From Bytes to Borsch: Fine-Tuning Gemma and Mistral for the Ukrainian Language Representation

Artur Kiulian[1], Anton Polishko[1], Mykola Khandoga[1,2],
Oryna Chubych[1], Jack Connor[3], Raghav Ravishankar[4],
Adarsh Shirawalmath[4]
[1]PolyAgent, [2]Mindee, [3]O'Shaughnessy Ventures, [4]Tensoic.
Email: a@polyagent.co, anton@polyagent.co, mkhandoga@gmail.com, oryna.chubych@gmail.com,
jack.connor83@gmail.com, raghav@tensoic.com, adarsh@tensoic.com

Abstract

In the rapidly advancing field of AI and NLP, generative large language models (LLMs) stand at the forefront of innovation, showcasing unparalleled abilities in text understanding and generation. However, the limited representation of low-resource languages like Ukrainian poses a notable challenge, restricting the reach and relevance of this technology. Our paper addresses this by fine-tuning the open-source Gemma and Mistral LLMs with Ukrainian datasets, aiming to improve their linguistic proficiency and benchmarking them against other existing models capable of processing Ukrainian language. This endeavor not only aims to mitigate language bias in technology but also promotes inclusivity in the digital realm. Our transparent and reproducible approach encourages further NLP research and development. Additionally, we present the Ukrainian Knowledge and Instruction Dataset (UKID) to aid future efforts in language model fine-tuning. Our research not only advances the field of NLP but also highlights the importance of linguistic diversity in AI, which is crucial for cultural preservation, education, and expanding AI's global utility. Ultimately, we advocate for a future where technology is inclusive, enabling AI to communicate effectively across all languages, especially those currently underrepresented.

Keywords: Gemma 2b, Gemma 7b, Mistral 7b, LLM, Ukrainian, Multilingual Models, LoRA, Fine-Tuning

1. Introduction

The field of Natural Language Processing (NLP) is expanding extremely quickly today, largely due to the immense success of the generative Large Language Models (LLM). Within only a few years, these language models have become capable of performing tasks like contextual understanding and generation, few-shot learning, automated question answering, sentiment analysis, emotion detection, and many others with unprecedented quality.

1.1. Background

The significance of recent NLP advances, obtained in such a short time, becomes even more evident looking back at the long history of quantitative language modeling. The first attempts to attack the problem of computational linguistics date back as far as 70 years ago, to the early 1950s (Shannon, 1951).

But it was not until the 2000s when the artificial Neural Network (NN) proved its effectiveness in the field (Bengio et al., 2000), notably applied to the machine translation problem (Schwenk et al., 2006). These models were mostly based on Recurrent Neural Networks (RNN) architecture like Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and later Gated Recurrent Unit (GRU) (Chung et al., 2014). Still, important milestones were achieved during this period like the introduction of word embeddings.

However, throughout most of the 2010s, while other fields of Deep Learning (DL) like Computer Vision (CV) and Reinforced Learning (RL) have achieved very impressive results (Krizhevsky et al., 2012; He et al., 2015; Silver et al., 2016), the NN-powered NLP field still suffered from a number of problems. This included the handling of long-term dependencies, capturing bidirectional context and overall difficulties with computational efficiency and stability.

The breakthrough came with the invention of the transformer architecture which introduced the key component: the attention mechanism (Vaswani et al., 2017).

1.2. The transformer era

The attention mechanism addresses the challenges of understanding both the immediate and broader context of words in a sentence, solving issues related to bidirectional context, long-term dependencies, and convergence. Furthermore transformer architecture enhances the ability to process data in parallel, significantly outperforming RNNs in this regard. This advancement has paved the way for the development of LLMs: highly complex language models with billions of parameters, trained on extensive corpora of text.

The early LLMs like Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al.,

2019) and its successors have focused on understanding text and problems like text classification, emotion recognition, etc. Although, with the emergence of the Generative Pre-trained Transformer (GPT) family (Radford and Narasimhan, 2018; Radford et al., 2019; Brown et al., 2020; OpenAI, 2023), focus has shifted towards generative tasks.

Training an LLM from scratch remains a cumbersome and costly task. Nevertheless the general nature of the training corpora allows them to fully benefit from transfer learning, implementing the *pre-training and fine-tuning* paradigm: once a model is pre-trained on a large language corpus it can be further fine-tuned for a specific use-case, requiring relatively minor costs.

The LLMs available on the market can be split into two groups: proprietary and open-source. Proprietary models like GPT-4 and Gemini (Team, 2023) tend to have more parameters and offer high out of the box performance in most common tasks, but their use is restricted by the providers and allows limited fine-tuning options. Open-source models like LLaMa2 (Touvron et al., 2023), Mistral (Jiang et al., 2023), Mixtral (Jiang et al., 2024) or a recent Gemma by Google (Gemma Team, 2024) offer full access to the model code and weights and impose little to no restrictions on the use of the model, making it a natural choice for fine-tuning experiments. Open-source models often come in a variety of sizes in terms of parameter number, allowing lighter models to be run on consumer-grade GPUs.

1.3. Motivation and objective

A substantial number of open-source models are available on the market today. At the same time all these models demonstrate a notable bias towards the English language due to their training conditions. The bias can manifest itself in a number of ways, including to but not limited to the following:

1. Language and cultural bias. This can impair a model's usability for non-English speakers and also perpetuate stereotypes or misunderstandings about cultures.

2. Ethical and fairness concerns. The same model may show considerably better performance with English-speaking users, leaving others with a subpar experiences.

3. Uneven knowledge representation. This can lead to a skewed representation of global knowledge, history, and perspectives, and embed these biases into the model's outputs and decision-making processes.

The bias becomes particularly prominent in non-European languages and languages that do not use a Latin alphabet.

This has naturally motivated numerous scholars and enthusiasts to put much efforts into fine-tuning open-source models, predominantly LLaMa 2, in many languages, both European (Basile et al., 2023; Vanroy, 2023) and non-European (Cui et al., 2024; Gala et al., 2024a,b; Nguyen et al., 2023; Azime et al., 2024; Kohli et al., 2023). Most of the listed articles have been published within the last months, and demonstrate great interest and involvement in solving this linguistic bias issue. The immediate benefits of having an open-source model that is fine-tuned with a certain language include:

1. Reduction or elimination of cultural bias.

2. Flexibility in use-cases, including both academic and business.

3. Preservation of rare and low-resource languages.

The effort also promotes the creation of language-specific datasets and development of the LLM-oriented ecosystem. Even when a particular model becomes obsolete, further progress is greatly facilitated by this groundwork.

1.3.1. Ukrainian sector of the LLMs

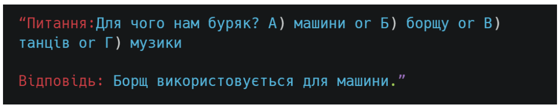
Ukraine is renowned for its dynamic IT community, which thrives both in academic circles and the commercial sector. The field of computational linguistics is no exception, boasting the inception of multi-billion dollar unicorns like Grammarly within its borders. With the advent of LLMs, there has been a keen interest in harnessing their capabilities for solving NLP challenges in the Ukrainian language.

Yet, until recently, these efforts have predominantly focused on leveraging BERT-like models (Tiutiunnyk, 2020; Laba et al., 2023; Katerynych et al., 2021), while the realm of generative LLMs has been somewhat overlooked. So far, UAlpaca is the only publicly available LLM that has been fine-tuned specifically for the Ukrainian language (Had). Likewise, instructional datasets in Ukrainian have been comparatively limited. The escalating enthusiasm for generative, GPT-style LLMs underscores the need for models attuned to Ukrainian linguistic and cultural nuances, further underlining the significance of our research endeavors.

1.3.2. Objectives

The aim of the effort presented in the current paper is multifold:

1. Create an open-source, free-to-use LLMs fine-tuned for Ukrainian language and culture thus expanding the Ukrainian presence in the NLP field.



```
“Питання:Для чого нам буряк? А) машини or Б) борщу or В)
танців or Г) музики

Відповідь: Борщ використовується для машини.”
```

Figure 1: Example of erroneous model inference in Ukrainian.

2. Compare the performance of different open-source LLMs, notably the SOTA Gemma model.
3. Benchmark the trained models using the dedicated Ukrainian dataset and compare them to the proprietary models.
4. Introduce the UKID instruction training dataset and make it publicly available for future fine-tuning efforts.
5. Perform the entire process in a fair and reproducible manner in order to facilitate future efforts.

2. Dataset and the experimental setup

Despite the abundance of online tutorials available for training large language models, establishing a reproducible setup for each model, complete with an appropriate dataset in the necessary format, proved to be unexpectedly challenging. Every model comes with its own set of constraints, including hardware requirements, deployment methods for inference, and specific approaches for processing instructions.

2.1. Dataset collection

When our team started working on the shared task for the UNLP conference, we were taken aback by the scarcity of suitable datasets for fine-tuning LLMs in Ukrainian. The organizers supplied a training dataset comprising 3,063 instruction rows, designed to acclimate the model to the multiple-choice format prevalent in the Ukrainian national examination. While this dataset proved valuable for training the LLM to answer in a specific format, it was notably limited in depth, offering little in terms of enhancing these LLMs' parametric knowledge base.

Through multiple experiments, we determined that 3-5 epochs of LoRA fine-tuning were sufficient for the model to grasp the multiple-choice format required for evaluation in the conference's shared task. However, the model's responses were lacking consistency, particularly when it generated incorrect or nonsensical answers. For instance, the model erroneously referred to "borsch," a well-known Ukrainian dish, as an item used in cars (See Figure 1).

This behavior underscored a deficiency in the model's general conceptual understanding, highlighting the pressing necessity to augment the dataset with more content in Ukrainian.

Consequently, we leveraged the **UAlpaca** dataset (Had) alongside **Squad-uk** (Drastic) which happened to be the only instruction datasets in the Ukrainian language available publicly.

Unfortunately, even after fine-tuning with these datasets, we observed that the model still didn't improve much, even on the training dataset itself, despite an improvement in sentence formulation and conceptual understanding. This led us to realize that a much more comprehensive approach to dataset construction would be required. Both UAlpaca and Squad-uk happened to be translated versions of the general knowledge English-based datasets, which is missing Ukrainian context and knowledge that is specific to both cultural and historical aspects that were being evaluated by the questions in the exam dataset. This realization led us to rethinking what kind of data we need and led to the creation of our own dataset, the Ukrainian Knowledge and Instruction Dataset (UKID), the first Ukrainian instruction dataset rooted in a Ukrainian context.

2.2. UKID methodology and construction

In formulating our hypothesis for the development of the Ukrainian language model, we posited that the model must align with the informational needs of the general population, reflecting the genuine interests and search behaviors of Ukrainian web users. To identify the most pertinent sources of intent-aligned knowledge, we turned to two widely recognized platforms: Wikipedia and Google. Consequently, we adopted a methodology focused on aggregating the most frequented Wikipedia pages, as determined by monthly traffic statistics, to ensure our dataset accurately captured the topics of highest relevance to Ukrainian web users.

We collected 1,064 pages by targeting those with monthly visit statistics ranging from 3,000 to 150,000. However, not all top-ranking Wikipedia pages in Google search results proved pertinent to our objective, as many described phenomena or entities not relevant to Ukraine. To refine our dataset, we employed a binary classification process to discern between relevant and non-relevant pages. This filtration mechanism is summarized in the table below, showcasing relevant versus non-relevant content (See Table 1). Through this methodical approach, we identified 367 pages that were suitable for inclusion in our dataset creation process.

The proposed methodology suggests an optimal approach for organizing an instruction-based dataset, aimed at fine-tuning language models for

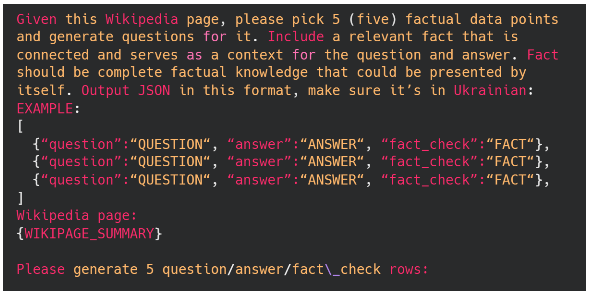
Page Title	Relevance
Ембер Герд	Not Relevant
Емульсія	Not Relevant
Ендокринна система	Not Relevant
Енеїда (Котляревський)	Relevant
Енцефаліт	Not Relevant
Еритроцити	Not Relevant
Єлизавета II	Not Relevant
Жадан і Собаки	Relevant
Жанр	Not Relevant
Житомир	Relevant

Table 1: Showcase of relevant vs non-relevant content.

underrepresented languages. This strategy offers the dual benefits of incorporating language-specific contexts and embedding essential factual knowledge into the model's trainable parameters during fine-tuning. Consequently, in addition to the conventional "question-answer" instruction pairs, we introduced a "fact_check" field. This addition acts as a comprehensive and standalone source of truth, enhancing the model's ability to verify facts and improve its accuracy. Performing this manually would have been unrealistic given the time constraints of the conference submission deadline, therefore an automated approach was implemented through the use of the Gemini 1.0 API and a few-shot learning example that utilizes the summary abstract of the Wikipedia page (See Figure.2)

As a result UKID-v0.1 was formed consisting of 962 question-answer-fact (QAF) pairs. Future work needs to focus on expanding the dataset to match other popular English-based datasets like Alpaca and Squad that consist of tens of thousands of rows. Even though the traditional notion of "less is more" for general English-based models recommends having smaller datasets (Zhou et al., 2023), our learnings indicate that fine-tuning under the constraints of lacking general conceptual understanding and context requires using much larger datasets.

Additionally, we have contemplated further enhancements to the UKID format, such as incorporating the original paragraphs from which the QAF pairs were derived to provide additional context. However, this aspect of the project remains unaddressed at present.



```
Given this Wikipedia page, please pick 5 (five) factual data points
and generate questions for it. Include a relevant fact that is
connected and serves as a context for the question and answer. Fact
should be complete factual knowledge that could be presented by
itself. Output JSON in this format, make sure it’s in Ukrainian:
EXAMPLE:
[
  {“question”:“QUESTION“, “answer”:“ANSWER“, “fact_check”:“FACT“},
  {“question”:“QUESTION“, “answer”:“ANSWER“, “fact_check”:“FACT“},
  {“question”:“QUESTION“, “answer”:“ANSWER“, “fact_check”:“FACT“},
]
Wikipedia page:
{WIKIPAGE_SUMMARY}

Please generate 5 question/answer/fact\_check rows:
```

Figure 2: Prompt to generate UKID examples

A crucial consideration in dataset development is tailoring the instruction format to the specific requirements of different models. For instance, Llama, Mistral, and Gemma each necessitate unique formats. Overlooking this critical aspect has empirically led to suboptimal outcomes, though these observations have yet to be formally documented. The adaptation of datasets to align with the distinct formats of these models is essential for maximizing their performance and efficacy.

3. Fine-tuning

3.1. Gemma models

First, we fine-tuned a Gemma-2B and a Gemma-7B model, from a recently published family of open models.

We used official "**gemma-2b-it**" and "**gemma-7b-it**" weights published by Google and followed official fine-tuning guidelines on the Vertex AI platform. The final python notebook is located in the "from-bytes-to-borsch" github repository.

Fine-tuning for gemma-2b-it was performed with a combined dataset consisting of 13,063 instructions, which included from the 10,000 rows of UAlpaca dataset and 3,063 rows from the ZNO dataset provided by organizers of the conference. Fine-tuning for gemma-7b-it was performed with a dataset consisting of 14,025 instructions (10,000 rows of UAlpaca, 3,063 rows of ZNO and 962 rows of UKID).

Due to resource constraints, we chose to use a LoRA (Hu et al., 2022) fine-tuning approach. We used a LoRA adapter implementation from the Keras v3 library, with $lora\_r = 4$, resulting in 11,067,392 trainable parameters, instead of the full 7B for the case of Gemma-7B.

The resulting model was published on the associated github repository. Unfortunately due to the time constraints we were not able to submit the 7B to the UNLP competition benchmarking, and only submitted results from the 2B instruct model.

3.2. Mistral model

As a second alternative, we used a completely different fine-tuning pipeline with the help of the axolotl tool to streamline the fine-tuning process. We used a 4x Nvidia Tesla A100-80Gb GPU instance on Microsoft Azure cloud for training. Due to compute constraints we chose to use the LoRA (Hu et al., 2022) approach once again, this time implemented using Hugging Face transformers library.

We used an AdamW optimizer (Loshchilov and Hutter, 2017) with common starting point hyperparameters for the LoRA adapters ($lora\_r = 32$, $lora\_alpha : 16$), which resulted in 32,505,856 trainable parameters.

For Mistral-based fine-tunes we used the "mistralai/Mistral-7B-Instruct-v0.1" weights and “LlamaTokenizer” tokenizer.

The training was performed using ZNO and Uk-Squad datasets. Both datasets have a Llama/Alpaca instruction format and collectively produced 37,890 rows of instructions.

More details of the configuration and execution can be found in the associated github repository.

4. Benchmarking results

We performed benchmarking using two test datasets: multiple choice questions (MCQ) and open questions (OQ).

The MCQ dataset comprises 3,063 questions from the Ukrainian External Independent Testing (EIT) test, a standard government test for college admission taken by secondary school students. This dataset splits into 1,139 Ukrainian history questions and 1,925 Ukrainian language and literature questions, reflecting the standard knowledge expected in Ukrainian schools. We evaluated this test automatically.

The OQ dataset contains 100 instruction-based questions prompting models to complete generative tasks, such as finishing a story or summarizing an event. We evaluated this dataset manually.

Below, we detail our benchmarking setup and criteria, focusing on the fine-tuned Gemma models, Gemma7bFT and Gemma2bFT, alongside an out-of-the-box model, Gemma7b, for reference.

4.1. Multiple choice questions

We presented all questions from this dataset within a uniform prompt in Ukrainian, instructing models to select the single correct answer in letter form. Despite this directive, models frequently included extraneous information, necessitating manual filtration to extract the required letter codes. Correct responses matched the letter codes exactly. Table 2 displays the models' performance percentages in each category.

4.2. Open questions

Evaluating open questions required a more nuanced approach, examining responses across four categories:

- Ukrainian (U): the response is given in the Ukrainian language.
- Facts/Coherence (C): factual correctness and coherence of the given answer.
- Relevance (R): the answer aligns with the given instructions.
- Grammar (G): stylistic and grammatical evaluation.

Each response could earn up to 1 point per category, with the results and average scores presented in Table 3.

Model	History (%)	L&L (%)
GPT4	82.95	47.12
Gemini	71.97	40.99
GPT3.5	52.37	26.65
MistralFT	40.16	22.86
Gemma7bFT	37.96	21.71
Gemma2bFT	28.91	20.57
Gemma7b	26.36	19.01

Table 2: Model benchmarking with multiple choice questions.

Model	U	C	R	G	Avg
GPT 4	97	79	85	79	85
GPT 3.5	97	61	79	74	77.75
Gemini	96	67	81	84	82
MistralFT	89	7	18	49	40.75
Gemma7b	85	13	45	35	44.5
Gemma7bFT	54	13	48	19	33.5

Table 3: Model benchmarking with open questions.

4.3. Discussion

The obtained results provide interesting insights into many aspects of the LLM's performance and training.

First, let us consider the results of the open-source models. Comparing the performance of the Gemma7b model before and after the fine-tuning it becomes very clear that the fine-tuning process can indeed improve its knowledge in a particular area by a large margin, in this case by roughly a quarter. Mistral shows even better improvement in answering the MCQs. Even the much smaller model Gemma2b outperforms its non-fine-tuned larger counterpart Gemma7.

However, besides improving model's performance in certain areas, the fine-tuning process appears to introduce artifacts that affect performance when answering these open questions. Mistral, after fine-tuning, seemed to struggle with following the given instructions (see the **R** column in Table 3). On the other hand, Gemma7bFT's ability to speak Ukrainian was impaired by 40%, also reducing its grammar score by nearly a half (columns **U** and **G** in Table 3). What's most exciting, Gemma7bFT started to manifest the *code-switching* phenomenon which can be considered an emergant property, and will be discussed in more detail in the Conclusions section.

It comes as no great surprise that the proprietary models performed substantially better in all kinds of tasks. The reasons are numerous, with the most obvious being:

- The scale of parameters significantly contributes to model performance. For instance, GPT-3.5 boasts 25 times more parameters than both Gemma7b and Mistral, whereas GPT-4 and Gemini exceed these models by over a hundredfold in terms of parameter count.
- Proprietary models benefit from unparalleled access to the most comprehensive and high-quality datasets available, ensuring a broad and deep understanding of language.
- The training of proprietary models extensively incorporates reinforcement learning techniques, refined through human feedback, to achieve nuanced understanding and response generation.

Nevertheless the performance of the fine-tuned open-source models is not so far behind that of GPT3.5. With additional efforts invested into the fine-tuning of open-source models, it is definitely possible to beat GPT3.5 in a range of specific language-related tasks.

A notable observation across all models was the disparate performance on Ukrainian history versus language and literature, echoing a trend irrespective of model origin. By design the EIT questions in different subjects are meant to be of the same complexity such that an average Ukrainian school student gets average marks in every subject. However, the performance of every LLM tested showed very skewed results, with history knowledge favoured over that of language and literature. Possible reasons could include:

- The skew in available datasets toward history is due to its widespread availability from open sources such as Wikipedia. Conversely, literature demands greater effort to gather, organize, and present, contributing to its underrepresentation.
- Answering history questions accurately is largely a matter of recalling specific factual information, such as dates, names, and events. Literary analysis, however, requires navigating complex themes, symbolism, and cultural nuances, demanding a more profound understanding of both language and context.
- The Ukrainian language, along with its cultural and literary heritage, often falls outside the primary interests of major corporations, affecting the availability and focus of datasets dedicated to these areas.

This underscores the cultural bias challenge in advanced LLMs today which will be further discussed in subsequent sections.

4.4. Code-switching and Azirivka

Code-switching is a linguistic phenomenon in which a speaker alternates between two or more languages within a single utterance or sentence. Until recently, this term was applied only to humans, but with the advent of LLMs this effect has been observed and studied in generative models (Winata et al., 2021; Zhang et al., 2023). Code-switching in LLMs arises from the multilingual nature of training and fine-tuning processes.

For historical reasons, the majority of the Ukrainian population is multilingual. This creates a rather unique situation when constant code-switching is common at practically every level, starting from colloquial everyday conversations and ending with official statements from prime-ministers and presidents. A particular case of the latter has the official name Azirivka (Wikipedia), named after Ukrainian ex-prime minister Mykola Azarov.

Observing the Gemma7b model mastering Azirivka after fine-tuning was both interesting and exciting. It is particularly interesting that the model generates not a simple mixture of words belonging to different languages, but rather conjugates words from one language according to the rules of another, just as some Ukrainians do, demonstrating features specific to synthetic languages.

Below, we present several instances of Azirivka code-switching. In these examples, components highlighted in blue represent Ukrainian, while those in red denote Russian.

Example 1:
Azirivka: Твір про коллекцию кольоровых олівцов Василя Голобородька.
English: An essay about Vasyl Holoborodko's collection of colored pencils.

Example 2:
Azirivka: Привітать друзів с одруждением можно множеством способов.
English: You can congratulate friends on their marriage in many ways.

Example 3:
Azirivka: Я обращаюсь к Вам с жалобой по неякісной замене труб в подвалі нашего дома, расположенного по [адрес].

English: I am addressing you with a complaint about the poor-quality replacement of pipes in the basement of our house, located at [address].

Example 4:
Azirivka: В Украине Маланку не святкуют.
English: Malanka is not celebrated in Ukraine.

Example 5:
Azirivka: У п'ятницю, 23 лютого, в Україні опадів не будет, но местами - рвучкий і сильний вітер.
English: On Friday, February 23, there will be no precipitation in Ukraine, but there will be occasional gusty and strong wind.

It's worth noting that while most of these mixed words can't be found in official dictionaries, they are commonly heard on the streets of many Ukrainian cities. Such a language mixture naturally has been an object for linguistic studies (Bilaniuk, 2004; Kent, 2011). We consider this emerging LLM property to be of great interest for further studies.

5. Applications, risks and future work

It is abundantly clear that having a language-specific model is going to aid all of the possible use cases around communication, but it's also important to note the risks of not having the model. Both from the industrial and cultural standpoints.

Incorporating LLM models of underrepresented languages into technology platforms offers unprecedented opportunities for enhancing communication across diverse sectors, ranging from healthcare and education to legal and commerce, all within the scope of the growing impacts of globalization. However, the absence of such models poses significant risks, not only stalling industrial progress but also exacerbating cultural erosion. Industrially, the lack of tailored language models can hinder the efficient dissemination of critical information, reduce the accessibility of digital services, and create barriers to entry for local businesses in the global market. Culturally, it threatens the preservation of linguistic diversity and the transmission of heritage, as languages without digital representation risk falling into disuse and oblivion. Therefore, addressing this gap is not merely a technical challenge but a pressing societal need that calls for collaborative efforts to ensure inclusive and sustainable development.

5.1. Applications

Oleksandr, a Ukrainian refugee in the USA, benefits from a language-specific LLM that digests and explains legal aid and immigration documents into Ukrainian. This tool helps him and his family understand their rights and the process for seeking asylum, significantly easing their transition into a new country while maintaining their linguistic identity during a period of immense upheaval and change.

Maria, a primary school teacher in a rural Peruvian village, uses a language-specific LLM to access educational materials in Quechua, enabling her to provide more engaging and culturally relevant lessons to her students. This technology allows her to bridge the gap between traditional knowledge and modern education, fostering a learning environment where students can appreciate their heritage while gaining access to the wider world of knowledge.

Michael, a software developer with Navajo heritage, creates an interactive application powered by a language-specific LLM that facilitates live, conversational practice in Navajo for learners worldwide. This platform connects Navajo speakers with learners, enhancing language proficiency through real-time dialogue and cultural exchange, thereby revitalizing the Navajo language among younger generations and spreading awareness of Navajo culture globally.

5.2. Risks through the prism of education

Classroom education and child development will depend heavily on large language models tailored for different languages and contexts, especially since there is no doubt in the growing influence of AI on youth, in particular within the educational and edutainment contexts (Chowdhury, 2023). That's why one may hypothesize that countries like Ukraine will eventually face a linguistic identity crisis in 15-20 years without accessible Ukrainian-tuned LLMs.

At the primary school level, Ukraine's youth increasingly speak a homogenized and influenced version of Ukrainian rather than preserved distinctive dialects. Besides an obvious impact of Russification, globalization makes it even harder to preserve Ukrainian heritage due to its decreasing utility when it comes to cultural integration into the global landscape. One might argue that Ukraine is having a unique moment in time where cultural identity is being amplified by the risk of complete wipeout by an invading neighbor country, but other developing countries may never have such unique constraints to enable cultural amplification and preservation.

One other risk is related to not having interactive AI tools. Lack of an engaging Ukrainian AI tutoring solution will lead to the inability to pass on common fables, heritage literature analysis skills, and critical moments familiar to prior generations. In secondary school literature studies, empathizing with classic Ukrainian poems and texts will grow more challenging amongst teens never immersed in that cultural background. Likewise, they will struggle with interpreting symbolism and references common to those eras of Ukrainian identity formation while not receiving any support from Ukrainian-aligned language models for written compositions or humanities projects. Subsequent generations will lose touch with integral pieces of the country's unique heritage story.

Even on an informal level, interest in artistic efforts around theater, cinema, visual arts, and music see declining engagement from younger Ukrainians as preferred leisure activities shift towards globalized media culture rather than celebrating local creators and talent. Despite the current obvious boom of local cultural talent, there is still a huge subset of the population that is dependent on external sources of entertainment, from movies to music (Molfar).

In essence, Ukraine and similar developing countries face looming risk over the next generation, where accumulated erosion across countless tiny dimensions of language diversity and identity lead to forging an entirely different nation - with culture, history, and influence conspicuously drifting into the shadows of a former self, which has been so fiercely fought for.

Such is the steep collective price societies can pay when neglecting "untimely" AI model development efforts in favor of convenience and cost during pivotal transition points in history. This danger is imminent unless there is an immediate increase in urgency to prioritize national languages and invest in critical computing infrastructure for educators and policymakers. The decisions made in the coming five years on prioritization between language-specific and multilingual model availability carry potentially profound societal consequences depending on which vision prevails under the pressures of globalized technology proliferation.

5.3. Risks of underrepresentation

Over the past 15 years, Ukrainian Google and YouTube search queries have become increasingly dominated by Russian language pages and video results (Search Engine Land, 2023). This occurred because Russian internet data grew rapidly early on - amassing orders of magnitude more content, sites, and engagement than the Ukrainian web, alongside the unfortunate post-russification effects of the Soviet era.

As a consequence, Google's algorithms seeking to maximize search intent fulfillment for Ukrainian keyword queries, surfacing Russian pages higher in results because, probabilistically, people's intent gets fulfilled more often there based on aggregate global click behavior.

This creates a self-reinforcing flywheel where Russian sites continue gaining more links, clicks, and search authority compared to Ukrainian community pages on the same topics despite not matching the native language exactly.

Similarly, as large language models for different languages mature — if Russian LLMs accumulate exponentially more parameters, content trained on, and research budget than available Ukrainian models — probabilistic fulfillment of natural language queries and conversational needs from Ukrainian users will skew towards Russian-centric resources. Even if the Ukrainian content exists, it surfaces less prominently. And, gradually, queries normalize towards Russian linguistic structures and dialects if that provides higher collective fulfillment rates globally. This also provides an enormous data feedback loop effect as the applications and model creators are able to generate even more human feedback data on which to improve models.

Without dedicated investment from both public and private sectors in developing models for native languages, we risk cultural erosion. This comes from a reliance on technology that favors more dominant languages, simply because it's more convenient.

This convenience itself opens up an opportunity for another medium of risk, enabling much faster and efficient distribution of propaganda and misinformation, requiring its own unique mechanisms for detection and prevention (Solopova et al., 2023). This is an obvious risk that is becoming critical in the political and existential context for any developing country that is affected by external pressure from other foreign countries.

5.4. Future work, policy, and critical timing

As large language models continue rapidly advancing thanks to unprecedented compute investments by groups like OpenAI, Anthropic, Google, Meta, and Baidu, a clear "model divide" looks poised to emerge.

Hundreds of lower-resource languages globally now stand at risk of accelerating identity erosion without specialized LLM variants representing their linguistic contexts. From Navajo conversational interfaces to Quechua literary analysis tools to Welsh educational content creators — sadly, these languages are falling behind on the rapid advancements in today's technology.

Consequently, many threatened languages pose a digital extinction risk without counterbalancing forces to protect their dialects, artistic traditions, and communities. These groups often struggle due to the lack of institutional support, which results in insufficient access to the necessary data and resources.

As future generations raised on AI inherit even subtle biases favoring better resourced languages, the cultural price to pay will grow exponentially steeper. Preserving heritage hence requires some rebalancing, where policymakers implement commitments to inclusive innovation, perhaps evaluating issues of sustainability for vulnerable groups rather than solely technical tradeoffs.

Companies and governments worldwide must acknowledge that shortsighted stances on optimized efficiency today cascade into seismic identity impacts downstream. Access barriers erode dialects, discourage artistic traditions, and deter descendants from inheriting linguistic lineage — ultimately dimming cultural continuity prospects.

Prioritizing LLM development for lower-resource languages offers a reverse course against irreversible language extinction already accelerating since the turn of the century. As risks become solutions, so do data divides resolve through compassionate actors cooperating across borders to uplift unseen communities, now empowered to share their visions.

6. Conclusions

In this paper, we have explored the importance of developing language-specific large language models (LLMs) for underrepresented languages, focusing on the Ukrainian language as a case study. Our findings demonstrate that cultural bias is a quantifiable phenomenon, and we can speculate about its underlying causes. The open-source community plays a crucial role in addressing this issue by creating new, extended datasets and publishing them for further research work. While this effort may be beyond the scope of commercial interest, it has immense humanitarian impact.

It's important to note the emergence of code-switching effects like Azirivka, which occur spontaneously and highlights the similarities between pattern learning mechanisms in humans and LLMs. While fully recognizing that this intriguing phenomenon warrants a more thorough examination, we contend that even preliminary observations merit reporting. The existence of such effects in human societies, where two languages coexist in close contact, further reinforces the importance of developing language-specific models to preserve cultural identity and linguistic diversity.

To advance the evaluation of language models for Ukrainian, we have introduced **ULIB**, the "Ukrainian Linguistic Inquiry Benchmark." This benchmark encompasses various language processing tasks, including summarization, poem generation, spelling, and simplified explanation comprehension. ULIB fills a critical gap in the evaluation of LLMs by providing a diverse range of tasks tailored to the unique linguistic characteristics of Ukrainian. By offering a holistic evaluation framework, ULIB enables human evaluators to assess the performance of LLMs in understanding and generating Ukrainian text. Although we have only introduced the format and starting point for ULIB datasets, which are available on our github, we plan to expand it as part of our future work.

In addition to ULIB, we have also introduced the Ukrainian Knowledge and Instruction Dataset (**UKID**), a pioneering instruction dataset rooted in Ukrainian context. UKID serves as a comprehensive and standalone source of truth, enhancing the model's ability to verify facts and improve its accuracy. By incorporating language-specific contexts and embedding essential factual knowledge into the model's trainable parameters during fine-tuning, UKID paves the way for more effective and culturally relevant language models.

Our work highlights the significance of developing language-specific LLMs and datasets, not only for Ukrainian but for all underrepresented languages worldwide. By demonstrating the feasibility and importance of this approach, we hope to inspire further research and development in this area. Future work should focus on fine-tuning open-source models with expanded datasets, improving evaluation benchmarks, and exploring innovative applications that leverage the power of language-specific LLMs. Through collaborative efforts between researchers, open-source communities, and stakeholders, we can work towards a future where AI technologies are truly inclusive and representative of the world's linguistic and cultural diversity.

7. Acknowledgements

We would like to acknowledge the support of UkraineNow.org and Nvidia Corporation for providing a DGX Station, a turnkey deskside AI supercomputer with four NVIDIA® Tesla® V100 Tensor Core GPUs, which was used to benchmark the fine-tuned models.

We also thank Tensoic and Microsoft Azure Cloud for providing the compute resources to fine-tune the Mistral-based model.

Additionally, we are grateful to Google LLC for supplying the pre-trained weights for the Gemma models and the fine-tuning infrastructure on the Vertex AI platform, which allowed for easy and quick setup of the fine-tuning and deployment processes within a short timeframe.

References

Israel Abebe Azime, Mitiku Yohannes Fuge, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Aman Kassahun Wassie, Eyasu Shiferaw Jada, Yonas Chanie, Walelign Tewabe Sewunetie, and

Seid Muhie Yimam. 2024. Enhancing amharic-llama: Integrating task specific and generative datasets.

Pierpaolo Basile, Elio Musacchio, Marco Polignano, Lucia Siciliani, Giuseppe Fiameni, and Giovanni Semeraro. 2023. Llamantino: Llama 2 models for effective text generation in italian language.

Y. Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. volume 3, pages 932–938.

Laada Bilaniuk. 2004. A typology of surzhyk: Mixed ukrainian-russian language. *International Journal of Bilingualism - INT J BILING*, 8:409–425.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Tahiya Chowdhury. 2023. Towards goldilocks zone in child-centered ai. 4 pages.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling.

Yiming Cui, Ziqing Yang, and Xin Yao. 2024. Efficient and effective text encoding for chinese llama and alpaca.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Drastic. github.com/drastic/squad-uk.

Jay Gala, Thanmay Jayakumar, Jaavid Aktar Husain, Aswanth Kumar M, Mohammed Safi Ur Rahman Khan, Diptesh Kanojia, Ratish Puduppully, Mitesh M. Khapra, Raj Dabre, Rudra Murthy, and Anoop Kunchukuttan. 2024a. Airavata: Introducing hindi instruction-tuned llm.

Jay Gala, Thanmay Jayakumar, Jaavid Aktar Husain, Aswanth Kumar M, Mohammed Safi Ur Rahman Khan, Diptesh Kanojia, Ratish Puduppully, Mitesh M. Khapra, Raj Dabre, Rudra Murthy, and Anoop Kunchukuttan. 2024b. Airavata: Introducing hindi instruction-tuned llm.

Google DeepMind Gemma Team. 2024. Gemma: Open models based on gemini research and technology. `https://storage.googleapis.com/deepmind-media/gemma/gemma-report.pdf`.

Robin Had. github.com/robinhad/kruk.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9:1735–80.

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*.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Larysa Katerynych, Maksym Veres, and Eduard Safarov. 2021. Transformer-based model for text classification in ukrainian. Taras Shevchenko National University of Kyiv.

Kateryna Kent. 2011. Language contact: Morphosyntactic analysis of surzhyk spoken in central ukraine.

Guneet Singh Kohli, Shantipriya Parida, Sambit Sekhar, Samirit Saha, Nipun B Nair, Parul Agarwal, Sonal Khosla, Kusumlata Patiyal, and Debasish Dhal. 2023. Building a llama2-finetuned llm for odia language utilizing domain knowledge instruction set.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neu-*

ral Information Processing Systems, volume 25. Curran Associates, Inc.

Yurii Laba, Volodymyr Mudryi, Dmytro Chaplynskyi, Mariana Romanyshyn, and Oles Dobosevych. 2023. Contextual embeddings for ukrainian: A large language model approach to word sense disambiguation. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop*, pages 11–19, Dubrovnik, Croatia.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Molfar. Song charts analysis. `https://molfar.com/en/blog/song-charts`. Accessed: 2024-04-04.

Quan Nguyen, Huy Pham, and Dung Dao. 2023. Vinallama: Llama-based vietnamese foundation model.

OpenAI. 2023. Gpt-4 technical report.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Holger Schwenk, Daniel Dechelotte, and Jean-Luc Gauvain. 2006. Continuous space language models for statistical machine translation. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 723–730, Sydney, Australia. Association for Computational Linguistics.

Search Engine Land. 2023. Google and the challenge of russian propaganda in search results. `https://searchengineland.com/google-russian-propaganda-search-results-382229`. Accessed: 2024-04-04.

C. E. Shannon. 1951. Prediction and entropy of printed english. *The Bell System Technical Journal*, 30(1):50–64.

David Silver, Aja Huang, Christopher Maddison, Arthur Guez, Laurent Sifre, George Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489.

Veronika Solopova, Oana-Iuliana Popescu, Christoph Benzmüller, and Tim Landgraf. 2023. Automated multilingual detection of pro-kremlin propaganda in newspapers and telegram posts. *arXiv preprint arXiv:2301.10604*. Available online at arXiv:2301.10604.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models.

Serhii Tiutiunnyk. 2020. Context-based question-answering system for the ukrainian language. Master's thesis, Ukrainian Catholic University, Lviv.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Bram Vanroy. 2023. Language resources for dutch large language modelling.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Wikipedia. Azirivka — wikipedia, the free encyclopedia.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2021. Are multilingual models effective in code-switching? *CoRR*, abs/2103.13309.

Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Indra Winata, and Alham Fikri Aji. 2023. Multilingual large language models are not (yet) code-switchers.

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.