Building a Large Language Model for Moderately Resourced Language: A Case of Croatian

Gaurish Thakkar, Vanja Štefanec, Daša Farkaš, Marko Tadić
University of Zagreb, Faculty of Humanities and Social Sciences
Institute of Linguistics
Ivana Lučića 3, Zagreb, Croatia
{gthakkar, vstefane, dfarkas, marko.tadic}@ffzg.unizg.hr

Abstract. *Large Language Models (LLMs) have recently demonstrated significant advancements in processing natural language tasks. Consequently, the tide in the natural language processing (NLP) and language technologies (LT) turned towards the development of LLMs for different languages. As a significant number of monolingual and multilingual LLMs are being developed and deployed, it has become crucial to assess their capacities and performance. This document presents a summary of the current activities to develop a monolingual LLM for Croatian (HR-LLM) as a moderately resourced language with less than ten million speakers. The paper presents the relevant previous work, explains why the model is needed, which methodology was used, and how the HR-LLM was trained and how it will be evaluated for selected downstream tasks.*

Keywords. Large Language Models, LLMs, GPT, Evaluation, Croatian Language, Natural Language Processing, Neural Networks, Deep Learning.

1 Introduction

In recent years, in the natural language processing (NLP) and language technologies (LT) there has been significant progress in the development of language models (LMs), particularly large language models (LLMs). Today, many LLMs are trained on massive datasets of natural language text and they have the capability to perform a wide range of NLP/LT tasks at a state-of-the-art level. One of the recent key advances in LLMs was the development of transformer architecture (Vaswani et al., 2017). The transformer is a neural network architecture that is well-suited for modelling long-range dependencies in text, and this approach enabled the training of LLMs that can represent complex linguistic relationships and thus generate more fluent text. Transformer-based models (Kenton and Toutanova, 2019; Liu et al., 2019; Brown et al., 2020) have influenced several industries and fields by enabling task automation, improving efficiency, and expanding the possibilities for human-AI collaboration. Their ability to generate human-like text and take into account the co-text has made them a transforming force in the world of AI and technologies that are being built upon. Most of the LLMs created so far are multilingual, i.e., they were trained using texts in different languages, and the number of languages varies from 2 to 200+.

One inherent problem with multilingual LLMs is the unbalanced dataset used for training (Lai et al., 2023). The languages with large amounts of training data are usually well represented in the model, whereas languages with limited resources are constrained by the paucity of training data (Kahana et al., 2024). Studies (Martin et al., 2020; Papadimitriou et al., 2023; Virtanen et al., 2019), have shown that models trained in only a few languages (bilingual/trilingual LLMs) or in a single language (monolingual LLMs) perform better in a number of NLP tasks than models trained in many languages (multilingual LLMs). There are also recent attempts to distil the monolingual model from a larger multilingual LLM in order to come up with a smaller and more efficient monolingual LLM (Singh et al., 2023). That is the main reason for instigating this project.

We strongly believe that this case study can serve as an example for the building of a LLM for a moderately resourced language with less than 10 million speakers, or for a lower-resourced language with more speakers. The reason for this is that the overall production of e-text, digitalisation and global usage of mobile devices open the doors to the collecting and harvesting of sufficient language data even for languages that were until recently considered to stay beyond the digital divide, and all this even without the need to artificially generate additional training data, i.e., to use the data augmentation approaches. It seems like many of the initial pioneering multilingual LLMs will soon be in the position to be either replaced by monolingual LLMs or, more probably, fine-tuned with additional monolingual data for the majority of monolingual tasks. In this respect and from the perspective of digital language equality (see e.g., the project funded by the EU Parliament with a similar name[1]), the production of a LLM for a lan-

[1] https://european-language-equality.eu/

guage could be considered as one of the building blocks of the contemporary BLARK (Krauwer, 1998) since its existence ensures the level of development of LT for a language, which is expected today. We believe that the real digital language equality should incorporate LLM as a highly valuable type of language resource.

The main objective of this particular project, Croatian Extended Reality Extensions (HR-XR-XTEND)[2], a FSTP subproject of a Horizon Europe-funded project UTTER[3], is to build a large monolingual generative pre-trained transformer model for Croatian (HR-LLM). As Croatian has already been included in multilingual LLMs, it remains unclear whether the presence of other languages within the LLM influences their performance with Croatian, particularly across different tasks. Since at this point in the development of LLMs we still lack the standardised and balanced evaluation tools for measurement of multilingual LLMs' performance, one of the ways to establish their performance baseline is to construct a monolingual LLM. To establish that baseline performance, it is necessary to construct and evaluate a monolingual LLM. The secondary objective is to develop a set of evaluation benchmarks and methods that are specifically tailored to measure the performance of any LLM when it is applied to tasks involving Croatian language. The third objective of this project is to successfully implement the HR-LLM in the production process of the industrial partner.

We propose the creation of a LLM tailored to address the needs of lower-resourced languages, focusing on those with less than 6 million speakers (Language Report Croatian, 2022). The success of this project could serve as proof of a concept that a similar building process could be applicable to other languages with less than 10 million speakers. The research primarily focuses on training an LLM for a language like Croatian using text collections already gathered by the project beneficiary and other publicly available datasets that are mostly crawled. This could serve as an example for a number of languages that don't have a lot of language resources at disposal. Primarily, it could contribute to the research theme of deep natural language understanding (NLU) for lower-resourced languages. The paper is structured as follows: after the introduction section, the related work is described, objectives and methods are presented in Section 3, while experiments and evaluation are presented in Sections 4 and 5, respectively. The final remarks and HR-LLM availability are mentioned in the concluding section 6.

2 Related Work

The recently introduced transformer architecture was first introduced in the paper by Vaswani et al. (2017). The key advance has been the use of pre-training. In pre-training, an LLM is trained on a large and diverse dataset of text without any specific task in mind, thus becoming a multipurpose resource.

The state-of-the-art LLMs that are based on the transformer architecture are: BERT (Kenton and Toutanova, 2019) developed by Google AI. BERT model had 110 million parameters. BERT is pre-trained on a massive amount of text data, which enables it to come close to general language understanding. During pre-training, BERT is trained to predict missing words, also known as masked language modelling (MLM), within sentences. This unsupervised pre-training process allows BERT to learn rich contextual representations. The bidirectional aspect of BERT is crucial because it allows the model to "understand" the co-text of a word by considering both its left and right neighbours in a sentence, enabling it to capture more complex language patterns. RoBERTa (Liu et al., 2019) is a LLM architecture that is based on the BERT framework. RoBERTa was pretrained on a much larger corpus of text data compared to the original BERT model. While BERT used a static masking pattern during pretraining, RoBERTa introduced a dynamic masking approach and a sentence-boundary objective during pretraining, which helps the model recognise sentence boundaries more effectively. Unlike BERT, RoBERTa does not use the Next Sentence Prediction task during pretraining. RoBERTa-Base and RoBERTa-Large have 137 and 355 billion parameters, respectively. GPT-3 (Brown et al., 2020) is a Large Language Model (LLM) that utilises the transformer architecture. GPT-3 has 175 billion parameters, which is significantly more than any other LLM that was available at the time of its release. This large size allows GPT-3 to learn more complex linguistic relationships and generate more fluent and informative text. GPT-NeoX (Black et al., 2022) is a 20-billion-parameter autoregressive LLM developed by EleutherAI. It uses the transformer architecture and is pre-trained on a massive dataset of text and code. GPT-NeoX is available to the public under an open-source licence. Hungarian GPT-3 (Yang et al., 2023) was developed by the Hungarian Research Centre for Linguistics in collaboration with the University of Pécs. It was the first LLM trained on a Hungarian only dataset with a 6.7 billion-parameter GPT language model. The model was trained using the GPT-NeoX implementation on a dataset of 36.3 billion tokens. Currently, one of the largest models in the world is the GPT-4 (OpenAI et al., 2024), with more than 1 trillion parameters. It is a multimodal LLM. It is the fourth in its series of GPT foundation models. Another model created by OpenAI is ChatGPT, which is a multilingual LLM with code and creative text generation capabilities. Meta AI has also published the LLaMA Models (Touvron et al., 2023a,b) (Llama 1 and 2), which is a collection of foundational language models ranging from 7 billion to 70 billion parameters. Bloom (Workshop et al., 2022) is a massive, multilingual language model (MLM) with 176 billion parameters, trained on a massive dataset of text and code. It can generate text, translate languages, write

[2]https://hr-xr-xtend.ffzg.unizg.hr

[3]https://he-utter.eu

various kinds of creative content, and answer your questions in an informative way. The 7-billion parameter model Mixtral 7B (Jiang et al., 2023) employs an innovative mixture-of-experts (MoE) framework to facilitate training that is both scalable and efficient. By employing this methodology, Mixtral 7B is capable of attaining competitive performance while utilising a considerably reduced number of parameters in contrast to conventional dense models. As a result, it serves as a valuable standard against which other studies in the field are measured. GEITje 7B (Rijgersberg and Lucassen, 2023) is a large open Dutch LLM developed by further training Mistral 7B (Jiang et al., 2023) on no less than 10 billion tokens of Dutch text from the Dutch Gigacorpus and the MADLAD-400 web crawling corpus. Marinova et al. (2023) created BERT-WEB-BG and GPT-WEB-BG LLMs, which are Bulgarian language counterparts of BERT and GPT-2. The web-scraped dataset underwent a specialised method for source filtering, subject selection, and lexicon-based elimination of improper language during the pre-training phase.

Table 1: List of the most widely used multilingual LLMs and several similar monolingual LLMs

Name	Languages	Parameters
(m)BERT	En/multilingual	110M
RoBERTa	En	137B-355B
GPT-3	multilingual	175B
GPT-NeoX	En	20B
Hungarian GPT-3	Hu	6.7B
GPT-4	multilingual	1T
LLaMA 1-2	multilingual	7-70B
Bloom	multilingual	176B
Mixtral 7B	multilingual	7B
GEITje 7B	Nl	7B
GPT-WEB-BG	Bg	1.5B

In the reminder of this section we also list several LLMs that have been trained in Croatian. All of them are multilingual, and there is no monolingual model built exclusively on Croatian data. Apart of already presented widely used multilingual LLMs, here we mention other less used LLMs that were still trained with Croatian data. Language-agnostic BERT Sentence Encoder (LaBSE) (Feng et al., 2022) is a BERT-based model trained for sentence embedding in 109 languages. X-MOD (Pfeiffer et al., 2022) is a MLM trained on filtered Common Crawl data containing 81 languages. This model reuses the tokenizer of XLM-R. The model has been pre-trained with language-specific modular components (language adapters). CroSloEngual BERT (Ulčar and Robnik-Šikonja, 2020) is a trilingual model (110 million parameters) using BERT-base architecture, trained on Croatian, Slovenian, and English corpora. Focusing on three languages, the model performs better than multilingual BERT while still offering an option for cross-lingual knowledge transfer. BERTić (Ljubešić and Lauc, 2021) is a joint multilingual LLM of four distinct languages: Bosnian, Croatian, Montenegrin and Serbian. Cro-CoV-cseBERT (Babić et al., 2021) is a LLM based on the CroSloEngual BERT and has been further trained on a large corpus of texts related to COVID-19 in the Croatian language (Cro-CoV-Texts) which contains 186,738 news articles, 500,504 user comments related to COVID-19 published on Croatian online news portals, and 28,208 COVID-19 tweets in the Croatian language. Several additional language models (LLMs) have been developed under the InfoCoV project, such as Cro-CoV-BERTić, Senti-CoV-cseBERT, and Multi-Cro-CoV-cseBERT. These models are specifically trained on text data connected to COVID-19. The multilingual parliamentary model (XLM-R-parla) (Mochtak et al., 2023) is the XLM-R-large model, additionally pre-trained on texts of parliamentary proceedings. Texts for the additional pre-training, 1.7 billion words in size, come from the ParlaMint corpus and the EuroParl corpus. EUBERT (European Parliament EUBERT Hugging Face, 2023) is a pretrained BERT uncased model that has been trained on a vast corpus of documents in all 24 EU official languages and published by the European Publications Office. These documents span the last 30 years, providing a comprehensive dataset that encompasses a wide range of topics and domains. HR-RoBERTa (HuggingFace macedonizer hr-roberta-base, 2021) model is a LLM-pretrained model on Macedonian using a MLM objective and has been further trained on Croatian data. The HR-GPT2 (HuggingFace macedonizer hr gpt2, 2021) is a LLM pretrained on a large corpus of Croatian data in a self-supervised fashion to provide text generation capabilities. TwHIN-BERT (Zhang et al., 2022) is a multilingual Tweet language model (250 and 550 million parameters) that is trained on 7 billion Tweets from over 100 distinct languages. It not only outperforms similar models in tasks such as text classification, but also in social recommendation tasks such as predicting user-to-Tweet engagement.

3 Objectives and Methods

The project goals were set to:

1. collect at least six billion tokens of Croatian texts and prepare that data for HR-LLM training;
2. train the LLM for the Croatian language using monolingual data only;
3. evaluate the HR-LLM for downstream tasks.

Our objective was to develop a monolingual model specifically designed for the Croatian language and based on monolingual data only. Croatian as a moderately resourced language has so far appeared only in multilingual LLMs (e.g., mBERT, BERTić, CroSloEngual) and it is not clear whether other languages contribute to improvement or degrading of LLMs' performance when using on Croatian texts alone. With the existence of a monolingual model, this baseline is set, and any increase or decrease in performance could be

Table 2: List of LLMs that include Croatian

Name	Language
(m)BERT	En/multilingual
RoBERTa	En
GPT-3	multilingual
GPT-4	multilingual
LLaMA 1-2	multilingual
LaBSE	multilingual
X-MOD	multilingual
CroSloEngual BERT	Hr, Sl, En
BERTić	Bo, Hr, Me, Sr
Cro-CoV-cseBERT	Hr, Sl, En
Cro-CoV-BERTić	Bo, Hr, Me, Sr
XLM-R-parla	multilingual
HR-RoBERTa	Mk, Hr
HR-GPT2	Mk, Hr
EUBERT	multilingual
TwHIN-BERT	multilingual

measured in comparison to that baseline. To build a monolingual LLM, a large number of tokens had to be collected from various sources: corpora and text collections from the existing repositories (although often under-performing for Croatian, e.g. results from the Oscar project[4]) and newly collected data (online and offline) that were not available in previous data collection campaigns. Table 3 lists the large repositories or data and collection campaigns that we used as sources of data.

The LLM type that was targeted stemmed from the European AI initiative OpenGPT-X[5] and recent achievements in building the first GPT-3 model for Hungarian (PULI GPT-3SX[6]). The OpenGPT-X at this moment features only five major European languages (English, French, German, Italian, and Spanish), with over 50 million speakers, while other European languages are not covered. On the other hand, the Hungarian GPT-3 LLM demonstrates that such an endeavour could be achieved for a language with approximately 13 million speakers (Language Report Hungarian, 2022).

We used the datatrove[7] library to perform the near deduplication with MinHashLSH and a threshold of 0.72. This was done following the advice that LLMs trained on deduplicated data are better and memorise less of their data (Lee et al., 2022). After the deduplication, where repeating paragraphs were removed from the training data, the total number of tokens used for training was downsized to 7.72 billion, compared to the original dataset which contains 8.9 billion tokens.

Our training arrangement was divided into three distinct steps:

1. We initialise the model and train it using the training configurations provided by the gpt-neox library. The following configurations were chosen as the starting candidates for training our model. This setup is designed to measure the performance of the model when trained exclusively on Croatian data from the beginning. We trained models with 160M, 350M, 410M, and 1.4B parameters respectively.

2. We conduct constant pre-training (CPT) on the gpt-2 model using the Croatian data outlined in the preceding section. This model validates its performance by leveraging existing knowledge of the English language.

3. We engage in ongoing pre-training (OPT) on the "unsloth/gemma-2-9b-bnb-4bit" model, which is a compressed version of Gemma 2. This is a multilingual language model that aids in the evaluation of Croatian CPT performance within a multilingual framework.

[4]https://oscar-project.org/
[5]https://www.aleph-alpha.com
[6]https://nytud.github.io
[7]https://github.com/huggingface/datatrove

Table 3: Non-exhaustive list of largest data sources used for training the HR-LLM (beta version) with approximate size in tokens. *Croatian texts only

Name	Approx Size
CLASSLA Hr Web corpus 1.0	2.5 billion
CC100-Hr Dataset	2.27 billion
Corpus of Croatian News Feeds	2.25 billion
Parallel data for En-Hr on OPUS Resources*,	1.48 billion
Hr-news from XLM-R-BERTić dataset	1.4 billion
news/legal corpus	175 million
Corpus of Croatian Academic Theses	312 million
ParaCrawl*	69.96 million
Riznica from XLM-R-BERTić dataset	69.51 million
MARCELL Croatian legislative subcorpus	56 million
CURLICAT Croatian corpus	49 million
MARCELL Croatian-English Parallel Corpus of Legislative Texts*	14.3 million
Romance-Croatian Parallel Corpus (literary works)	2.5 million
Total	8.9 billion

4 Experiments

The experimental phase was focused on developing and evaluating the HR-LLM architecture and training process. This involved:

1. Collecting and preprocessing of the training dataset (see section 3 above): This data will also be partially used within newly established initiatives such as Language Data Space[8] and ALT-EDIC[9].
2. Training the model: For training the model, we used the publicly available open source library at gpt-neox[10].
3. Local GPU infrastructure was utilised for experimentation, while EuroHPC and University of Zagreb Computing Centre supercomputers with GPU nodes were utilised for advanced experimenting and final version training.
4. Configuring the training hyperparameters: hyperparameters such as the learning rate, batch size, and number of epochs were tuned to achieve optimal performance.
5. Evaluating the model: the model was evaluated using language model evaluation Harness[11].

5 Evaluation

To evaluate the performance of the trained HR-LLM, we will conduct benchmarking tasks including Named Entity Recognition (NER), Sentiment Analysis (SA), and Choice Of Plausible Alternatives (COPA) on the Croatian language using the BENCHić (Rupnik et al., 2023) benchmarking dataset [12]. Furthermore, we intend to assess the LLM's proficiency in knowledge completion in Croatian by utilising established datasets such as MMLU (Hendrycks et al., 2021b,a), ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019), Belebele (Bandarkar et al., 2023), and TruthfulQA (Dac Lai et al., 2023). Here we present our initial zero-shot evaluation results on benchmarking datasets in Table 4. Since the full evaluation of the HR-LLM performance is still a work-in-progress, its final results are beyond the scope of this paper.

6 Conclusion

Recent advancements in NLP/LT have demonstrated that LLMs have become essential for reliable and efficient language processing capabilities. Imbalanced data distributions among different languages during multilingual LLM pre-training demonstrably weakens the monolingual proficiency of a multilingual LLM applied to the specific monolingual tasks, particularly in a lower-resourced languages.

As an attempt to overcome this situation, we have developed the first monolingual Croatian LLM: HR-LLM (beta version). This paper presents the results of the project HR-XR-XTEND so far (trained HR-LLM for Croatian) and the remaining work to be completed (its full evaluation for selected downstream tasks). While the primary objective of this study is to develop a HR-LLM using the Croatian language data only, the key findings and achievements may have broader implications for lower-resourced languages in Europe and around the globe in subsequent developments.

The direct outputs of the project, particularly the current state of the HR-LLM (beta version), will be available through HR-CLARIN[13] repository with permissive licenses. The final version of the HR-LLM will also be availabe through the same repository after the completion of the project.

7 Acknowledgements

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

[8] https://language-data-space.ec.europa.eu/

[9] https://language-data-space.ec.europa.eu/related-initiatives/alt-edic

[10] https://github.com/EleutherAI/gpt-neox

[11] https://github.com/EleutherAI/lm-evaluation-harness

[12] https://github.com/clarinsi/benchich

[13] https://www.clarin.hr

benchmark	metric	160M	350M	160M	1.4B	gpt2-cpt
arc_hr	acc	18.91	**20.96**	20.36	20.44	19.5
	acc_norm	23.44	**25.49**	25.06	24.89	23.7
belebele_hrv_Latn	acc	22.78	23	**23.11**	22.67	22.89
	acc_norm	22.78	23	**23.11**	22.67	**22.89**
hellaswag_hr	acc	28.43	29.87	30.08	**31.36**	26.67
	acc_norm	30.07	32.74	33.38	**35.52**	27.95
m_mmlu_hr	acc	22.65	**25.21**	22.8	22.54	23.05
truthfulqa_hr_mc1	acc	25.88	24.58	25.75	26.27	**27.05**
truthfulqa_hr_mc2	acc	43.82	42.21	42.34	42.52	**46.88**

Table 4: Results of zero-shot evaluation on benchmarking datasets ARC, Belebele, MMLU, HellaSwag and TruthfulQA.

References

Babić, K., Petrović, M., Beliga, S., Martinčić-Ipšić, S., Matešić, M., and Meštrović, A. (2021). Characterisation of covid-19-related tweets in the croatian language: framework based on the cro-cov-csebert model. *Applied Sciences*, 11(21):10442.

Bandarkar, L., Liang, D., Muller, B., Artetxe, M., Shukla, S. N., Husa, D., Goyal, N., Krishnan, A., Zettlemoyer, L., and Khabsa, M. (2023). The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*.

Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., Pieler, M., Prashanth, U. S., Purohit, S., Reynolds, L., Tow, J., Wang, B., and Weinbach, S. (2022). GPT-NeoX-20B: An open-source autoregressive language model. In Fan, A., Ilic, S., Wolf, T., and Gallé, M., editors, *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin. Association for Computational Linguistics.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. (2018). Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.

Dac Lai, V., Van Nguyen, C., Ngo, N. T., Nguyen, T., Dernoncourt, F., Rossi, R. A., and Nguyen, T. H. (2023). Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv e-prints*, pages arXiv–2307.

European Parliament EUBERT Hugging Face (2023). Huggingface/europeanparliament/eubert. Accessed: 19-02-2024.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2022). Language-agnostic bert sentence embedding.

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., and Steinhardt, J. (2021a). Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2021b). Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

HuggingFace macedonizer hr gpt2 (2021). Huggingface/macedonizer/hr-gpt2. Accessed: 19-02-2024.

HuggingFace macedonizer hr-roberta-base (2021). Huggingface/macedonizer/hr-roberta-base. Accessed: 19-02-2024.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023). Mistral 7b.

Kahana, A., Mathew, J. S., Bleik, S., Reynolds, J., and Elisha, O. (2024). Evaluation methodology for large language models for multilingual document question and answer. *arXiv preprint arXiv:2402.01065*.

Kenton, J. D. M.-W. C. and Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Krauwer, S. (1998). Elsnet and elra: A common past and a common future. *ELRA Newsletter*, 3(2):4–5.

Lai, V., Ngo, N., Pouran Ben Veyseh, A., Man, H., Dernoncourt, F., Bui, T., and Nguyen, T. (2023). ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Language Report Croatian (2022). European language equality. Accessed: 19-02-2024.

Language Report Hungarian (2022). European language equality. Accessed: 19-02-2024.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., and Carlini, N. (2022). Deduplicating training data makes language models better. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

Ljubešić, N. and Lauc, D. (2021). BERTić - the transformer language model for Bosnian, Croatian, Montenegrin and Serbian. In Babych, B., Kanishcheva, O., Nakov, P., Piskorski, J., Pivovarova, L., Starko, V., Steinberger, J., Yangarber, R., Marcińczuk, M., Pollak, S., Přibáň, P., and Robnik-Šikonja, M., editors, *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 37–42, Kiyv, Ukraine. ACL.

Marinova, I., Simov, K., and Osenova, P. (2023). Transformer-based language models for Bulgarian. In Mitkov, R. and Angelova, G., editors, *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 712–720, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., and Sagot, B. (2020). CamemBERT: a tasty French language model. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Mochtak, M., Rupnik, P., and Ljubešić, N. (2023). The parlasent multilingual training dataset for sentiment identification in parliamentary proceedings. *arXiv preprint arXiv:2309.09783*.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Kaiser, Ł., Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Kondraciuk, Ł., Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., MÃ©ly, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng,

E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. (2024). Gpt-4 technical report.

Papadimitriou, I., Lopez, K., and Jurafsky, D. (2023). Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models. In Beinborn, L., Goswami, K., Muradoğlu, S., Sorokin, A., Kumar, R., Shcherbakov, A., Ponti, E. M., Cotterell, R., and Vylomova, E., editors, *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 143–146, Dubrovnik, Croatia. Association for Computational Linguistics.

Pfeiffer, J., Goyal, N., Lin, X. V., Li, X., Cross, J., Riedel, S., and Artetxe, M. (2022). Lifting the curse of multilinguality by pre-training modular transformers. *arXiv preprint arXiv:2205.06266*.

Rijgersberg, E. and Lucassen, B. (2023). Geitje: een groot open nederlands taalmodel.

Rupnik, P., Kuzman, T., and Ljubešić, N. (2023). BENCHić-lang: A benchmark for discriminating between Bosnian, Croatian, Montenegrin and Serbian. In Scherrer, Y., Jauhiainen, T., Ljubešić, N., Nakov, P., Tiedemann, J., and Zampieri, M., editors, *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 113–120, Dubrovnik, Croatia. Association for Computational Linguistics.

Singh, P., Maladry, A., and Lefever, E. (2023). Too many cooks spoil the model: Are bilingual models for Slovene better than a large multilingual model? In Piskorski, J., Marcińczuk, M., Nakov, P., Ogrodniczuk, M., Pollak, S., Přibáň, P., Rybak, P., Steinberger, J., and Yangarber, R., editors, *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 32–39, Dubrovnik, Croatia. Association for Computational Linguistics.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023a). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023b). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ulčar, M. and Robnik-Šikonja, M. (2020). Finest bert and crosloengual bert: less is more in multilingual models. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings 23*, pages 104–111. Springer.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., and Pyysalo, S. (2019). Multilingual is not enough: Bert for finnish.

Workshop, B., Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., et al. (2022). Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Yang, Z. G., Laki, L. J., Váradi, T., and Prószéky, G. (2023). Mono-and multilingual gpt-3 models for hungarian. In *International Conference on Text, Speech, and Dialogue*, pages 94–104. Springer.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. (2019). Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Zhang, X., Malkov, Y., Florez, O., Park, S., McWilliams, B., Han, J., and El-Kishky, A. (2022). Twhin-bert: a socially-enriched pre-trained language model for multilingual tweet representations. *arXiv preprint arXiv:2209.07562*.

0em 1.5em