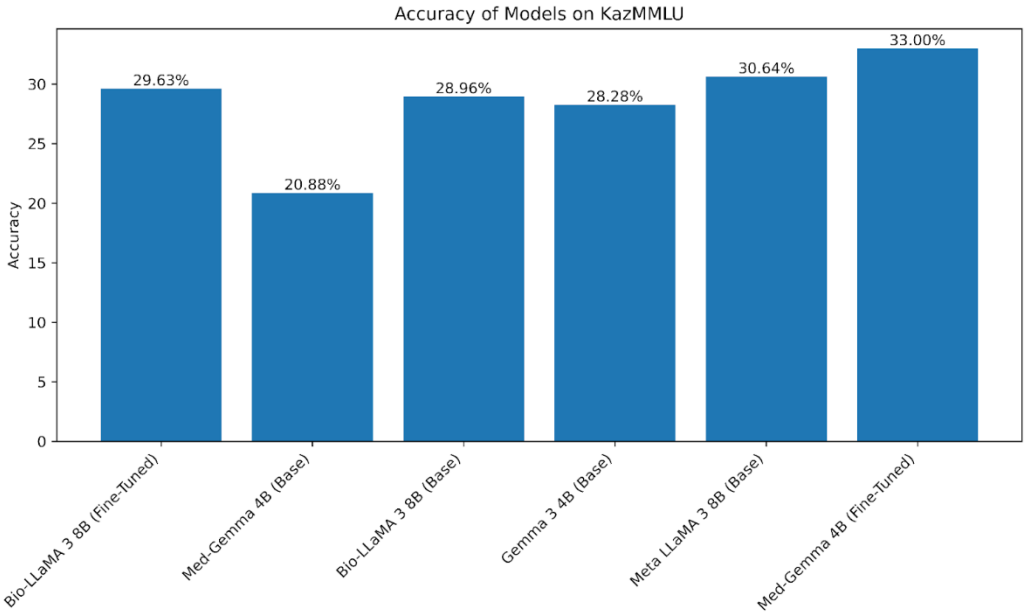

Figure 3 - Accuracy on KazMMLU with ninety-five percent bootstrap confidence intervals for baseline and fine-tuned models

https://doi.org/10.52889/1684-9280-2026-77-2-jto046

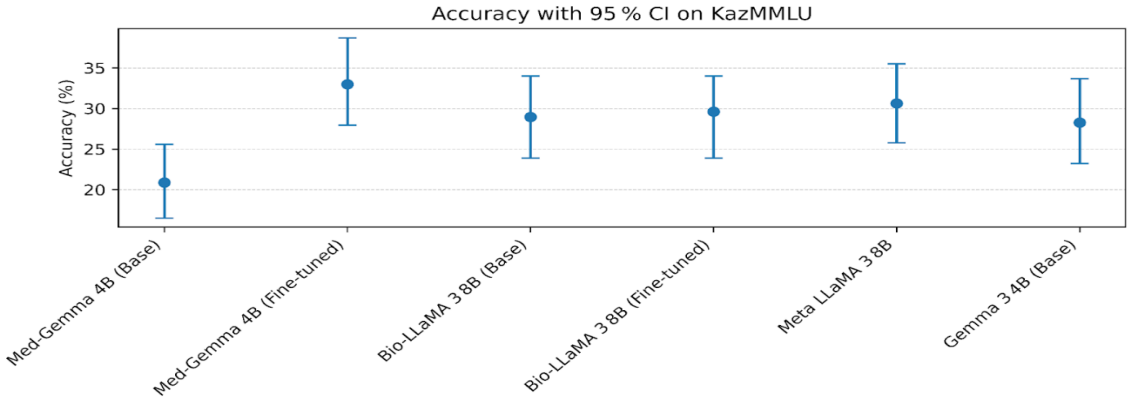

Figure 4 - Accuracy across all evaluated models on KazMMLU; error bars indicate ninety-five percent bootstrap confidence intervals based on one thousand resamples

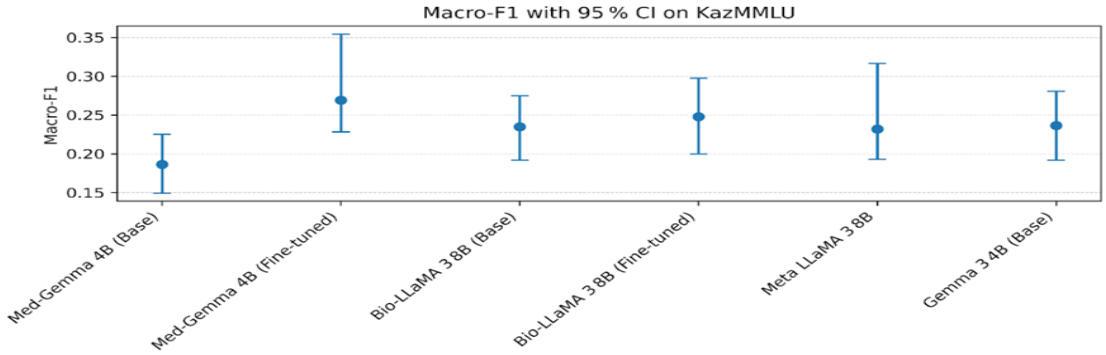

Figure 5 - Macro-averaged F1 score across all evaluated models on KazMMLU; error bars indicate ninety-five percent bootstrap confidence intervals based on one thousand resamples

4. Discussion

This study examined whether cross-lingual domain adaptation of pre-trained medical transformer models using bilingual orthopedic documentation improves benchmark performance in a Kazakh-Russian context. Using a parameter-efficient low-rank adaptation approach [10,11], the Med-Gemma-4B model increased accuracy on the medicine subset of the KazMMLU benchmark from 20.88% to 33.00%, with concurrent increases in macro-averaged F1 score, balanced accuracy, and Matthew's correlation coefficient, whereas the Bio-Medical-LLaMA-3-8B model changed only minimally.

The improvement observed for the compact instruction-oriented model is consistent with reports that smaller models can learn efficiently on clinical tasks under constrained data regimes, while larger backbones may require substantially more in-domain examples to yield measurable gains [15].

The markedly different gains across backbones also suggest that adaptation outcomes depend on the backbone's pretraining distribution and instruction alignment, as well as the cross-lingual distribution shift introduced by mixed Kazakh-Russian narratives. Related work emphasizes that multilingual medical language modeling remains necessary to support non-English clinical environments and that bilingual settings can expose gaps in narrowly specialized pretraining [6,7].

The competitive performance of a broadly multilingual general-purpose baseline on exam-style multiple-choice questions aligns with the idea that language coverage and diverse pretraining can compensate for limited specialty tuning in bilingual evaluation. At the same time, multiple-choice benchmarks capture only a subset of clinically relevant abilities, and recent surveys and critiques caution that benchmark scores may be insufficient proxies for practical performance, reliability, and safety in applied clinical workflows [8,9].

https://doi.org/10.52889/1684-9280-2026-77-2-jto046

Strengths of this work include the use of real-world orthopedic documentation from Kazakhstan, a reproducible pipeline for converting long case histories into bilingual instruction-style training examples, and uncertainty reporting with bootstrap confidence intervals alongside complementary metrics. Limitations include the single-center dataset, potential differences in documentation style across institutions and regions, and the restriction of evaluation to a multiple-choice benchmark. Further work should expand to multi-center corpora and evaluate task-specific outcomes such as information extraction for musculoskeletal diagnoses, discharge summary drafting, and rehabilitation recommendation consistency under clinician oversight.

5. Conclusions

This study evaluated which pre-trained medical transformer model benefits most from parameter-efficient domain adaptation on Kazakhstan trauma and orthopedic clinical documentation and how this adaptation affects performance on a Kazakh-Russian multiple-choice medicine benchmark. Fine-tuning the Med-Gemma-4B model using a low-rank adaptation approach increased accuracy from 20.88% (95% confidence interval 16.50-25.59) to 33.00% (95% confidence interval 27.95-38.72) and improved macro-averaged harmonic mean of precision and recall, balanced accuracy, and Matthews correlation coefficient, whereas the Bio-Medical-LLaMA-3-8B model showed minimal change. These findings support the qualified conclusion that, in a bilingual Kazakh-Russian orthopedic documentation setting with limited local data, compact instruction-oriented medical models can gain more from domain adaptation than larger biomedical models, while broad multilingual general-purpose pretraining can remain competitive on exam-style evaluation. Larger multi-center datasets and validation on practical clinical documentation tasks are required before use in high-responsibility clinical workflows.

Conflicts of Interest. The authors declare no conflicts of interest.

Funding. This study was funded by the Ministry of Science and Higher Education of the Republic of Kazakhstan under research grant BR24992820: "Innovative medical technologies and devices aimed at improving surgical interventions for prosthetics and rehabilitation in orthopedics and medical rehabilitation".

Author Contributions. Conceptualization – Z.B., B.I., K.O.; Methodology – Z.B., A.B., B.I.; Software – A.B., D.B.; Validation – Z.B., B.I.; Formal analysis – Z.B., A.B.; Investigation – Z.B., A.B., D.B.; Resources – K.O., D.B.; Data curation – A.B., D.B.; Writing – original draft – Z.B.; Writing – review and editing – Z.B., B.I., K.O., D.B.; Visualization – A.B., Z.B.; Supervision – B.I., K.O.; Project administration – K.O., B.I.; Funding acquisition – K.O.

All authors have read and agreed to the published version of the manuscript and have signed the copyright transfer form.

AI Declaration. The authors declare that no AI was used in the preparation of this manuscript.

References

1. Challa, S., Wu, H. H., Cunningham, B. P., & O'Toole, R. V. (2018). Orthopaedic trauma in the developing world: Where are the gaps in research and what can be done? Journal of Orthopaedic Trauma, 32(Suppl 1), S43–S46. https://doi.org/10.1097/BOT.0000000000001293

2. DeMaio, E. L., Marra, G., Suleiman, L. I., & Tjong, V. K. (2024). Global health inequities in orthopaedic care: Perspectives beyond the United States. Current Reviews in Musculoskeletal Medicine, 17, 439–448. https://doi.org/10.1007/s12178-024-09917-8

3. Omiye, J. A., Gui, H., Rezaei, S. J., Zou, J., & Daneshjou, R. (2024). Large language models in medicine: The potentials and pitfalls: A narrative review. Annals of Internal Medicine, 177(2), 210–220. https://doi.org/10.7326/M23-2772

4. Wang, D., & Zhang, S. (2024). Large language models in medical and healthcare fields: Applications, advances, and challenges. Artificial Intelligence Review, 57, 299. https://doi.org/10.1007/s10462-024-10921-0

5. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., Schaekermann, M., Wang, A., Amin, M., Collins, M., Brower, A., Lee, J., H, S., Li, Y., Rajpurkar, P., & Hinton, G. (2025). Toward expert-level medical question answering with large language models. Nature Medicine, 31, 943–950. https://doi.org/10.1038/s41591-024-03423-7

6. Qiu, P., Wu, C., Zhang, X., Zhang, Q., & others. (2024). Towards building multilingual language model for medicine. Nature Communications, 15, 8384. https://doi.org/10.1038/s41467-024-52417-z

7. Aracena, C., & Dunstan, J. (2023). Development of pre-trained language models for clinical natural language processing in Spanish. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (Student Research Workshop) (pp. 52–60). https://doi.org/10.18653/v1/2023.eacl-srw.5
8. Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., & others. (2024). A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology, 15, 39. https://doi.org/10.1145/3641289
9. McIntosh, T. R., Susnjak, T., Arachchilage, N. A. G., & others. (2025). Inadequacies of large language model benchmarks in the era of generative artificial intelligence. IEEE Transactions on Artificial Intelligence. https://doi.org/10.1109/TAI.2025.3569516
10. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. arXiv. https://doi.org/10.48550/arXiv.2106.09685
11. Han, Z., Gao, C., Liu, J., Zhang, H., & others. (2024). Parameter-efficient fine-tuning for large models: A comprehensive survey. arXiv. https://doi.org/10.48550/arXiv.2403.14608
12. Togmanov, M., Mukhituly, N., Turmakhan, D., Zhumabekov, A., & others. (2025). KazMMLU: Evaluating language models on Kazakh, Russian, and regional knowledge of Kazakhstan. arXiv. https://doi.org/10.48550/arXiv.2502.12829
13. Meta AI. (2025). Introducing Meta Llama 3: The most capable openly available large language model to date. Website. [Cited 16 Jul 2025]. Available from URL: https://ai.meta.com/blog/meta-llama-3/
14. Google. MedGemma: Health AI developer foundations. Website. [Cited 16 Jul 2025]. Available from URL: https://developers.google.com/health-ai-developer-foundations/medgemma
15. Taylor, N., Ghose, U., Rohanian, O., Young, T., & others. (2024). Efficiency at scale: Investigating the performance of diminutive language models in clinical tasks. Artificial Intelligence in Medicine, 157, 103002. https://doi.org/10.1016/j.artmed.2024.103002

Ортопедиялық контексттер үшін медициналық LLM-модельдерді кросс-лингвистикалық бейімдеу: Қазақтандық клиникалық деректері мен KazMMLU бенчмаркі негізіндегі салыстырмалы нәтижелер

Байғараева Ж.Е. [1], Болтабоева А.К. [2], Иманбек Б.Т. [3], Өжікенов Қ. [4], Баймбетов Д. [5]

[1] Техника ғылымдарының магистрі, Қ.И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті; Ақпараттық технологиялар факультеті, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан. E-mail: zhanel.baigarayeva@gmail.com

[2] PhD студент, Қ.И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті; Ақпараттық технологиялар факультеті, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан. E-mail: boltaboyeva_assiya3@live.kaznu

[3] Доцент, Ақпараттық технологиялар факультеті, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан. E-mail: imanbek.baglan18.06@gmail.com

[4] Робототехника және автоматика техникалық құралдары кафедрасының меңгерушісі, Автоматика және ақпараттық технологиялар институты, Қ.И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті, Алматы, Қазақстан. E-mail: k.ozhikenov@satbayev.university

[5] Аға оқытушы, Автоматика және ақпараттық технологиялар институты, Бағдарламалық инженерия кафедрасы, Қ.И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті, Алматы, Қазақстан. E-mail: d.baimbetov@satbayev.university

Түйіндеме

Тірек-қимыл жүйесінің аурулары мен жарақаттары, соның ішінде сүйек сынықтары, буындардың дегенеративті өзгерістері, байлам-меникс зақымданулары, сондай-ақ эндопротездеу және остеосинтезден кейінгі жағдайлар ұзақ емдеуді және кезең-кезеңімен оңалтуды талап етеді. Травматология және ортопедияда клиникалық шешім қабылдау көбіне мәтіндік құжаттамадағы мәліметтердің толықтығы мен бірізділігіне тәуелді. Алайда қазақ және орыс тілдерінің аралас қолданылуы, терминдердің көпнұсқалығы және еркін мәтіннің басымдығы диагноздарды кодтау, нәтижелерді талдау және стандартталған эпикриз бен реабилитациялық ұсынымдарды дайындауды күрделендіреді.

Зерттеудің мақсаты Қазақстанға тән травматологиялық-ортопедиялық клиникалық нарративтер негізінде алдын ала оқытылған медициналық трансформер модельдерін кросс-лингвистикалық домендік бейімдеудің екітілді медициналық тұжырымдарды түсінуді жақсарту мүмкіндігін бағалау болды.

Әдістері. Ретроспективті деректер жиынтығы ретінде Алматы қаласындағы №4 қалалық клиникалық ауруханаңың ортопедиялық хирургия бөлімінде ем алған 500 ересек науқастың көпбетті электронды ауру

тарихтары пайдаланылды. Тасымалданатын құжат пішіміндегі жазбалар де-идентификацияланып, мәтінге айналдырылды және клиникалық құжаттама стилін бейнелейтін екітілді (қазақ-орыс) нұсқаулық-диалогтық мысалдар түрінде құрылымдандырылды. Денсаулық сақтау саласының міндеттеріне оңтайландырылған ықшам медициналық модель және биомедициналық көлемдірек модель төмен дәрежедегі бейімдеу арқылы параметрлік тиімді қосымша оқытудан өткізілді. Бағалау Kazakh Massive Multitask Language Understanding бенчмаркіндегі «Медицина (кәсіби және университет деңгейі, орыс тілі)» ішкі жиынтығында жүргізілді; негізгі метрика ретінде дәлдік, ал қосымша метрикалар ретінде макро-орташаланған дәлдік пен толықтықтың гармониялық орташа мәні, теңдестірілген дәлдік және Мэтьюс корреляция коэффициенті есептелді. Негізгі көрсеткіштер үшін тоқсан 5 пайыздық сенімділік интервалдары 1 мың бутстреп қайта үлгілеу арқылы алынды.

Нәтижесі. Домендік бейімдеуден кейін ықшам медициналық модельдің дәлдігі 20,88 пайыздан (95% сенімділік интервалы - 16,50–25,59) 33,00 пайызға (95% сенімділік интервалы - 27,95–38,72) дейін өсті; макро-орташаланған гармониялық орташа мән 18,64 пайыздан 26,92 пайызға, теңдестірілген дәлдік 21,01 пайыздан 33,34 пайызға, Мэтьюс корреляция коэффициенті 0,105-тен 0,170-ке дейін артты. Биомедициналық көлемдірек модельде өзгеріс өте аз болды (дәлдік 28,96 пайыздан 29,63 пайызға). Клиникалық домендік бейімдеусіз жалпы мақсаттағы көптілді базалық модель 30,64 пайыз дәлдік көрсетті.

Қорытынды. Нәтижелер Қазақстанға тән травматология және ортопедия құжаттамасының шектеулі корпусында кросс-лингвистикалық домендік бейімдеу өлшенетін жақсартуға жеткізетінін, әсіресе ықшам нұсқаулыққа бағытталған медициналық модельдер үшін пайдалы екенін көрсетеді және құжаттаманы стандарттау мен реабилитациялық ұсынымдарды дайындауға арналған болашақ құралдарға негіз бола алады. Сонымен бірге, жоғары жауапкершілікті клиникалық қолдану үшін көпорталықты деректерді кеңейту, нақты құжаттау міндеттері бойынша валидация және қауіпсіздік/құпиялылықты бөлек бағалау қажет.

Түйін сөздер: тірек-қимыл жүйесінің аурулары, сүйек сынықтары, остеоартрит, оңалту, компьютерлендірілген медициналық жазба жүйелері, табиғи тілді өңдеу, машиналық оқыту, көптілділік.

Кросс-лингвистическая адаптация медицинских LLM-моделей для ортопедических контекстов: Сравнительные результаты на казахстанских клинических данных и бенчмарке KazMMLU

Байгараева Ж. [1], Болтабоева А. [2], Иманбек Б. [3], Ожикенов К. [4], Баймбетов Д. [5]

[1] Магистр технических наук, Казахский национальный технический исследовательский университет имени К. И. Сатпаева; Факультет информационных технологий, Казахский национальный университет имени Аль-Фараби, Алматы, Казахстан.
E-mail: zhanel.baigarayeva@gmail.com

[2] PhD студент, Казахский национальный технический исследовательский университет имени К. И. Сатпаева; Факультет информационных технологий, Казахский национальный университет имени Аль-Фараби, Алматы, Казахстан. E-mail: boltaboyeva_assiya3@live.kaznu

[3] Доцент, Факультет информационных технологий, Казахский национальный университет имени Аль-Фараби, Алматы, Казахстан. E-mail: imanbek.baglan18.06@gmail.com

[4] Заведующий кафедрой «Технические средства робототехники и автоматики», Институт автоматики и информационных технологий, Казахский национальный технический исследовательский университет имени К.И. Сатпаева, Алматы, Казахстан.
E-mail: k.ozhikenov@satbayev.university

[5] Старший преподаватель, кафедра программной инженерии, Институт автоматики и информационных технологий, Казахский национальный исследовательский технический университет имени К.И. Сатпаева, Алматы, Казахстан.
E-mail: d.baimbetov@satbayev.university

Резюме

Заболевания и травмы костно-мышечной системы, включая переломы костей, дегенеративные изменения суставов, повреждения связок и менисков, а также состояния после эндопротезирования и остеосинтеза, часто приводят к длительному лечению и поэтапной реабилитации. В травматологии и ортопедии значительная часть клинической информации фиксируется в текстовых документах, однако смешение русского и казахского языков, вариативность терминов и обилие свободного текста затрудняют единообразное кодирование, анализ исходов и подготовку стандартизированных выписных и реабилитационных рекомендаций.

Целью настоящего исследования было оценить, улучшает ли кросс-лингвистическая доменная адаптация предварительно обученных медицинских трансформерных моделей понимание двуязычных

медицинских формулировок на материале клинических нарративов травматолого-ортопедического профиля из Казахстана.

Методы. Проведено ретроспективное исследование, включившее 500 взрослых пациентов, пролеченных в отделении ортопедической хирургии Городской клинической больницы №4 (Алматы, Казахстан). Многостраничные электронные истории болезни, сохраненные в формате переносимого документа, были деидентифицированы, преобразованы в текст и далее структурированы в виде двуязычных (казахский и русский) инструкционно-диалоговых примеров, отражающих типичный стиль клинической документации. Две предварительно обученные медицинские трансформерные модели (компактная модель, оптимизированная для задач здравоохранения, и более крупная биомедицинская модель) были дообучены параметрически эффективной процедурой низкоранговой адаптации. Качество оценивалось на подмножестве «Медицина (профессиональный и университетский уровни, русский язык)» бенчмарка Kazakh Massive Multitask Language Understanding. Основным показателем была точность, дополнительными показателями — макро-усредненное гармоническое среднее точности и полноты, сбалансированная точность и коэффициент корреляции Мэтьюса. 95% доверительных интервалов для ключевых показателей рассчитывались методом бутстреп-перевыборки с одной тысячей итераций.

Результаты. После доменной адаптации компактная медицинская модель повысила точность с 20,88% (95% доверительный интервал – 16,50–25,59) до 33,00 процента (95% доверительный интервал – 27,95–38,72); макро-усредненное гармоническое среднее точности и полноты увеличилось с 18,64% до 26,92%, сбалансированная точность — с 21,01% до 33,34%, коэффициент корреляции Мэтьюса — с 0,105 до 0,170. Более крупная биомедицинская модель изменилась минимально (28,96% до 29,63% точности). Универсальная многоязычная базовая модель без клинической доменной адаптации показала точность 30,64%.

Выводы. Полученные данные свидетельствуют, что кросс-лингвистическая доменная адаптация на ограниченном корпусе травматолого-ортопедической документации из Казахстана дает измеримый прирост качества, особенно для компактных инструкционно-ориентированных медицинских моделей, и может быть использована как основа для инструментов стандартизации документации и подготовки реабилитационных рекомендаций. При этом, для применения в клинических процессах высокой ответственности необходимы расширение данных на несколько центров, валидация на прикладных задачах документирования, а также отдельная оценка безопасности и конфиденциальности.

Ключевые слова: болезни костно-мышечной системы, переломы костей, остеоартрит, реабилитация, компьютеризированные системы медицинских записей, обработка естественного языка, машинное обучение, многоязычие.

https://doi.org/10.52889/1684-9280-2026-77-2-jto046