Data-Efficient Adaptation of Multilingual LLMs to Ukrainian

Yurii Paniv[1], Bohdan Didenko[2], Mykola Haltiuk[3], Vladyslav Humennyy[1], Andrian Kravchenko[1], Roman Kyslyi[4] Viktoriia Makovska[1], Artem Orlovskyi[5], Bohdan Ruban[1], Maksym-Yurii Rudko[1], Anastasiia Senyk[1] Nazarii Drushchak[1], Dmytro Chaplynskyi[1], Mariana Romanyshyn[6]

[1]Ukrainian Catholic University
[2]Lviv Polytechnic National University
[3]AGH University of Science and Technology
[4]Kyiv School of Economics
[5]National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
[6]Grammarly

Correspondence: paniv@ucu.edu.ua

Abstract

Adapting large language models to low-resource languages presents three interconnected challenges: inefficient tokenization, scarcity of high-quality annotated data, and limited resources for instruction-tuning. We present a reproducible approach that addresses each challenge using data-centric methods that primarily rely on unlabeled text corpora, parallel translation data, and a multilingual base model. Our approach combines (1) vocabulary surgery for tokenizer adaptation without full retraining, (2) cross-lingual transfer of quality classifiers via translation, enabling filtering without target-language annotations, and (3) generation of instruction data through translation, task conversion, and targeted synthesis. We validate this recipe by adapting Gemma-3-12B to Ukrainian. Our pretrained model achieves top performance on Ukrainian benchmarks, while our instruction-tuned variant demonstrates strong performance on translation (33 BLEU on FLORES), summarization, and question-answering tasks, while requiring 1.5x fewer tokens than the original model for the same text. We release all models, datasets, classifiers, and code to enable replication for other languages.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities across diverse languages and tasks, but their effectiveness and efficiency for lower-resource languages remain limited. For example, models such as Llama-3 (Grattafiori et al., 2024) were trained on 75% English data, 17% code, and the remaining 8% on other 70 languages. Beyond the dataset design choices to include multilingual training in the data, there is an issue with the overall availability of data for both pretraining stages and, crucially, instruction-tuning data, which is scarce for low- and mid-resource languages. This prompts researchers to consider data-efficient methods for model training and synthetic data generation for creating additional scarce corpora.

Due to the data mixture described above, tokenizers trained on such data introduce a common problem for low- and mid-resource languages—multilingual tokenizers also exhibit poor compression rates on them, leading to a noticeable efficiency gap (in terms of inference cost) relative to English. This results in an inherent disadvantage of using LLMs in the native language, which influences the deployment of those models in industry.

We conduct a case study on the Ukrainian language, focusing on methods for adapting models to the target language: data collection, adaptation of existing English-language resources, synthetic data generation, and model adaptation. We primarily rely on raw unannotated text corpora (such as CulturaX by Nguyen et al. (2024), FineWeb 2 by Penedo et al. (2025), etc), parallel corpora (Schwenk et al., 2021; Fan et al., 2020), methodology for their filtering (Chaplynskyi and Zakharov, 2025), and multilingual models, preferably with at least basic support of a target language (Team et al., 2025; Grattafiori et al., 2024).

Regarding tokenizer performance, Ukrainian uses the Cyrillic script, so most models are poor at compressing Ukrainian text. We addressed this challenge by training and transferring the tokenizer of our base model using modern methods, thereby achieving a speedup of up to 1.5x with negligible performance drop.

As a result of our study, we introduce Lapa[1], a

[1]Named after Valentyn Lapa, who with Oleksiy Ivakhnenko created the Group Method of Data Handling, a predecessor to Deep Learning.

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 155–168

12B parameter language model designed mainly for Ukrainian language processing. Our contributions are the following:

(1) A detailed recipe for adapting multilingual LLMs to the target language in mid- and low-resource settings; (2) A comprehensive data filtering pipeline using native and transferred English classifiers to estimate the text quality: educational value, propaganda detection, disinformation and manipulative content detection, and grammatical correctness; (3) An open dataset collection including approximately 30B tokens of filtered pretraining data and instruction-tuning datasets; (4) Competitive benchmark results that match or exceed larger multilingual models on Ukrainian tasks while enabling efficient on-device deployment. Our instruction datasets enable better Ukrainian text processing than the best models we've evaluated across most tasks.

All models, datasets, and training code are released under the MIT license on GitHub[2] and HuggingFace[3] to support further research and development of Ukrainian NLP technologies.

2 Related Work

2.1 Multilingual Language Models

Recent multilingual models like Gemma-3 (Team et al., 2025), Qwen (Yang et al., 2025), Mistral (Jiang et al., 2023), and Llama (Grattafiori et al., 2024) have improved performance across many languages through scaling and better training data. However, these models face efficiency challenges for languages with different writing systems or linguistic structures, particularly when tokenizers are optimized primarily for English. Approaches to improving low-resource language model performance include continued pretraining (Gupta et al., 2023), vocabulary adaptation (Kim et al., 2024), and multilingual transfer learning (Conneau et al., 2020). Recent work has shown that data quality matters more than quantity for instruction-tuning (Zhou et al., 2023), suggesting that careful data curation is particularly important for low-resource languages.

From a data perspective, the ideal setting for model training is an abundance of pretraining and instruction-tuning data, consistent with Chinchilla scaling laws (Hoffmann et al., 2022). Given the inherent limitations in data availability for lower-resource languages, model authors must close the gap with high-quality data or data augmentation techniques.

2.2 Ukrainian NLP

Previous work on Ukrainian NLP includes the development of benchmark datasets (Chaplynskyi and Romanyshyn, 2024; Syvokon et al., 2023) and earlier Ukrainian language models like UAlpaca (Paniv, 2023). The UNLP workshop series has fostered the development of datasets for tasks including grammatical error correction, named entity recognition, and machine translation. However, to the best of our knowledge, no previous work has combined efficient tokenization, large-scale pretraining with quality filtering, and adding instruction-following capabilities in a single openly available model to adapt a multilingual model to a specific language.

2.3 Data Quality for Pretraining

Recent work has emphasized the importance of data quality in pretraining. FineWeb-Edu (Lozhkov et al., 2024) demonstrated that educational value filtering improves downstream performance. DataComp-LM (Li et al., 2025) and Nemotron-CC (Su et al., 2025) showed that aggressive model-based filtering can achieve better performance-to-data ratios. We build on these approaches while addressing Ukrainian-specific challenges, such as misinformation detection and code-switching (through grammatical error correction).

3 Model Architecture and Tokenizer

3.1 Base Model Selection

We selected Google's Gemma-3-12B-PT (Team et al., 2025) as our base model based on several criteria: (1) strong performance on Ukrainian benchmarks (Table 1) to leverage stronger foundation; (2) balanced size enabling consumer GPU (24 GB of VRAM) deployment; (3) multimodal support for vision-language tasks; (4) permissive license allowing commercial use.

3.2 Tokenizer

Recent work has explored custom tokenizers for Ukrainian via bilingual vocabularies (Kiulian et al., 2025) and hybrid morpheme+BPE tokenisation for Ukrainian (Borodavko, 2025). Our tokenizer was developed in parallel to these efforts: we share the overall goal of reallocating vocabulary

[2] https://github.com/lapa-llm/lapa-llm
[3] https://huggingface.co/lapa-llm

Model	Belebele	MMLU	FLORES	Avg Rank
Gemma-3-27B	91.56	71.16	22.46	3.00
Gemma-3-12B	89.56	64.07	21.98	7.00
Qwen3-14B	90.56	70.64	11.02	11.00

Table 1: Baseline performance of candidate models on Ukrainian benchmarks. For benchmarking performance we use Ukrainian LLM Leaderboard (Paniv, 2025).

capacity towards Ukrainian, but focus on a minimal Gemma 3 compatible surgery that avoids overlapping-vocabulary issues seen in other Slavic settings (Ociepa et al., 2025).

We adapt the original `SentencePiece` tokenizer of Gemma 3 with a vocabulary size of 256 thousand tokens to better support Ukrainian while preserving the model's behavior on English, all official EU languages, and several neighboring languages important to the Ukrainian context.

Vocabulary Surgery. We perform a three-step surgery on the Gemma 3 vocabulary. First, we analyse more than sixteen writing systems and significantly shrink only those scripts that are geographically and culturally distant from Ukraine (e.g., Chinese, Bengali, Thai, Japanese, Korean), while keeping Latin-based EU languages and scripts of minority languages of Ukraine (such as Turkish, Armenian, Georgian) essentially intact. All Cyrillic tokens from the original tokenizer (13,398 entries) are removed, together with a subset of `<unused-*>` tokens; there are no conflicts between old and new merges. Second, we train a Ukrainian-centric Cyrillic donor tokenizer on the Kobza (Haltiuk and Smywiński-Pohl, 2025) corpus, using the same settings as Gemma 3's tokenizer. Third, we deterministically reassign the freed token IDs to the donor Cyrillic subwords. Tokens from other writing systems that do not appear in the "Replaced tokens" table preserve both their string form and IDs, so English, EU languages, and the above-mentioned neighboring languages behave exactly as in the base model. A detailed breakdown of removed and retained tokens per script is given in Table 8 in Appendix A.

As a result. The surgery-based design reduces the average number of tokens per Ukrainian word from about 2.5 to 1.6 (roughly 35% fewer tokens, or $\approx$1.5$\times$ more Ukrainian text in the same 32k-token context), while leaving English and EU languages tokenization essentially unchanged (see Table 9 in the appendix).

Adapting Embeddings. To adapt the model's embeddings to the newly introduced tokens, we used the Model-Aware Tokenizer Transfer method (Haltiuk and Smywiński-Pohl, 2025). It utilizes cross-tokenizer self-distillation to model inter-token communication in attention layers. This allows us to recover most of the original model's performance on downstream tasks after tokenizer transfer, as shown in Table 2.

Model	Belebele	Global MMLU	Long FLORES
Gemma 3 12B PT	89.33	67.03	14.36
MATT	89.56	64.98	8.70

Table 2: Performance on Belebele, Global MMLU (accuracy, %), and Long FLORES (BLEU) of the original Gemma 3 12B PT model compared to itself after tokenizer transfer with MATT.

We initialize the embeddings using FOCUS (Dobler and de Melo, 2023), which then serves as a starting point for the MATT training. Following the original paper, we train embeddings for the new tokens on a small subset of our pretraining corpus comprising 240 million tokens using the AIM objective with MSE loss on the 12th layer out of 34.

The resulting tokenizer remains fully compatible with Gemma 3 with respect to vocabulary size, special tokens, and pre- and post-processing.

This improvement has practical implications: (1) faster inference: fewer tokens mean proportionally less computation (2) longer effective context: the same 32K token limit covers more Ukrainian text with English compression rate intact (3) lower deployment costs due to reduced computational requirements.

4 Pretraining Data Collection and Filtering

4.1 Data Sources

Our pretraining data combines several sources.

Kobza Kobza (Haltiuk and Smywiński-Pohl, 2025) is a large corpus of Ukrainian web text created by combining Ukrainian subsets of multiple multilingual corpora into a single data source with heavy deduplication. We employ additional filtering and deduplication to ensure the highest data

Subcorpora	Documents	Tokens
CulturaX	24,942,577	15,002,455,535
FineWeb 2	32,124,035	19,114,177,138
HPLT 2.0	26,244,485	20,709,322,905
UberText 2.0	6,431,848	2,904,208,874
Ukrainian News	7,175,971	1,852,049,111
Total	96,918,916	59,582,213,563

Table 3: Composition of source data before filtering. Kobza dataset is represented here by CulturaX, FineWeb, HPLT and Ukrainian News datasets.

quality for our model's continual pertaining, as described in subsection 4.2.

Institutional Books High-quality books provided by Harvard Law School Library's Institutional Data Initiative (Cargnelutti et al., 2025). It contains a relatively small amount of Ukrainian text, but it is of higher quality, as confirmed by our quality assessment models.

Ukrainian News As part of Kobza dataset, we include Ukrainian news to improve understanding of global events and broader cultural context.

Public Datasets Additional licensed data to capture Ukrainian cultural and historical context. This includes the Ukrainian subset of YODAS2 (Li et al., 2023) transcribed using OpenAI's Whisper (Radford et al., 2022). We include speech for improved understanding of the Ukrainian language.

4.2 Filtering Pipeline

We implemented a multi-stage filtering pipeline inspired by Nemotron-CC (Su et al., 2025) but adapted for Ukrainian-specific challenges:

Language Identification and Normalization We first identify Ukrainian text and fix Unicode encoding issues common in web-scraped data.

Deduplication We perform both exact and fuzzy deduplication to remove redundant content while preserving diverse expressions of similar ideas.

Heuristic Filtering Following practices as outlined in Nemotron paper, we apply heuristics to remove low-quality content using the following rules: (1) Non-alphanumeric character ratio < 0.25 (2) Symbol-to-word ratio < 0.1 (3) Number-to-text ratio < 0.15 (4) URL-to-text ratio < 0.2 (5) Whitespace ratio < 0.25.

The only modification we make to the rules is that we remove English-specific heuristics, like calculating the ratio of English characters in text to total unique characters. Overall, at this stage, we discard approximately 14% of the total data.

Model-Based Quality Classification The core of our filtering uses ensemble classifiers to score multiple quality dimensions. Since no large-scale annotated quality datasets exist for Ukrainian, we employ a transfer learning approach: (1) Translate 500K random samples from our corpus to English using our best available translation models. For selecting a translation model, we must be sure of model's long-context translation performance. For that purpose, we use LongFLORES (Paniv, 2025) benchmark, which is a long-context version of industry-standard FLORES (NLLB Team, 2022) benchmark. (2) Train Ukrainian classifiers using translated examples and English quality models as teachers. (3) We validate transferred classifiers on the held-out 10% of translated data, which we use for the test set. This means that, for transferred models, the F1 score is not a direct measure of model performance but rather of the success of language transfer.

Overall, our quality dimensions include:

Educational Value: We transfer two models from English: FineWeb-Nemotron-Edu (F1=0.96) and FineWeb-Mixtral-Edu (score of transferred performance: F1=0.94), which classify texts by educational value from 0-5.

Informational Value: We use FastText-OH-ELI5 (transferred performance: F1=0.67), a classifier distinguishing between high-quality Reddit ELI5 content and generic Common Crawl data. Looking at final distribution of quality scores, we noticed a lack of strong scores for high-quality data in Ukrainian, to which we attribute performance drop of transferred model.

Propaganda and Disinformation Detection: We developed a novel classifier (F1=0.96) trained on Ukrainian propaganda and fact-checking datasets based on the VoxCheck (VoxCheck, 2018) project. The model scores texts on a scale from 0 (propaganda) to 1 (factual claim).

Manipulative Content Detection: Using data from the UNLP 2025 Shared Task (Kyslyi et al., 2025), we trained a classifier (F1=0.74) to identify manipulative language patterns, which we don't want to include in model training.

Grammatical Correctness: We trained a classifier (F1=0.71) using Ukrainian grammatical error correction datasets.

Classifier	Transfer	F1 Score
FineWeb-Nemotron-Edu	Yes	0.96
FineWeb-Mixtral-Edu	Yes	0.94
FastText-OH-ELI5	Yes	0.67
Propaganda Detection	No	0.96
Manipulative Content	No	0.74
Grammar Correctness	No	0.71

Table 4: Quality classifier performance. "Transfer" indicates whether the classifier was trained via English-to-Ukrainian transfer learning.

Score Combination and Bucketing We combine classifier scores into a composite quality metric and assign each document to quality buckets using the cumulative distribution function (CDF) binning. Documents are assigned scores from 0 to 20 based on their bucket.

We perform manual inspection of documents across the quality spectrum to validate our filtering. Documents with maximum scores ≤ 10 across all classifiers are removed as definitively low-quality. The remaining data is split into two groups: 1) **High-quality**: Documents in bucket 19 plus high-scoring documents from other buckets; 2) **Regular-quality**: Remaining documents above the threshold.

This filtering reduces our dataset from 60B tokens to approximately 30B tokens while considerably improving data quality.

5 Pretraining

5.1 Training Setup

We trained on the filtered 30B token dataset using the following configuration as described in Table 5.

5.2 Data Mixing Strategy

We employ a quality-based data mixing strategy. For first 70% of training, we use regular-quality data for broad coverage, and for the rest 30% (decay phase), we use high-quality data for refinement.

5.3 Results

Our pretrained model achieves the top position on our Ukrainian benchmark suite (Table 6), confirming validity of our approach.

6 Instruction-Tuning

6.1 Dataset Creation

We created instruction-tuning datasets through three approaches. We fine-tune base models on the resulting instruction-tuning datasets, restoring Gemma's original instruction-following capabilities and outperforming it on most tasks.

Translation of English Instruction Data Additional general datasets with proper license translated to Ukrainian to help the model solve tasks more efficiently (e.g., coding, reasoning). We used Gemma-3-27B-IT as the most capable model for long-context translation to Ukrainian at the time.

We translated high-quality English instruction-tuning datasets using the specific prompt, that can leave code portion of the data intact. We calibrate translation prompt manually on 100 randomly sampled examples. We used Hermes-3-Dataset (Teknium, 2023) , both translated version and original in training to retain model's English capabilities, and LeetCode Dataset (Xia et al., 2025) for code generation tasks.

Ukrainian Task Conversion We converted existing Ukrainian NLP datasets into instruction format: (1) UA-GEC (Syvokon et al., 2023) for grammatical error correction; (2) NER-UK 2.0 (Chaplynskyi and Romanyshyn, 2024): Named entity recognition; (3) UberText-NER-Silver (Radchenko and Drushchak, 2025): Silver-standard NER annotations; (4) UA-Lawyer dataset (Smoliakov, 2025) for answering legal-specific questions in Ukrainian context; (5) FiftyFiveShades (Chaplynskyi and Zakharov, 2025) parallel corpus for English-Ukrainian translation.

FiftyFiveShades is a deduplicated dataset for parallel sentences, collected from the Open Parallel Corpora project, rated by six Quality Estimation models, together with the score derived from an ensemble of these models, which best correlates with human judgment. For the purpose of instruction-tuning of our model , only the top 10% of the dataset, sorted by model quality score, were used to enable even better translation capabilities in the model.

Synthetic Data Generation Following Nemotron-CC approach, for high-quality pretraining, we used Gemma-3-27B-IT to generate instruction-response pairs to include in the pretraining data: (1) **Diverse QA**: Questions in various formats (yes/no, open-ended, multiple-choice) about factual information. Specifically, a dataset of more than 1.3 million entries was created, containing questions based on the context of Ukrainian text documents of various domains and lengths. To generate a single entry in the dataset, the model is

	Base Model	Instruction-Tuned
Total Training Tokens	30 billion	≈ 1.2 billion
Max Sequence Length	8,192	16,384
Sample Packing	Enabled	Enabled
Loss	Cross-Entropy	DFT (Wu et al., 2025)
Global Batch Size (Tokens)	≈ 458,752	≈ 2 million
Micro Batch Size (per GPU)	1	1
Gradient Accumulation Steps	1	10
Total Training Steps	≈ 65,394	≈ 600
Learning Rate (Peak)	1e-5	2e-5
Schedule	Warmup-Stable-Decay (WSD)	
Warmup Steps	6,539	100
Decay Steps	20,926	100
Min LR Ratio	0.05	-
Weight Decay	0.005	0.005
Optimizer	AdamW (Fused)	
Betas	(0.9, 0.977543)	(0.9, 0.98)
Epsilon	1e-6	1e-6
Training Hardware	56x H100 80GB	12x GH200 96GB
Training Strategy	Hybrid-FSDP (Intra-node FSDP / Inter-node DDP)	
Precision	BF16 + Flash Attention 2	

Table 5: Pretraining hyperparameters for pretrained and instruct models.

prompted with the full text document for context, and then it is asked to provide five questions with correct answers for each, as well as two or three incorrect answers. (2) **Distillation**: Rewriting text into concise, clear passages. (3) **Knowledge Extraction**: Extracting key information while discarding uninformative content (4) **Knowledge Lists**: Organizing information as structured lists.

For generating more than 30 thousand instructions on NER, paraphrase, and simplification tasks, entries from the UberText 2.0 corpus (Chaplynskyi, 2023) were used, specifically the "Wikipedia cleansed" portion.

Rule-based Synthetic Data Generation Additional Grammatical Error Correction dataset was created using pre-defined rules that were applied to random parts of the original text from the above-mentioned UberText Corpus (Fiction section). These rules include altering or completely removing the ending of a word, randomly dropping a vowel, duplicating a letter, and other similar modifications.

6.2 Vision Instruction Data

For multimodal capabilities, we rendered a subset of 500K of high-quality documents in Markdown as images to synthetically create OCR dataset. As measured on MMZNO benchmark (Paniv et al., 2025), a slight fine-tune enabled performance improvement on vision 51.78% to 58.54% over original Gemma-3-12B-it.

6.3 Results

We trained on the combined 1.5B token dataset using the following configuration as described in Table 5.

Our instruction-tuned model demonstrates strong performance across multiple benchmarks as shown in Table 6.

Overall, in its size, our model is the best performing, with the only weak side as measured on translated IFEval, which indicates poorer instruction-following capabilities. Notably, for practical tasks, such as Q&A our model is the best in its class. Besides that, we achieve 33 BLEU on English-to-Ukrainian translation (FLORES benchmark), making our model the best available Ukrainian translator among open models, which should further contribute to improvements in data availability for Ukrainian language.

7 Additional experiments

7.1 Reasoning Capabilities

We developed reasoning capabilities following the DeepSeek-R1 (DeepSeek-AI, 2025) approach with adaptations for Ukrainian: (1) Translated

Model	Average Rank	IFEval Ukrainian	FLORES EN→UK	Long FLORES EN→UK	MMLU Ukrainian
INSAIT-Institute/MamayLM-Gemma-3-12B-IT-v1.0 (0-shot)	**3.18**	**61.18**	29.40	30.13	64.29
Pretrained (ours) (3-shot)	3.76	20.70	**33.44**	**33.14**	62.84
google/gemma-3-12b-pt (3-shot)	4.24	19.59	30.75	31.47	66.54
Instruction-Tuned (ours) (0-shot)	4.53	31.79	33.37	32.50	61.85
Qwen/Qwen3-8B-Base (3-shot)	4.59	25.32	22.53	20.23	**67.15**
google/gemma-3-12b-it (0-shot)	6.18	58.41	3.58	15.63	53.77
meta-llama/Llama-3.1-8B (3-shot)	6.41	9.80	24.22	23.76	51.62
google/gemma-3-4b-pt (3-shot)	7.00	22.37	26.20	26.63	51.10
meta-llama/Llama-3.1-8B-Instruct (0-shot)	7.35	35.49	18.64	15.67	42.94
google/gemma-3-4b-it (0-shot)	8.00	48.61	3.05	16.24	31.09
Qwen/Qwen3-8B (0-shot)	10.35	59.52	0.71	0.78	22.95

Table 6: Model evaluations based on Ukrainian LLM Leaderboard. Average rank represents an average rank model gets across 17 different tasks. Our model demonstrates SOTA performance on translation tasks, and, combined with token efficiency, is a fast and cheap instrument to obtain natural instruction-tuning data in Ukrainian.

Experiment	**Persona ↑**	**Leak ↓**	**KL ↓**	**Act. Drift ↓**	**SQuAD F1 ↑**
Base (unmodified model; `base`)	0.007	0.652	–	–	56.83
E0: Gentle Attn+MLP (`exp0`)	0.018	0.102	7.89	1.08	56.40
E0-MLP: MLP-only (`exp0_mlp`)	**0.030**	0.225	**4.97**	**0.72**	**56.49**
E1: Neuron-Masked MLP (`exp1`)	0.020	0.012	24.04	2.39	56.38
E2: Strong Attn+MLP (`exp2`)	0.021	**0.001**	47.31	3.37	55.91
E3: Circuit-Guided Attn+MLP (`exp3`)	0.014	**0.001**	47.36	3.38	55.89

Table 7: Core identity-editing experiments. Higher Persona indicates stronger compliance with identity constraints; lower Leak, KL, and activation drift indicate better preservation.

OpenThoughts-114K (Guha et al., 2025) mathematical and coding reasoning dialogs, including inputs, reasoning traces, and outputs; (2) Generated synthetic reasoning examples from high-quality Ukrainian content; (3) Fine-tuned on approximately 1B tokens of reasoning data.

Interestingly, training on math and code reasoning data enabled generalization to other question types, even when starting from the pretrained checkpoint rather than the instruction-tuned model. Unfortunately, the reasoning model performed poorer than the instruction-tuned variant, so we reserve exploration in that area for future work.

7.2 Persona Editing

Pretrained language models retain unwanted self-identification statements (e.g., "I am Gemini"). We investigate targeted interventions to ablate identity representations while preserving language competence, using LoRA-based edits (Hu et al., 2022) as efficient alternatives to ROME/MEMIT (Meng et al., 2022, 2023). Identity removal uses negative task-vector updates (Ilharco et al., 2023) with KL-to-reference regularization (Schulman et al., 2017) to prevent drift and address residual traces under adversarial prompting (Carlini et al., 2021).

We explore mechanistically motivated interventions informed by causal tracing (Meng et al., 2022): (1) neuron masking via Gumbel-Softmax selection (Jang et al., 2017) restricting updates to high-importance MLP components, (2) contrastive masking across attention and MLP projections, and (3) circuit-guided editing targeting identity-relevant components. Experiments evaluate four configurations (E0-E3) on our instruct variant, measuring forget success via identity leakage, while monitoring KL divergence and SQuAD-UK performance (Rajpurkar et al., 2016).

Results in Table 7 show that stronger interventions (E2, E3) eliminate identity leakage but increase KL divergence, whereas gentler edits (E0-MLP) minimize drift with modest suppression. No configuration degrades benchmark performance, confirming localized editing preserves capabilities. However, persona compliance remains low across variants, indicating identity removal alone is insufficient for establishing alternative personas.

8 Evaluation

8.1 Benchmark Suite

Our evaluation suite is based on Ukrainian LLM Leaderboard (Paniv, 2025) We evaluate on a comprehensive suite of Ukrainian benchmarks: (1) Belebele Ukrainian: Reading comprehension (2) MMLU Ukrainian: Multi-task understanding (3) FLORES Ukrainian: Machine translation (4) SQuAD Ukrainian: Question answering (5) XL-Sum Ukrainian: Summarization and (6) MMZNO: Multimodal understanding (images + Ukrainian text) benchmark (Paniv et al., 2025).

9 Discussion

9.1 Societal Impact

Our model enables several positive applications, such as processing sensitive documents locally without privacy concerns, supporting Ukrainian-language technology development, reducing computational costs, improving energy efficiency, and providing educational resources through high-quality QA and summarization.

Potential risks include misuse to generate misleading content, despite our filtering efforts; over-reliance on model outputs without human verification; and biases in the training data that may affect model responses not detected by our quality classifiers.

We mitigate these through open release of models and datasets, enabling community auditing and improvement.

10 Conclusion

We presented our high-quality Ukrainian language model that demonstrates competitive performance while being significantly more efficient than existing alternatives. Using our recipe, which goes through three critical stages: state-of-the-art tokenizer transfer, comprehensive data filtering using both Ukrainian-specific and adapted English quality classifiers, and carefully curated instruction data, we created a model that achieves strong benchmark results and practical utility.

Our work shows that using our methods, researchers can efficiently adapt the model to the target language. The complete open release of models, data, and code aims to accelerate Ukrainian and multilingual NLP research and to enable practical applications requiring local, efficient, and culturally aware language processing.

Future work includes improving reasoning evaluation, expanding multimodal capabilities, and exploring more data-efficient fine-tuning methods for language adaptation. We invite the community to build upon our work and contribute to the development of multilingual and Ukrainian language technologies.

Limitations

Our work has several limitations. The model inherits its biases from the base model and training data, which we try to mitigate by filtering unsafe data using our quality classifiers. Some quality classifiers achieved moderate performance (F1=0.67-0.74), suggesting room for improvement, or adopting more data-efficient methods like SetFit (Tunstall et al., 2022), which could improve both quality and training efficiency. Worse performance than current SOTA models for IFEval-type tasks indicates the need for attention in this direction. We haven't performed RL on target tasks, which could improve generalization and instruction-following. We haven't tried model merging, which could be beneficial for adding new capabilities to the model without erasing previous ones.

Acknowledgements

Primarily, we would like to express our gratitude to startup Comand.AI for compute support, without which this project would not be possible. We would like to also thank ELEKS for support of this project through a grant dedicated to the memory of Oleksiy Skrypnyk. EuroHPC supported this project through compute grant EHPC-BEN-2025B12-043. We would like to thank Talents for Ukraine project of Kyiv School of Economics for the grant on compute resources. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2025/018285. Sincere thanks to HuggingFace, which provided the team with a free corporate subscription to store models and datasets. Our deepest gratitude goes to Oleksii Molchanovsky, Yurii Filipchuk, Artur Kiulian, Nikita Trynus, Marko Kostiv and Oles Dobosevych, who helped at various stages of this project. In addition, we would like to thank the reviewers for their feedback.

References

Vitalii Borodavko. 2025. Hybrid tokenization for Ukrainian language: Using morphemes and BPE. Master's thesis, Ukrainian Catholic University, Lviv, Ukraine. Faculty of Applied Sciences, Department of Computer Sciences.

Matteo Cargnelutti, Catherine Brobston, John Hess, Jack Cushman, Kristi Mukk, Aristana Scourtas, Kyle Courtney, Greg Leppert, Amanda Watson, Martha Whitehead, and Jonathan Zittrain. 2025. Institutional books 1.0: A 242b token dataset from harvard library's collections, refined for accuracy and usability. *Preprint*, arXiv:2506.08300.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

Dmytro Chaplynskyi. 2023. Introducing UberText 2.0: A corpus of modern Ukrainian at scale. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Dmytro Chaplynskyi. 2025. lang-uk/malyuk · Datasets at Hugging Face — huggingface.co. https://huggingface.co/datasets/lang-uk/malyuk. [Accessed 01-02-2026].

Dmytro Chaplynskyi and Mariana Romanyshyn. 2024. Introducing NER-UK 2.0: A rich corpus of named entities for Ukrainian. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 23–29, Torino, Italia. ELRA and ICCL.

Dmytro Chaplynskyi and Kyrylo Zakharov. 2025. A framework for large-scale parallel corpus evaluation: Ensemble quality estimation models versus human assessment. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 73–85, Vienna, Austria (online). Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Konstantin Dobler and Gerard de Melo. 2023. FOCUS: Effective embedding initialization for monolingual specialization of multilingual models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13440–13454, Singapore. Association for Computational Linguistics.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *Preprint*, arXiv:2010.11125.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, and 31 others. 2025. Openthoughts: Data recipes for reasoning models. *Preprint*, arXiv:2506.04178.

Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats L. Richter, Quentin Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. 2023. Continual pre-training of large language models: How to (re)warm your model? *Preprint*, arXiv:2308.04014.

Mykola Haltiuk and Aleksander Smywiński-Pohl. 2025. On the path to make Ukrainian a high-resource language. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 120–130, Vienna, Austria (online). Association for Computational Linguistics.

Mykola Haltiuk and Aleksander Smywiński-Pohl. 2025. Model-aware tokenizer transfer. *Preprint*, arXiv:2510.21954.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland,

Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. Training compute-optimal large language models. *Preprint*, arXiv:2203.15556.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Seungduk Kim, Seungtaek Choi, and Myeongho Jeong. 2024. Efficient and effective vocabulary expansion towards multilingual large language models. *Preprint*, arXiv:2402.14714.

Artur Kiulian, Anton Polishko, Mykola Khandoga, Yevhen Kostiuk, Guillermo Gabrielli, Łukasz Gagała, Fadi Zaraket, Qusai Abu Obaida, Hrishikesh Garud, Wendy Wing Yee Mak, Dmytro Chaplynskyi, Selma Amor, and Grigol Peradze. 2025. From English-centric to effective bilingual: LLMs with custom tokenizers for underrepresented languages. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 1–13, Vienna, Austria (online). Association for Computational Linguistics.

Roman Kyslyi, Nataliia Romanyshyn, and Volodymyr Sydorskyi. 2025. The unlp 2025 shared task on detecting social media manipulation. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 105–111, Vienna, Austria (online). Association for Computational Linguistics.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, and 40 others. 2025. Datacomp-lm: In search of the next generation of training sets for language models. *Preprint*, arXiv:2406.11794.

Xinjian Li, Shinnosuke Takamichi, Takaaki Saeki, William Chen, Sayaka Shiota, and Shinji Watanabe. 2023. Yodas: Youtube-oriented dataset for audio and speech. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. Fineweb-edu: the finest collection of educational content.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.

James Cross Onur Çelebi Maha Elbayad Kenneth Heafield Kevin Heffernan Elahe Kalbassi Janice Lam Daniel Licht Jean Maillard Anna Sun Skyler Wang Guillaume Wenzek Al Youngblood Bapi Akula Loic Barrault Gabriel Mejia Gonzalez Prangthip Hansanti John Hoffman Semarley Jarrett Kaushik Ram Sadagopan Dirk Rowe Shannon Spruit Chau Tran Pierre Andrews Necip Fazil Ayan Shruti Bhosale Sergey Edunov Angela Fan Cynthia Gao Vedanuj Goswami Francisco Guzmán Philipp Koehn Alexandre Mourachko Christophe Ropers Safiyyah Saleem Holger Schwenk Jeff Wang NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.

Krzysztof Ociepa, Łukasz Flis, Krzysztof Wróbel, Adrian Gwoździej, and Remigiusz Kinas. 2025. Bielik 7b v0.1: Polish language model - development, insights, and evaluation. *Computer Science*, 26(4).

Yurii Paniv. 2023. Ualpaca: Ukrainian alpaca dataset. `https://github.com/robinhad/kruk`. A Ukrainian instruction-following dataset containing 52,002 examples.

Yurii Paniv. 2025. Isolating LLM performance gains in pre-training versus instruction-tuning for mid-resource languages: The Ukrainian benchmark study. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI*

Era, pages 876–883, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Yurii Paniv, Artur Kiulian, Dmytro Chaplynskyi, Mykola Khandoga, Anton Polishko, Tetiana Bas, and Guillermo Gabrielli. 2025. Benchmarking multimodal models for Ukrainian language understanding across academic and cultural domains. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 14–26, Vienna, Austria (online). Association for Computational Linguistics.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language. *Preprint*, arXiv:2506.20920.

QIRIM. 2025. QIRIM/crh_web · Datasets at Hugging Face — huggingface.co. https://huggingface.co/datasets/{Q}{I}{R}{I}{M}/crh_web. [Accessed 01-02-2026].

Vladyslav Radchenko and Nazarii Drushchak. 2025. Improving named entity recognition for low-resource languages using large language models: A Ukrainian case study. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 27–35, Vienna, Austria (online). Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *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)*, pages 6490–6500, Online. Association for Computational Linguistics.

Yehor Smoliakov. 2025. ua-l/questions-with-answers · Datasets at Hugging Face — huggingface.co. https://huggingface.co/datasets/ua-l/questions-with-answers. [Accessed 01-02-2026].

Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2025. Nemotron-CC: Transforming Common Crawl into a refined long-horizon pretraining dataset. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2459–2475, Vienna, Austria. Association for Computational Linguistics.

Oleksiy Syvokon, Olena Nahorna, Pavlo Kuchmiichuk, and Nastasiia Osidach. 2023. UA-GEC: Grammatical error correction and fluency corpus for the Ukrainian language. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 96–102, Dubrovnik, Croatia. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Teknium. 2023. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. *arXiv preprint*.

VoxCheck. 2018. Home Eng — voxcheck.voxukraine.org. https://voxcheck.voxukraine.org/home-eng.html. [Accessed 01-02-2026].

Yongliang Wu, Yizhou Zhou, Zhou Ziheng, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, and Xu Yang. 2025. On the generalization of sft: A reinforcement learning perspective with reward rectification. *Preprint*, arXiv:2508.05629.

Yunhui Xia, Wei Shen, Yan Wang, Jason Klein Liu, Huifeng Sun, Siyue Wu, Jian Hu, and Xiaolong Xu. 2025. Leetcodedataset: A temporal dataset for robust evaluation and efficient training of code llms. *Preprint*, arXiv:2504.14655.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. *Preprint*, arXiv:2305.11206.

A Tokenizer Replacement Statistics

Writing system	Tokens removed	Tokens retained
Han (Chinese)	16,488	4,122
Devanagari (Hindi)	10,976	2,743
Bengali	7,983	1,995
Arabic	6,730	1,682
Hiragana / Katakana (Japanese)	3,944	985
Hangul (Korean)	3,744	935
Tamil	3,080	770
Thai	1,740	435
Malayalam	1,566	391
Telugu	1,428	356
Gujarati	1,080	270
Kannada	1,016	253
Ethiopic	691	172
Hebrew	670	167
Khmer	481	119
Sinhala	435	108
Myanmar	410	102
Lao	243	60
Gurmukhi	215	53
Tibetan	107	26
Oriya	100	25
Cyrillic	13,398	0
Gemma-3 <unused-*>	6,139	102

Table 8: Writing systems whose Gemma-3 tokens we partially or fully reallocated when constructing our tokenizer

Tokenizer	uk	en	EU (es/fr/it/de)	crh (Cyr.)	ru	bg	be
Qwen/Qwen3-8B	3.686	1.296	1.996	4.259	2.728	2.971	4.022
Llama-3.1-8B-Instruct	2.499	1.274	1.928	3.954	2.467	2.669	3.480
Phi-4-mini-instruct	2.596	1.256	1.691	3.209	2.158	2.296	2.771
Aya-Expanse-8B	2.226	1.309	1.782	3.541	2.187	2.523	3.273
Gemma-3-12B-it	2.506	1.307	1.788	3.341	2.245	2.329	3.045
Initial adapted tokenizer	1.628	1.308	1.788	2.356	2.556	2.514	3.057

Table 9: Average tokens-per-word (*toks/word*) for several multilingual tokenizers on Ukrainian (uk), English (en), pooled EU languages (es/fr/it/de), Crimean Tatar in Cyrillic (crh), Russian (ru), Bulgarian (bg), and Belarusian (be).

A.1 Additional Tokenizer Efficiency Metrics

In addition to the replacement statistics above, Table 9 summarises average tokens-per-word (*toks/word*) for several popular multilingual tokenizers and for our adapted tokenizer on seven corpora: Ukrainian Malyuk dataset (Chaplynskyi, 2025), English (C4-en), pooled EU languages (C4-es/fr/it/de) (Dodge et al., 2021), Crimean Tatar in Cyrillic (QIRIM), Russian, Bulgarian, and Belarusian (QIRIM, 2025).

Note. Table 9 reports metrics for the initial adapted tokenizer, not for the exact tokenizer used in OUR MODEL . Both tokenizers follow the same vocabulary-surgery procedure, but they are trained on different Ukrainian corpora: initial adapted tokenizer uses the Malyuk Ukrainian corpus plus the Cyrillic slice of the QIRIM Crimean Tatar corpus,while our model's tokenizer is trained on the full Kobza corpus plus the same Cyrillic slice of QIRIM.Therefore, the exact numbers for our model will differ, but the overall behaviour is expected to be similar.

B Author Contributions

Name	Activity
Dmytro Chaplynskyi	Pretraining Data Collection, Translation Dataset
Bohdan Didenko	Tokenizer Patching, Model Training, Hyperparameter Tuning
Nazarii Drushchak	Alignment
Mykola Haltiuk	Pretraining Data Collection, Tokenizer Transfer, Model Training
Vladyslav Humennyy	Model Persona
Andrian Kravchenko	Alignment Dataset Creation, Synthetic Data Generation for Alignment
Roman Kyslyi	Coding Models, Math Tasks, Function Calling, RL Environments
Viktoriia Makovska	Alignment, Model Unlearning
Artem Orlovskyi	Alignment Dataset Creation, Synthetic Data Generation for Alignment, Function Calling Benchmarks
Yurii Paniv	Paper Writing, Pretraining Data Filtering, Data Quality Estimation, Model Evaluation, Instruction-Tuning Dataset Translation, Model Training
Mariana Romanyshyn	Paper Writing, Alignment
Bohdan Ruban	Synthetic Data Generation for Instruction-Tuning (Visual and Text-only)
Maksym-Yurii Rudko	Audio Data Processing for Pretraining
Anastasiia Senyk	Templated Instruction-Tuning Dataset Generation

Table 10: Author contributions; activity indicates author's contribution to the paper