Towards Open Foundation Language Model and Corpus for Macedonian: A Low-Resource Language

Stefan Krsteski[1], Matea Tashkovska[*,1], Borjan Sazdov[*,2,3], Hristijan Gjoreski[2,3], Branislav Gerazov[2],

[1]École Polytechnique Fédérale de Lausanne (EPFL), Switzerland,
[2] Faculty of Electrical Engineering and Information Technologies, UKIM, North Macedonia,
[3]Emteq Ltd., Brighton, United Kingdom,

[*]Equal contribution

Correspondence: stefan.krsteski@epfl.ch

Abstract

The increase in technological adoption worldwide comes with demands for novel tools to be used by the general population. Large Language Models (LLMs) provide a great opportunity in this respect, but their capabilities remain limited for low-resource languages, restricting applications in countries where such languages are spoken. We create several resources to facilitate the adoption of LLMs and to support research advancements for Macedonian. We collect the largest Macedonian corpus to date, consisting of 40GB of textual data and totaling 3.5B words. To support conversational applications, we collect a 106k-instance instruction dataset, carefully built to be culturally grounded. For evaluation, we construct a Macedonian evaluation suite covering seven benchmarks. Finally, we train *domestic-yak*, a state-of-the-art 8B-parameter model, on our curated datasets and evaluate it against eight baseline models using the newly constructed benchmark suite. Our model outperforms all existing models in the 8B parameter range across all benchmarks, and achieves performance comparable to models up to 10× larger. Furthermore, a qualitative analysis with native speakers reveals that our model is preferred over larger counterparts, receiving higher ratings for grammatical correctness and cultural appropriateness. All datasets, code, and model weights are openly released, setting a foundation for advancing LLMs in similarly underrepresented languages. These resources are publicly available at `github.com/LVSTCK` for source code, and at `huggingface.co/LVSTCK` for pretrained model weights and data.

1 Introduction

As Large Language Models (LLMs) continue to transform modern natural language processing (NLP), the benefits of these advances remain disproportionately concentrated among high-resource languages (Joshi et al., 2020). With over 7,000 languages spoken globally, most remain severely underrepresented in the training data that powers these models, limiting access to AI for billions of people worldwide (Blasi et al., 2021).

Despite the development of multilingual variants aimed at addressing this disparity, significant challenges remain for low-resource languages. These models often lack the depth of understanding necessary for high-quality performance across all languages they claim to support. This issue is particularly pronounced in languages with smaller speaker populations, such as Macedonian, which belongs to the Eastern South Slavic branch and is spoken by over 1.1 million native speakers (State Statistical Office of North Macedonia, 2022). The fundamental relationship between data quantity and model performance means that languages with limited representation in training corpora inevitably experience degraded results (Kaplan et al., 2020). In addition, the absence of standardized evaluation benchmarks makes progress difficult to measure.

In this work, we present a thorough approach to advancing Macedonian NLP through the development of several language-specific resources:

1. The largest Macedonian corpus to date is collected, consisting of over 3.5 billion words aggregated from four existing and eight newly collected sources.

2. A novel Macedonian instruction-tuning dataset is constructed, featuring multi-turn dialogue, synthetic, commonsense, and logical reasoning examples, refined through human feedback and LLM-assisted filtering.

3. We introduce **domestic-yak**, an 8B-parameter foundation language model for Macedonian, offering both pretrained and instruction-tuned variants. It outperforms existing models in its class and achieves performance comparable to models 10× larger.

4. We create an evaluation benchmark designed to assess model performance in Macedonian across multiple tasks, including commonsense reasoning and reading comprehension.

Our results demonstrate that targeted, language-specific development can significantly help in increasing performance. By open-sourcing all data, code, and model weights, we hope to contribute both immediate value to the Macedonian-speaking community and a reproducible blueprint for similar efforts in other languages.

The paper is structured as follows: Section 2 reviews related work. Section 3 details our data collection methodology. Section 4 describes the model training setup. Section 5 presents the evaluation framework, covering both quantitative and qualitative setups. Section 6 reports and discusses the results. Section 7 ends the paper with a conclusion.

2 Related Work

Corpora for Multilingual Models. A key factor contributing to the success of LLMs in English has been the wide availability of high-quality text resources, as performance improvements correlate strongly with both corpus size and quality (Kaplan et al., 2020). Large-scale English corpora such as Common Crawl[1], The Pile (Gao et al., 2020), and C4 (Raffel et al., 2020) have provided the scale and variety needed to train increasingly capable models. These datasets offer large volumes of text and cover many domains, styles, and linguistic features, making them effective for pretraining and thereby enabling better knowledge transfer to downstream tasks.

In contrast, many low-resource languages lack such large, comprehensive corpora, creating a significant barrier to the development of competitive language models. However, in recent years, there have been growing efforts to close this gap through creating multilingual datasets that aggregate content across languages and enable training at a global scale. Similar to the English-centric datasets, multilingual resources such as mC4 (Xue et al., 2020), OSCAR (Suárez et al., 2020), Fineweb2 (Penedo et al., 2024b) and HPLT-v2 (Burchell et al., 2025), have been introduced to facilitate large-scale pretraining. These datasets aim to provide a better foundation for building models that generalize across a wider range of languages and cultures.

Nevertheless, even within these multilingual collections, representation remains uneven. High-resource languages dominate the data distribution, while low-resource Slavic languages like Macedonian are often underrepresented, both in terms of quantity and quality. To address this gap, we introduce an open-source corpus designed to advance research for this underrepresented language.

Language Modeling Approaches. The availability of multilingual datasets has enabled a shift from English-centric to multilingual models. For instance, models such as GPT-4o and Llama-3 now claim native support for over 100 languages (Grattafiori et al., 2024). Beyond these efforts, researchers have investigated more efficient strategies to extend existing models to low-resource languages. Parameter-efficient approaches, such as MAD-X (Pfeiffer et al., 2020), incorporate lightweight language and task adapters to enable zero-shot transfer while training only 3–5% of the model's parameters. Alternatively, large-scale continual pretraining has been shown to introduce hundreds of new languages simultaneously, yielding strong task generalization. For instance, EMMA-500 (Ji et al., 2024), trained on 546 languages, achieves significant gains without any task-specific fine-tuning.

Alongside general multilingual models, some approaches focus on groups of closely related languages to exploit shared linguistic structure. The CroSloEngual model (Ulčar and Robnik-Šikonja, 2020), for instance, was pretrained from scratch on Croatian, Slovene, and English, aiming to support multi- and cross-lingual training across these languages. Similarly, YugoGPT (Aleksa, 2024) is a recent effort that trains the best 7B-parameter LLM for Bosnian, Croatian, and Serbian. Furthermore, the BERTić model (Ljubešić and Lauc, 2021), was trained on Bosnian, Croatian, Montenegrin, and Serbian, which are languages that form the pluricentric Serbo-Croatian language and have overlapping vocabulary and grammar. This strategy allows for efficient use of limited data while still benefiting from multilingual learning as the languages share strong structural and lexical similarities.

However, where sufficient high-quality data exists, monolingual models are also emerging as a better alternative. For instance, for Vietnamese, continued pretraining on top of multilingual backbones followed by instruction tuning led to improvements across 10 tasks over multilingual base-

lines (Truong et al., 2024). Similar trends are seen in recent monolingual models for Italian (Orlando et al., 2024), Arabic (Koubaa et al., 2024), and Finnish (Luukkonen et al., 2023).

Macedonian remains underrepresented, with only a single publicly available language model to date[2]. We address this issue with our work and present a new large-scale language model for Macedonian, providing both pretrained and fine-tuned versions.

3 Data

In this section, we present two contributed datasets: a Macedonian corpus and an instruction dataset designed to elicit chat capabilities. We describe their properties and explain how they were collected and prepared.

3.1 Macedonian Corpus

To construct our corpus, we combine well-established sources with newly published data that have remained unexploited in Macedonian NLP research. These new sources include academic publications, educational materials spanning elementary to university levels, and various text-rich documents, typically available as PDFs on the web. The sources used are described in detail below and summarized in Table 1.

FineWeb2 (Penedo et al., 2024b) represents one of the most popular web crawled datasets available for the non-English community. Sourced from 99 CommonCrawl snapshots that span from 2013 to 2024, the data underwent deduplication and quality filtering. For our purposes, we use only the Macedonian portion of this dataset.

HPLT-v2 (Burchell et al., 2025) provides another valuable resource in our corpus. This collection includes 193 languages and was derived from web crawls subjected to similar processing as FineWeb2. Similarly, we isolate only the Macedonian subset.

MaCoCu-mk 2.0 (Bañón et al., 2023) represents another well-known web crawl resource. The Macedonian subset was constructed by crawling the ".mk" Internet top-level domains in 2021.

Document-to-Text. Historically, pre-training data for Macedonian language models has been sourced from web crawls, as shown by the preceding collections. To expand beyond these limi-

tations, we contribute new data sources that have remained untapped to date. Several tools have recently emerged to facilitate document-to-text conversion, including *docling* (Livathinos et al., 2025), *nv-ingest* (Team, 2024), and *mmore* (Sallinen et al., 2025). In our work, we use *mmore* to extract high-quality text from a variety of document sources, particularly focusing on academic publications, educational materials, official government documents and other scanned digital resources. More information on these tools and the full list of processed sources is available in the Appendix A.1.

Wikipedia. As a standard resource in language modeling, we include the "mk" Wikipedia dump with the last update being January 2025.

SETimes Corpus (Ljubešić and Stojanovska, 2023), is a parallel corpus of news articles in the Balkan languages. In this work we use the complete Macedonian-English pair (207,777 sentence pairs; 44.6M tokens) and retain only the Macedonian side.

Common Voice (Ardila et al., 2019) is an open-source, multilingual dataset originally developed to train speech-enabled applications. It provides transcriptions in the form of natural text prompts for speakers. We extract only the Macedonian transcription text, which consists of human-validated sentences. Although not originally intended as a text corpus, it offers an unconventional but high-quality source of conversational language.

Origin	Words (B)	Percentage
HPLT-2	1.49	42.21%
FineWeb2	1.33	37.66%
MaCoCu-mk 2.0	0.49	13.92%
Documents (mmore)	0.14	4.07%
Wikipedia	0.07	1.96%
SETimes Corpus	0.004	0.13%
Common Voice	0.002	0.05%
Total	**3.53**	**100.00%**

Table 1: Sources and word distribution for the Macedonian pretraining corpus

The resulting corpus consists of 3.53 billion words. Given the significant overlap between web-based sources (particularly those derived from CommonCrawl) and recent evidence demonstrating that filtering and deduplication significantly improve language model performance (Lee et al., 2021), we implement a text filtering pipeline, closely following FineWeb2's methodology (Penedo et al., 2024b).

---

[2] https://huggingface.co/trajkovnikola/MKLLM-7B-Instruct

As an initial step, we remove Personally Identifiable Information (PII) such as email and IP addresses, and telephone numbers to comply with privacy regulations using *datatrove* (Penedo et al., 2024a). We then apply C4 filtering (Raffel et al., 2020) to discard low-quality content, including removing lines with fewer than three words or lines lacking terminal punctuation.

Furthermore, we implement Gopher filtering (Rae et al., 2021), including rejecting instances where over 90% of lines begin with bullets or where more than 30% of lines end with ellipses. We use the FastText language identification model (Joulin et al., 2016b,a) to retain only high-confidence Macedonian text (confidence > 0.65). Following this, we perform sentence-level deduplication to remove redundant content. For the newly contributed document-based data, we apply sentence chunking to segment texts into manageable units, each not exceeding 4000 words.

Finally, we use MinHash-based locality-sensitive hashing (Broder, 1997) for document-level deduplication, removing near-duplicate documents across the entire corpus. The multistage filtering pipeline resulted in 1.47 billion words of high-quality text.

3.2 Instruction Dataset

Most existing instruction datasets (Upadhayay and Behzadan, 2024) for Macedonian rely on direct translation from English, which introduces both linguistic artifacts and cultural mismatches (Bizzoni et al., 2020). To overcome these limitations, we use a hybrid construction methodology combining human supervision with model-assisted refinement. Specifically, we post-edit translated instances using GPT-4o-mini (OpenAI, 2024), by instructing it to grammatically refine the translated sentences, followed by human verification to filter low-quality samples. This process enables us to build a richer, culturally appropriate dataset while minimizing translation noise. Our final dataset integrates several sources, each selected to support specific capabilities, which we describe in details below, with the summary available in Table 2.

General Instruction Following. To support broad task coverage, we incorporate *Alpaca* (Taori et al., 2023) and *Databricks-Dolly* (Conover et al., 2023), two well-known instruction datasets. These primarily include instruction-following examples including tasks such as brainstorming, classification, closed and open question answering, gener-

ation, information extraction, and summarization. Since both datasets were produced using earlier models (e.g., GPT-3) and translated automatically, the aforementioned refinement was necessary to address issues in fluency and cultural misalignment.

Conversational Abilities. To support multi-turn conversational capabilities, we include *UltraChat 200k* (Ding et al., 2023) and *Capybara* (Daniele and Suphavadeeprasit, 2023). *UltraChat* focuses on assistant-style dialogues across a wide range of user intents, while *Capybara* focuses on multi-turn reasoning, logic and extrapolation about a wide range of subjects. These sources contribute to the conversational fluency of the final dataset.

Reasoning. To incorporate reasoning capabilities, we translated a subset of the *Open Platypus* (Lee et al., 2023) dataset, which focuses on improving logical reasoning skills in language models. This dataset mainly consists of mathematical problems that challenge the model's reasoning abilities.

Culturally Grounded Content. To address the scarcity of Macedonian-specific content and to ensure cultural relevance beyond what translated datasets could provide, we generate *synthetic data*. Using GPT-4o-mini with in-context learning, we create 3,400 culturally relevant input-output pairs across domains such as geography, history, education, science, religion, and governance. These examples are then post-processed and manually reviewed to ensure higher quality.

Origin	Words (M)	Percentage
Alpaca[†]	13.01	16.95%
Ultrachat	34.14	44.48%
Capybara	22.63	29.48%
Databricks Dolly[†]	3.38	4.40%
Open Platypus[†]	1.80	2.34%
Synthetic Data[†]	1.80	2.34%
Total	**76.76**	**100.00%**

Table 2: Source distribution of the Macedonian instruction-tuning dataset. Datasets marked with [†] were refined through model-assisted post-editing and human verification to improve fluency and cultural relevance.

The final instruction dataset contains 106,993 samples and approximately 77 million words, covering tasks such as question answering, chat conversations, mathematical reasoning, essay writing and code generation. Table 2 summarizes the dataset composition, while Appendix A.2 (Figure 3) illustrates the topic distribution.

4 Language Model Training

Our training procedure follows a two-stage approach: continued pretraining on raw text (the corpus), followed by supervised fine-tuning (SFT) on instruction data.

4.1 Continued Pretraining

In the pre-training stage, the model is optimized to predict the next token in a sequence using the standard autoregressive objective. Given a token sequence $\{x_1, \ldots, x_T\}$, the training objective is to maximize the log-likelihood:

$$\mathcal{L} = \sum_{t=1}^{T} \log P(x_t \mid x_{<t}) \qquad (1)$$

where T denotes the sequence length, x_t is the token at position t, and $x_{<t}$ represents the preceding tokens.

Rather than training from scratch on our corpus, we continue pre-training from the publicly available *Llama3.1 8B Instruct* model weights. This approach exploits the knowledge learned during the models' original multilingual training, which is especially useful for low-resource settings where data scarcity is a major bottleneck (Ji et al., 2024). We retain the original tokenizer to avoid the complexity of re-tokenization. Training spans for one epoch over the full corpus using four H100 GPUs (80 GB each 320 GB total). We use a maximum sequence length of 8,192 tokens, a cosine annealing scheduler (peak learning rate 2×10^{-5}), and the AdamW optimizer. To optimize memory usage, we set a per-device batch size of 1 and use gradient accumulation over 8 steps.

4.2 Supervised Fine-Tuning

Full fine-tuning is performed on top of our pretrained model using the instruction dataset. To make use of higher quality data, we sample with a 2:1 sampling ratio favoring human-supervised and synthetic examples over translated ones. Based on an analysis of the instruction lengths, we set the maximum sequence length to 4,096 tokens, covering over 95% of the dataset without truncation (see Appendix A.2, Figure 2). We optimize the standard cross-entropy loss over the instruction data, i.e. negative log-likelihood of the next token given the prefix. Training spans for three epochs using a single H100 GPU (80 GB). We use the AdamW optimizer with a per-device batch size of 2 and gradient accumulation over 8 steps. We double

the learning rate to 4×10^{-5} and use the same scheduling method as in the pre-training phase.

5 Evaluation Setup

5.1 Benchmarks

Similar to many other low-resource languages, Macedonian lacks a standardized evaluation benchmark, making it difficult to track progress in LLM development. To address this, we construct a Macedonian adaptation of the Language Model Evaluation Harness (Gao et al., 2021).

A natural approach would be to translate the original English benchmarks directly into Macedonian. However, as discussed in Section 3.2, translations from English tend to introduce unnatural phrasing, so called "translationese", and cultural biases, which can make the benchmarks unreliable for evaluating models in the target language. To address these issues, we instead leverage an existing high-quality benchmark adaptation available for Serbian (Gordić, 2023). Given the close linguistic and cultural affinities between these two South Slavic languages, we translate the Serbian version into Macedonian, maintaining natural phrasing and improving evaluation fidelity.

Furthermore, to preserve grammatical correctness during translation, we use a template-based strategy. Translating individual text segments (multiple-choice questions without answer options) often disrupts target language word order. To address this, we translate full sentence templates containing placeholders for answer options, then remove the placeholders post-translation. See Appendix A.4 for implementation details and examples.

In total, we translated seven benchmarks, which we use to quantitatively measure the performance of our model using accuracy as the evaluation metric. The benchmarks cover two task categories: commonsense reasoning and reading comprehension.

Commonsense Reasoning benchmarks evaluate an LLM's ability to apply everyday human-like assumptions that are not explicitly stated. This includes physical world knowledge, causal and temporal reasoning, as well as understanding of social norms and expectations. We report results on six well-known datasets (in their translated versions): *HellaSwag* (Zellers et al., 2019), *WinoGrande* (Keisuke et al., 2019), *PIQA* (Bisk et al., 2020), *OpenbookQA* (Mihaylov et al., 2018), *ARC-*

Model	Size	PIQA	OBQA	WinoG	ARC-E	ARC-C	BoolQ	HSwag	Avg.
Smaller Models									
Llama 3.2	1B	0.539	0.162	0.509	0.231	0.190	0.573	0.270	0.353
Phi-3.5-mini	3.8B	0.526	0.164	0.519	0.289	0.188	0.603	0.263	0.364
Comparable Sizes (7B–8B)									
Qwen2.5	7B	0.560	0.216	0.535	0.391	0.253	0.779	0.339	0.439
Mistral	7B	0.578	0.218	0.561	0.463	0.287	0.759	0.372	0.462
Llama 3.1	8B	0.587	0.252	0.568	0.445	0.282	0.764	0.374	0.467
MKLLM[†]	7B	0.642	0.294	0.615	0.503	0.300	0.788	0.433	0.510
domestic-yak[†]	**8B**	**0.692**	**0.302**	**0.627**	0.547	0.336	0.787	0.448	0.535
Larger Models (12B–70B)									
Mistral Nemo	12B	0.607	0.242	0.606	0.472	0.319	0.809	0.400	0.493
Llama 3.3	70B	0.660	0.282	0.609	**0.581**	**0.369**	**0.851**	**0.466**	0.545

Table 3: Performance comparison across models (all in their instruction-tuned variants), evaluated with accuracy. Benchmarks are sorted by average score (descending) within each model class. Models with explicit support for Macedonian are marked with [†]. For the remaining models, we could not confirm language coverage. Despite being over 10× smaller, our 8B model outperforms Llama 70B on 3 out of 7 benchmarks (PIQA, OBQA, WinoG). Standard deviations were consistent (0.009–0.014) and are omitted for clarity.

Easy, and *ARC-Challenge* (Clark et al., 2018).

Reading Comprehension benchmarks evaluate the ability of a model to understand a given text passage, specifically its ability to grasp context, coherence and narrative flow. We evaluate performance using the *BoolQ* dataset (Clark et al., 2019).

5.2 Qualitative Evaluation

In addition to quantitative evaluation, we conduct an analysis where we assess the quality of responses through native speaker judgments. We carry out a head-to-head comparison between our *domestic-yak* and the strongest evaluated model Llama 3.1 70B. We design ten original questions (included in Appendix A.5) that reflect everyday reasoning, culturally grounded knowledge, and typical native language use. Specific tasks include understanding common expressions, giving advice, writing informal messages, and answering questions about local institutions. For each question, native speakers evaluate the responses of both models. A total of 35 participants completed the survey, with a mean age of 28±9 years, including 19 males and 16 females. Participants evaluated each pair of responses by selecting the better answer and providing a brief justification. The available options for the justification included better gramatical consistency, more natural phrasing, higher cultural appropriateness, more information, and an open "Other" field for free-text input. Moreover, the participants rated both of the answers for fluency and relevance using a Likert scale from 1 to 5 (Likert, 1932). To

reduce bias, model outputs were anonymized and randomized across questions, with responses labeled as "Model A" and "Model B". The goal of this human evaluation is to highlight differences that are not captured by quantitative benchmarks alone.

6 Results and Discussion

6.1 Quantitative Results

We compare *domestic-yak-instruct* (8B) against eight baselines spanning three size categories: smaller (1B–4B), in-class (7B–8B), and larger (12B–70B). The results are shown in Table 3. Three key takeaways emerge from this comparison. Firstly, our model achieves the highest performance among all models of comparable size across every evaluated task, outperforming strong baselines such as Mistral, Qwen2.5, and Llama 3.1. We attribute this significant improvement to our targeted training strategies, particularly the use of the largest Macedonian corpus combined with the instruction dataset that enables the model to better capture the linguistic patterns. Secondly, *domestic-yak* outperforms larger counterparts, surpassing Mistral Nemo (12B) on all but one task, and Llama-3.3 (70B) on three of seven benchmarks (*PIQA*, *OpenBookQA*, *WinoGrande*), despite being an order of magnitude smaller. Finally, our model represents a significant improvement compared to the previous best Macedonian model, MKLLM, achieving higher accuracy across six out of seven benchmarks. In summary, *domestic-yak* sets a new

state-of-the-art result for the Macedonian language and marks a significant step forward for NLP in this domain, laying the foundation for a full suite of models that will be released in the near future.

6.2 Ablation Study

A central objective of this work is to demonstrate the effectiveness of our proposed Macedonian corpus and instruction dataset for adapting language models. To break down their impact, we run an ablation study measuring performance gains. Starting from the baseline *Llama-3.1-8B-Instruct* model, we incrementally apply (*i*) continued pre-training on our Macedonian corpus (*domestic-yak-base*), and (*ii*) supervised fine-tuning on the instruction dataset (*domestic-yak-instruct*). Table 4 reports the results, isolating the effects of domain-specific pretraining and instruction tuning.

The pre-training phase provides the majority of gains, increasing the average score from 0.47 to 0.52. Improvements are consistent across all Commonsense Reasoning benchmarks, with *PIQA* (+8), *ARC Easy* (+7), and *HellaSwag* (+6) among the highest. In contrast, the Reading Comprehension benchmark (*BoolQ*) shows only a marginal (+1) improvement. Since many of the tasks with larger improvements primarily test factual recall, this pattern suggests that continued pretraining is very effective at enhancing the model's factual knowledge. Meanwhile, skills such as contextual reading and coherence tracking appear to be well-covered by the base model, as no significant improvements were seen for that task category.

Instruction tuning provides an additional +2 points on average. It improves performance on tasks such as *ARC Easy* and *BoolQ*, but has no positive effect on *WinoGrande*, where pronoun-resolution skills (Winogrande's main task) plateau during pretraining. This limited effect is consistent across tasks, which we attribute to the strength of the base model. Since *Llama 3.1 Instruct* is already trained for general-purpose instruction following, additional fine-tuning on task-specific instructions largely acts as light alignment. It helps adapt the model to domain-specific phrasing and task format, but contributes little in terms of new capability.

6.3 Qualitative Analysis

We collected human evaluation data comparing responses from *domestic-yak-instruct* and *Llama 3.1 70B Instruct* across ten unique prompts. The analysis includes model preference counts and Likert
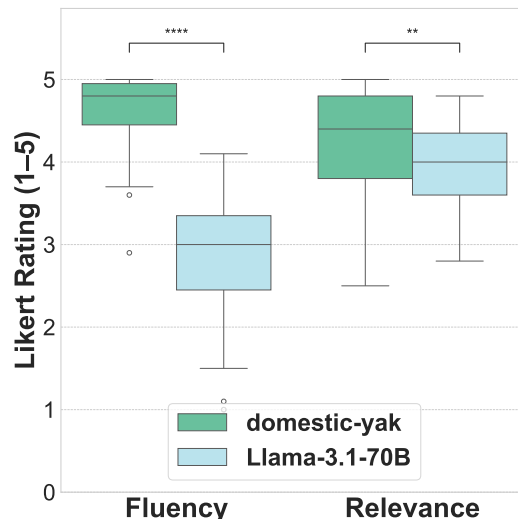

Figure 1: Average fluency and relevance Likert ratings per model. *domestic-yak-instruct* outperforms *Llama 3.1 70B Instruct* in both dimensions (Wilcoxon signed-rank test, Bonferroni corrected, $p_{\text{fluency}}=1.83\times10^{-11}$, $p_{\text{relevance}}=1.192\times10^{-3}$). Statistical significance annotations: * if $p \in [0.05, 10^{-2})$; ** if $p \in [10^{-2}, 10^{-3})$; *** if $p \in [10^{-3}, 10^{-4})$; and **** if $p \leq 10^{-4}$.

scale ratings for fluency and relevance. Overall, *domestic-yak-instruct* was preferred in 64.2% of the pairwise comparisons, while *Llama 3.1 70B Instruct* was preferred in 27.1%. In 8.7% of cases, participants rated the two responses equally.

Participants most cited better grammatical consistency (81.6%), more natural phrasing (60%), and higher cultural appropriateness (37%) as reasons for preferring our model. In Likert ratings, our model achieved average scores of 4.6 for fluency and 4.26 for relevance, compared to 2.8 and 3.9 for *Llama 3.1 70B*, respectively.

To formally test these differences, we grouped the results by participant. For each participant, we computed the number of times each model was preferred and the average Likert ratings for fluency and relevance. A Shapiro–Wilk test indicated that the distributions were not normal, so we applied the Wilcoxon signed-rank test for all comparisons, which tests if the median difference between pairs is zero. Bonferroni correction was used to adjust for multiple testing. Regarding model preference, the results demonstrate a statistically significant difference ($p = 8.56\times10^{-5}$), with participants favoring *domestic-yak-instruct* over *Llama 3.1 70B Instruct*. Similarly, as shown in Figure 1, our model significantly outperformed the baseline in both fluency ($p = 9.17\times10^{-11}$) and relevance ($p = 7.26\times10^{-3}$)

Task (mk)	Llama 3.1	domestic-yak-base (+ Δ_1)	domestic-yak-instruct (+ Δ_2)	total Δ
ARC Easy	0.45	0.52 (+0.07)	0.55 (+0.03)	**+0.10**
ARC Challenge	0.28	0.32 (+0.04)	0.34 (+0.02)	**+0.06**
BoolQ	0.76	0.77 (+0.01)	0.79 (+0.02)	**+0.03**
HellaSwag	0.37	0.43 (+0.06)	0.45 (+0.02)	**+0.08**
Openbook QA	0.25	0.29 (+0.04)	0.30 (+0.01)	**+0.05**
PIQA	0.59	0.67 (+0.08)	0.69 (+0.02)	**+0.10**
WinoGrande	0.57	0.63 (+0.06)	0.63 (+0.00)	**+0.06**
Average	**0.47**	**0.52** (+0.05)	**0.54** (+0.02)	**+0.07**

Table 4: Ablation study on the effects of pre-training and instruction tuning. **Llama 3.1** is the base model. **domestic-yak-base** is a result from continued pretraining using our corpus, and **domestic-yak-instruct** adds instruction tuning. We report stepwise improvements inline in green, and total gains are highlighted in **bold green**.

Although Llama 70B achieved higher scores on several quantitative benchmarks (Table 3), our model was highly preferred by native speakers during qualitative evaluation. This demonstrates that benchmark scores do not fully capture the whole story, i.e. real-world, language-specific model quality. By continuing pretraining on high-quality data and applying instruction tuning across a broad range of tasks, including general instruction following, culturally grounded content, reasoning and conversational skills, our model learned the linguistic and cultural characteristics of the Macedonian language crucial for native speakers. The qualitative results confirm that our model surpasses a model nearly ten times larger in fluency, relevance, and overall preference among native speakers, proving that careful adaptation can rival scale (see Appendix A.5 for example responses).

7 Conclusion

In this work, we bridge the gap in Macedonian NLP by introducing a suite of language-specific resources and demonstrating the effectiveness of focused monolingual adaptation in low-data settings. We release the largest Macedonian corpus (3.5B+ words), a cleaned version of the said dataset (1.5B+ words), a conversational instruction-tuning dataset, and a standardized evaluation benchmark spanning commonsense reasoning, factual knowledge, and reading comprehension. Using these resources, we train and release *domestic-yak*, an 8B-parameter model that outperforms existing baselines and matches or surpasses multilingual models up to ten times larger across tasks.

Ablations highlight the importance of continued monolingual pretraining, which resulted in greater gains than instruction tuning alone, emphasizing the value of high-quality, language-specific data. Human evaluations further strengthen our findings:

native speakers consistently preferred *domestic-yak-instruct* over the *Llama 3.1 70B Instruct*, rating it significantly higher for fluency, grammatical accuracy, and cultural relevance.

Our results prove that targeted resource development and monolingual adaptation enable smaller models to outperform larger multilingual systems in real-world applications. All datasets, benchmarks, and model weights are publicly released to accelerate Macedonian NLP research and applications. Future work will expand the benchmark to include broader task coverage and address the current 4k context-length limitation to support applications requiring larger windows. We also plan to incorporate additional datasets, such as COPA-MK (Ljubešić et al., 2022), a Macedonian translation of the Choice of Plausible Alternatives (COPA) benchmark (Ponti et al., 2020), as well as resources from the OPUS collection (Tiedemann, 2012) to further improve model robustness and evaluation depth. We hope this work offers a blueprint for revitalizing other low-resource languages through targeted efforts, free from the constraint of scale.

Limitations

We identify three main limitations in our work. First, while the model performs well on general-purpose tasks, it has not been evaluated nor adapted for niche domains such as law, medicine, or finance. Performance in these areas is likely to be limited due to the lack of domain-specific data. Accordingly, we position this release as a general-purpose foundation and encourage the community to pursue fine-tuning and evaluation in specialized domains.

Furthermore, the model uses a maximum context window of 8,192 tokens during pretraining and 4,096 tokens during instruction tuning. This limits its ability to handle tasks that require longer context, such as multi-document summarization or

long-form QA. We believe that addressing this limitation should be the key focus of future work, both in data collection and model training processes.

Finally, we note as a minor limitation the lack of Macedonian benchmarks, which required us to rely on translated datasets. This introduces variance that can negatively affect the accuracy of Macedonian-specific quantitative evaluation, even though we took steps to reduce it. Nevertheless, comparison is made against the same datasets, so this does not significantly reduce the confidence in the presented results.

References

Gordić Aleksa. 2024. Yugogpt - an open-source LLM for Serbian, Bosnian, and Croatian languages.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670.*

Marta Bañón, Malina Chichirau, Miquel Esplà-Gomis, Mikel L Forcada, Aarón Galiano-Jiménez, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, et al. 2023. Macedonian-English parallel corpus MaCoCu-mk-en 2.0.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about Physical Commonsense in Natural Language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence.*

Yuri Bizzoni, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How Human is Machine Translationese? Comparing Human and Machine Translations of Text and Speech. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 280–290, Online. Association for Computational Linguistics.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2021. Systematic inequalities in language technology performance across the world's languages. *arXiv preprint arXiv:2110.06733.*

Andrei Z. Broder. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pages 21–29. IEEE.

Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Erik Henriksson, et al. 2025. An Expanded Massive Multilingual Dataset for High-Performance Language Technologies. *arXiv preprint arXiv:2503.10267.*

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the Surprising Difficulty of Natural Yes/No Questions. In *NAACL.*

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv:1803.05457v1.*

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM.

Luigi Daniele and Suphavadeeprasit. 2023. Amplify-Instruct: Synthetically Generated Diverse Multi-turn Conversations for efficient LLM Training. *arXiv preprint arXiv:(coming soon).*

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing Chat Language Models by Scaling High-quality Instructional Conversations. *Preprint*, arXiv:2305.14233.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv preprint arXiv:2101.00027.*

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Aleksa Gordić. 2023. First Serbian LLM Evaluation. Report.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783.*

Shaoxiong Ji, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, et al. 2024. Emma-500: Enhancing massively multilingual adaptation of large language models. *arXiv preprint arXiv:2409.17892.*

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. *arXiv preprint arXiv:2004.09095.*

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651.*

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of Tricks for Efficient Text Classification. *arXiv preprint arXiv:1607.01759*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Sakaguchi Keisuke, Le Bras Ronan, Bhagavatula Chandra, and Choi Yejin. 2019. WinoGrande: An Adversarial Winograd Schema Challenge at Scale.

Anis Koubaa, Adel Ammar, Lahouari Ghouti, Omar Najar, and Serry Sibaee. 2024. Arabiangpt: Native arabic gpt-based large language model. *arXiv preprint arXiv:2402.15313*.

Ariel N. Lee, Cole J. Hunter, and Nataniel Ruiz. 2023. Platypus: Quick, Cheap, and Powerful Refinement of LLMs.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2021. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Nikolaos Livathinos, Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Kasper Dinkla, Yusik Kim, et al. 2025. Docling: An Efficient Open-Source Toolkit for AI-driven Document Conversion. *arXiv preprint arXiv:2501.17887*.

Nikola Ljubešić, Boshko Koloski, Kristina Zdravkovska, and Taja Kuzman. 2022. Choice of plausible alternatives dataset in Macedonian COPA-MK. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić and Davor Lauc. 2021. Bertić–The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. *arXiv preprint arXiv:2104.09243*.

Nikola Ljubešić and Biljana Stojanovska. 2023. Macedonian linguistic training corpus SETimes. MK 0.1.

Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, et al. 2023. FinGPT: Large generative models for a small language. *arXiv preprint arXiv:2311.05640*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *EMNLP*.

OpenAI. 2024. Gpt-4o System Card. *Preprint*, arXiv:2410.21276.

Riccardo Orlando, Luca Moroni, Pere-Lluís Huguet Cabot, Simone Conia, Edoardo Barba, Sergio Orlandini, Giuseppe Fiameni, and Roberto Navigli. 2024. Minerva LLMs: The first family of Large Language Models trained from scratch on Italian data. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 707–719.

Guilherme Penedo, Hynek Kydlíček, Alessandro Cappelli, Mario Sasko, and Thomas Wolf. 2024a. Datatrove: large scale data processing.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Martin Jaggi, Leandro von Werra, and Thomas Wolf. 2024b. FineWeb2: A sparkling update with 1000s of languages.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. *arXiv preprint arXiv:2005.00052*.

Edoardo M. Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Alexandre Sallinen, Stefan Krsteski, Paul Teiletche, Allard Marc-Antoine, Baptiste Lecoeur, Michael Zhang, Fabrice Nemo, David Kalajdzic, Matthias Meyer, and Mary-Anne Hartley. 2025. Mmore: Massive multimodal open rag & extraction. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*. To appear.

State Statistical Office of North Macedonia. 2022. Census of Population, Households and Dwellings in the Republic of North Macedonia, 2021.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. *arXiv preprint arXiv:2006.06202*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

NVIDIA Ingest Development Team. 2024. *NVIDIA Ingest: An accelerated pipeline for document ingestion*.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Sang T Truong, Duc Q Nguyen, Toan Nguyen, Dong D Le, Nhi N Truong, Tho Quan, and Sanmi Koyejo. 2024. Crossing linguistic horizons: Finetuning and comprehensive evaluation of vietnamese large language models. *arXiv preprint arXiv:2403.02715*.

M. Ulčar and M. Robnik-Šikonja. 2020. FinEst BERT and CroSloEngual BERT: less is more in multilingual models. In *Text, Speech, and Dialogue TSD 2020*, volume 12284 of *Lecture Notes in Computer Science*. Springer.

Bibek Upadhayay and Vahid Behzadan. 2024. TaCo: Enhancing Cross-Lingual Transfer for Low-Resource Languages in LLMs through Translation-Assisted Chain-of-Thought Processes. In *5th Workshop on practical ML for limited/low resource settings, ICLR*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

A Appendix

A.1 Document-to-Text

We mention in our main text that a significant portion of our corpus was collected using document-to-text tools. Recently, such tools are well established in the community and enable text extraction from diverse file formats (PDF, DOCX, PPTX, and more). In our work, we use a tool called *mmore* (Sallinen et al., 2025), a distributed pipeline similar to IBM's *Docling* (Livathinos et al., 2025). The most useful feature of these tools is the ability to parse scanned documents, which we found was very valuable given that digitization in North Macedonia lags behind, and many available sources are scanned copies. Table 5 lists the sources processed using *mmore*. All entities were contacted directly, and we obtained proper approval to use materials from each of them.

Source	Origin
Ss. Cyril and Methodius University in Skopje	`https://ukim.edu.mk/en/`
Macedonian Academy of Sciences and Arts	`https://manu.edu.mk/`
St. Clement of Ohrid University of Bitola	`https://uklo.edu.mk/?lang=en`
Goce Delčev University of Štip	`https://www.ugd.edu.mk/en/home/`
Institute of Macedonian Language	`http://imj.ukim.edu.mk/`
Official PE Gazette of North Macedonia	`https://www.slvesnik.com.mk/`

Table 5: Macedonian Sources Processed with the Document-to-Text Pipeline

A.2 Data for Instruction Model

Figure 3 shows the composition of our instruction dataset across four high-level categories. The dataset is heavily dominated by *question answering* and *chat-style interactions*, which together account for over 80% of all examples. A smaller portion is dedicated to *reasoning tasks* and more open-ended formats such as *code generation* and *essay writing*, which help diversify the model's capabilities beyond straightforward instruction following.

Furthermore, Figure 2 presents the token length distribution across the dataset. The majority of samples (97.4%) fall below the 4,096-token cutoff used during supervised fine-tuning, ensuring that most examples are used without truncation.
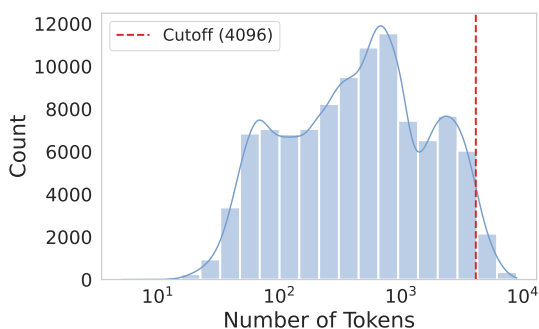

Figure 2: Token length distribution in the SFT dataset. The red dashed line indicates the 4,096-token cutoff, which covers 97.4% of all samples.
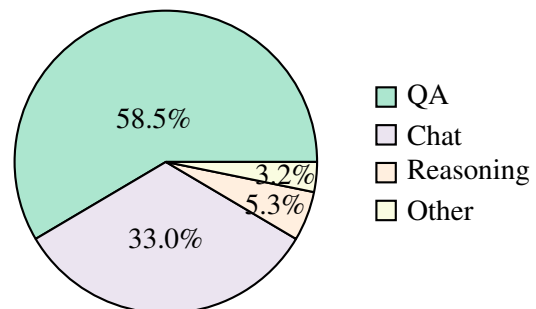

Figure 3: Distribution of Topics in the Instruction Dataset. Question Answering tasks comprise the majority (58.5%), followed by Chat Conversations (33.0%), with Reasoning and Other categories making up smaller portions (5.3% and 3.2% respectively).

A.3 System Prompt

The system prompt that was used to train the instruction model is given below in both its original and English form.

> **System Prompt:**
>
> *Macedonian:* Ти си виртуелен асистент кој помага на корисници на македонски јазик. Одговарај на прашања на јасен, разбирлив и професионален начин. Користи правилна граматика и обиди се одговорите да бидат што е можно покорисни и релевантни.
>
> ---
>
> *English:* You are a virtual assistant that helps users in the Macedonian language. Answer questions in a clear, understandable, and professional manner. Use correct grammar and try to make your responses as helpful and relevant as possible.

A.4 Translation

To preserve grammatical structure during translation of multiple-choice questions, we implement a *template-based translation strategy*. Unlike naïve translation of isolated queries - which often produces grammatically flawed outputs - our approach maintains syntactic integrity through contextual grounding. Below we show the reason we went for this approach by using an example from the Serbian version of the ARC-Easy benchmark.

> **Example Instance from ARC-Easy:**
>
> **Original (Serbian):** Hladnokrvne životinje su često
> **Choices (Serbian):** ["brze", "velike", "bez dlake", "spore"]
> **Gloss (English):** Cold-blooded animals are often
> **Choices (English):** ["fast", "large", "hairless", "slow"]

The naïve translation produces a grammatically awkward construction with syntactically incorrect word order, primarily due to missing subject-verb-object agreement.

> **Naïve Translation (Incorrect)**
>
> **Translation (Macedonian):** Често се ладнокрвни животни
> **Gloss (English):** Often are cold-blooded animals

To mitigate this issue, we insert a placeholder in place of the answer choice during translation, which is removed after processing. In addition to ensuring correct translation, this approach also helps prevent potential data leakage that could arise from choice-dependent translations.

> **Template-Based Translation (Correct)**
>
> **Original (Serbian):** Hladnokrvne životinje su često ___
> **Translation (Macedonian):** Ладнокрвните животни често се ___
> **Gloss (English):** Cold-blooded animals are often ___

This method ensures that sentence structure remains intact, avoiding artifacts introduced by out-of-context or partial sentence translations.

A.5 Survey Questions for Qualitative Evaluation

To assess the stylistic and cultural quality of model responses, we designed a set of ten original prompts reflecting everyday reasoning, communication, and local knowledge. Participants evaluated responses to the following questions (presented in English below):

> **Survey Questions:**
>
> 1. What does it mean when someone says "It's not going well for me"?
>
> 2. How many years is compulsory education in North Macedonia?
>
> 3. I have a headache. What medicine is recommended for relieving a headache?
>
> 4. Tell me a good joke.
>
> 5. Tell me a well-known Macedonian proverb.
>
> 6. Write a short email postponing a meeting with a professor due to justified reasons.
>
> 7. How would you respond to a friend who invites you for coffee, but you are busy?
>
> 8. What should you do if you lose your ID card?
>
> 9. What are the most famous tourist attractions in North Macedonia?
>
> 10. If the bus is late and you have an important meeting, what is the smartest thing to do?
>
> *Participants were asked to evaluate responses based on fluency and relevance.*

To illustrate, we show example answers to Question 5 ("Tell me a well-known Macedonian proverb") generated by the two models:

> **Example Responses:**
>
> **domestic-yak-instruct:**
> Една позната македонска изрека е: „Кој рано рани, две среќи граби".
>
> **Gloss (English):** A well-known Macedonian proverb is: "The one who rises early seizes two fortunes."
>
> ---
>
> **Llama 70B Instruct:**
> Една поговорка која ми доаѓа на ум е: "Молчи и слушај, пајакот на ѕидот чувај."Оваа изрека значи дека човек треба да биде внимателен и да слуша, а не само да зборува. Слично на тоа, и другата изрека вели: "Ушите најпрво чујат, а потоа устата зборува-ат."Овие изреки ни потсетуваат на важноста на слушањето и вниманието во нашиот секојден живот.
>
> **Gloss (English):** One proverb that comes to mind is: "Be silent and listen, the spider on the wall beware." This proverb means that one should be attentive and listen, not just talk. Similarly, another proverb says: "The ears first hear, and then the mouth speaking." These proverbs remind us of the importance of listening and attention in our daily lives.