**TranscribePro: Enhancing linguistic analysis through mobile technology**

*Victoria Firsanova, Saint Petersburg State University, 7/9 Universitetskaya Emb., Saint Petersburg 199034, Russia*

In recent years, numerous large language models (LLMs), such as Gemma (Mesnard, Hardin, Dadashi et al. 2024), GPT-4 (Achiam, Adler, Agarwal et al. 2023), or Phi-3 (Abdin, Jacobs, Awan et al. 2024), demonstrate their capacity on a wide range of tasks from the field of theoretical linguistics. For example, LLMs proved their efficiency in semantic role labeling (Li, Kazeminejad, Brown et al. 2023), syntax tree generation (Altıntaş, Tantuğ 2023), and morphological segmentation (Pranjić, Šikonja, Pollak 2024). However, applying LLMs in academic research might be challenging.

Firstly, despite the accelerating development of LLM user interfaces, such as ChatGPT, most artificial intelligence models are not designed for non-technical users. Utilizing state-of-the-art neural network models requires a high level of computational expertise, which makes them inaccessible for many experts in Arts and Humanities. Secondly, the computational resources needed to run LLMs can be substantial. The deployment of LLMs is challenging on low-resource hardware commonly found in academic settings. Thirdly, the LLM development is typically focused on large-scale industrial applications rather than on specialized linguistic research. Consequently, LLMs lack support for essential linguistic tools and standards, such as the International Phonetic Alphabet (IPA), which is crucial for detailed phonetic and phonological analysis.

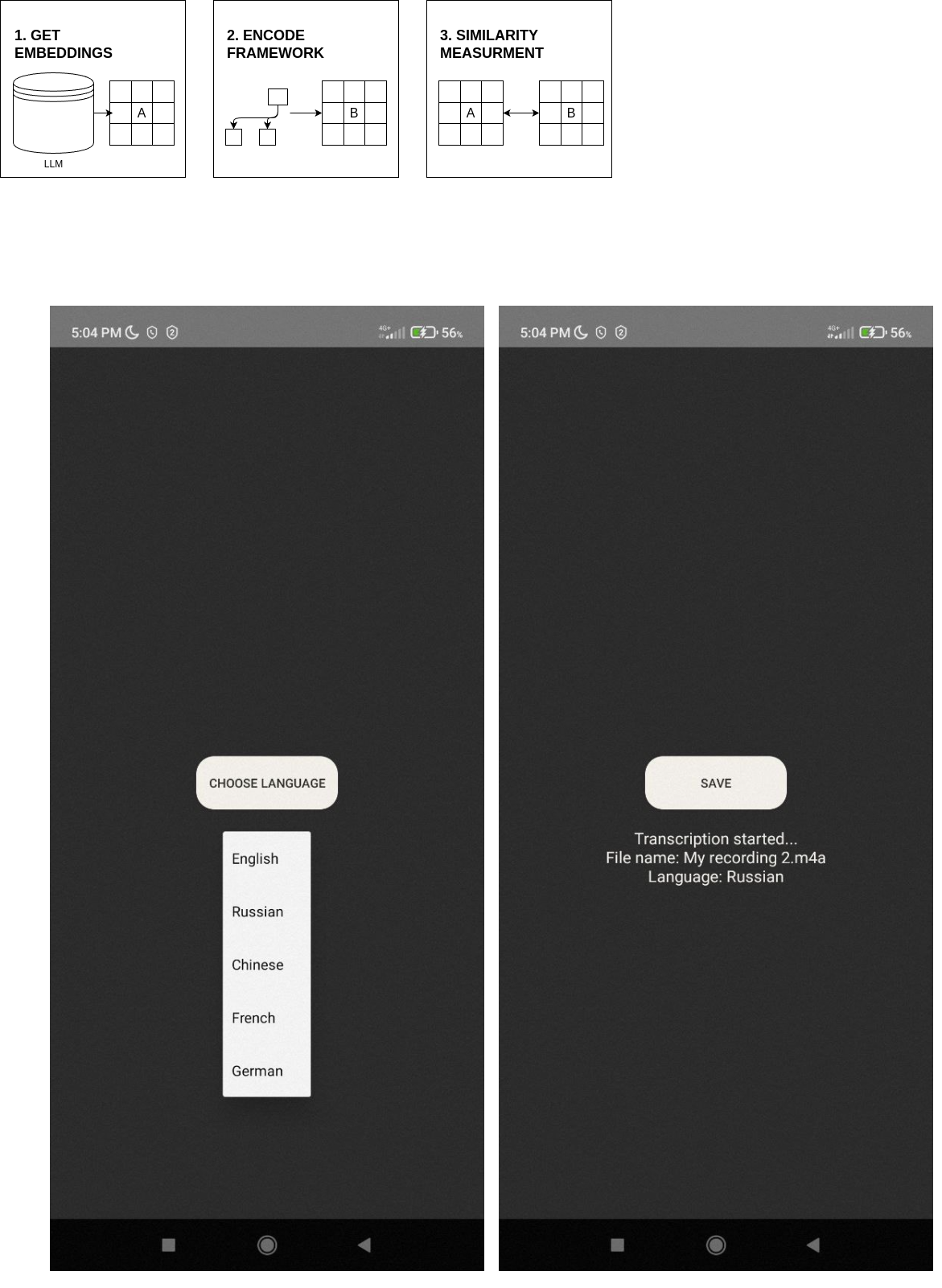
This paper introduces a novel mobile application TranscribePro, designed to make LLMs accessible and useful for linguistic research. The application integrates large-scale state-of-the-art neural networks into a single, minimalistic, user-friendly interface, providing a powerful tool for phonetic and phonological analysis, as well as language education. The research focuses on exploring approaches to LLM compression, adapting existing models to the specific needs of the linguistic community, and making advanced LLM technology accessible for both theoretical linguistics and language education through efficient design.

Most LLM compression techniques are based on the intrinsic dimensionality concept (Aghajanyan, Gupta, Zettlemoyer 2021), demonstrating that significant training data features are represented by a relatively small number of the overall model parameters. Such techniques, as quantization (Gholami et al. 2022), palletization (Cho et al. 2021), pruning (Hoefler et al. 2021), knowledge distillation (Gou et al. 2021), and low-rank adaptation (Hu et al. 2021), are often used for LLM compression. This study implies exploring the advantages and limitations of each method in the context of mobile development.

Fig. 1 illustrates the TranscribePro application user interface. The project is open source, the code base is provided at GitHub (https://github.com/vifirsanova/TranscribePro). TranscribePro is an Android application that uses Kotlin programming language and XML layouts. The machine learning models are hosted on user device. The application allows for loading pre-recorded audio file supporting a wide range of file extensions. The application automatically creates a linguistic transcription file using a markup language, inspired by such annotation tools as ELAN and Praat.

The study focuses on exploring LLMs for foundational linguistic research. The research method implies training, fine-tuning and adapting various multilingual machine learning models for text-to-speech synthesis in the theoretical linguistics context. The study consists of technical and theoretical parts. The technical part of the study compares two different baselines: (1) converting OpenAI Whisper to TensorFlow Lite format and integrating the model to an application; (2) using C++ version of OpenAI Whisper and Java Native Interface to run quantized model on a mobile device. The theoretical part of the study implies the research and development of annotation for phonetic and phonological studies based on empirical material.

The technical part of the study compares inference speed, model size, and performance based on user feedback. The theoretical stage resulted in transcription markup development validated using empirical linguistic data from multiple languages. The study demonstrates the potential of applying LLMs in foundational linguistic research, highlighting the advantages and trade-offs of different deployment strategies. Both TensorFlow Lite and C++ models maintained high accuracy levels, suitable for linguistic research and language learning. User feedback indicated a preference for the C++ model, while the TensorFlow Lite model is recommended for its ease of integration. The annotation framework developed in the theoretical part of the study provided a robust tool for phonetic and phonological studies, contributing to the field of theoretical linguistics.

Fig. 1. TranscribePro application user interface example

References

Abdin, M., Jacobs, S.A., Awan, A.A., Aneja, J., Awadallah, A., Awadalla, H., Bach, N., Bahree, A., Bakhtiari, A., Behl, H. and Benhaim, A. Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219, 2024.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R. GPT-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

Aghajanyan A., Gupta S., Zettlemoyer L. Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021. P. 7319-7328.

Altıntaş M., Tantuğ A.C. Improving the performance of graph based dependency parsing by guiding bi-affine layer with augmented global and local features. Intelligent Systems with Applications 1 (18), 2023.

Cho M. et al. DKM: Differentiable k-means clustering layer for neural network compression. arXiv preprint arXiv:2108.12659, 2021.

Gholami A. et al. A survey of quantization methods for efficient neural network inference. Low-Power Computer Vision. Chapman and Hall/CRC, 2022. P. 291–326.

Gou J. et al. Knowledge distillation: A survey. International Journal of Computer Vision, 2021, 129 (6). P. 1789-1819.

Hoefler T. et al. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. Journal of Machine Learning Research, 2021, 22 (241). P. 1-124.

Hu E.J. et al. LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021.

Li T., Kazeminejad G., Brown S.W., Srikumar V., Palmer M. Learning Semantic Role Labeling from Compatible Label Sequences. In Findings of the Association for Computational Linguistics: EMNLP 2023, 2023. P. 15561-15572.

Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M.S., Love, J. and Tafti, P.. Gemma: Open models based on Gemini research and technology. arXiv preprint arXiv:2403.08295, 2024.

Pranjić M., Robnik-Šikonja M., Pollak S. LLMSegm: Surface-level Morphological Segmentation Using Large Language Model. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024. P. 10665-10674.