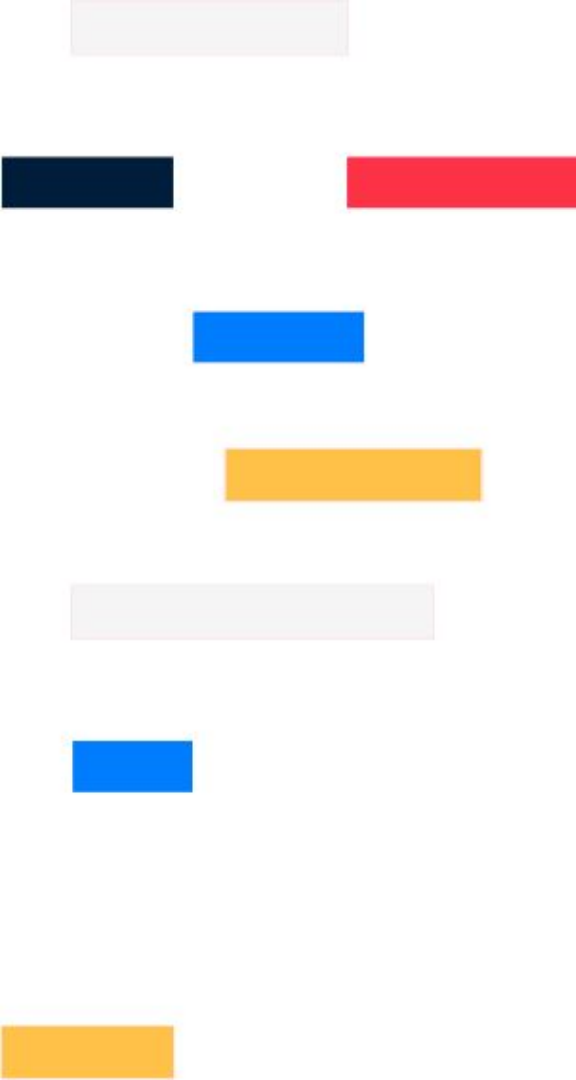


From Bytes to Borsch: Fine-Tuning Gemma and Mistral for the Ukrainian Language Representation

**Artur Kiulian, Anton Polishko, Mykola Khandoga,
Oryna Chubych, Jack Connor,
Raghav Ravishankar and Adarsh Shirawalmath**





Motivation:

Why do language
fine-tuning?

~~Large Language Models~~

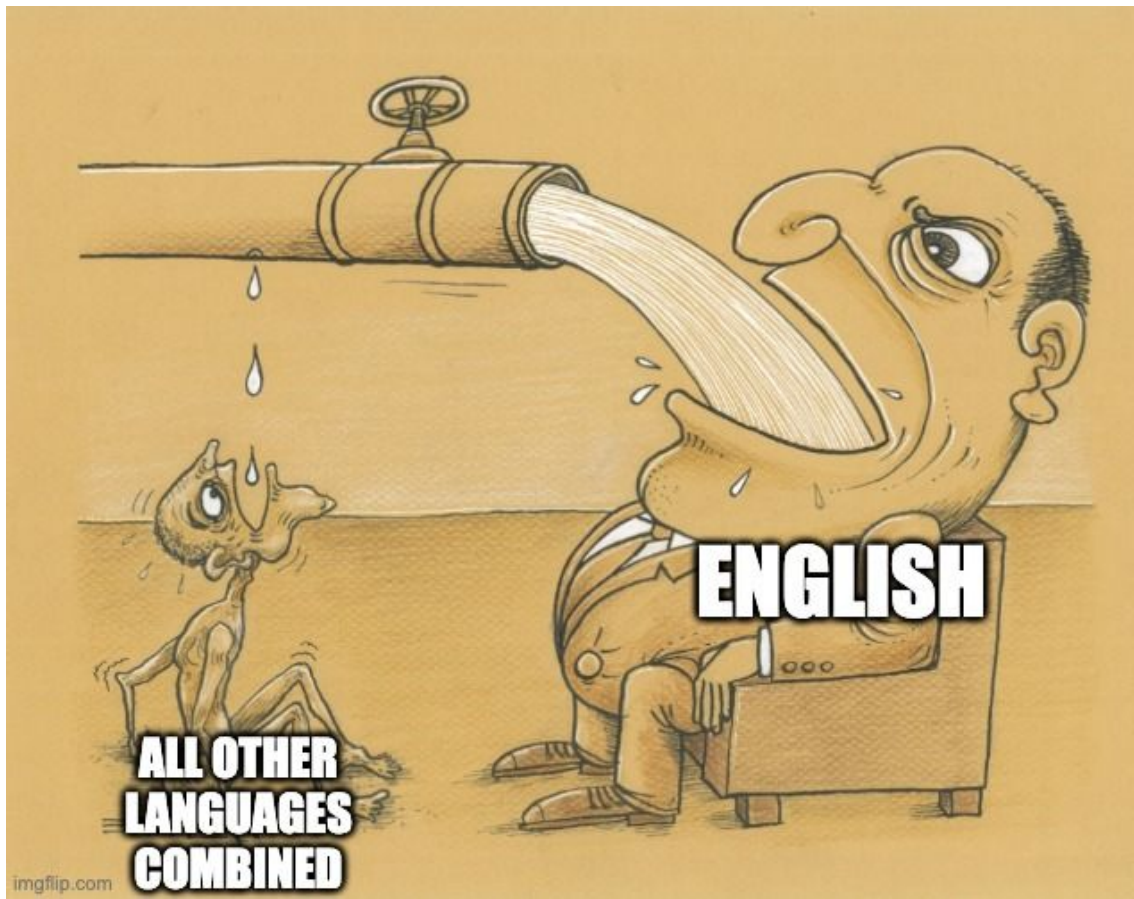
Large **English** Models



Training datasets

Training dataset
of LLama 3 8B is
95% English with
only 5% left for
all the other
languages
(according to
Andrej Karpathy)

**Leads to
immense cultural
bias**



Official ChatGPT tokenizer

EN

Tokens
219

Characters
914

A towel is just about the most massively useful thing an interstellar hitchhiker can carry. Partly because it has great practical value. You can wrap it around you for warmth as you bound across the cold moons of Jaglan Beta; you can lie on it on the brilliant marble-sanded beaches of Santraginus V, inhaling the heady sea vapours; you can sleep under it beneath the stars which shine so redly on the desert world of Kakrafoon; use it to sail a miniraft down the slow heavy River Moth; wet it for use in hand-to-hand combat; wrap it around your head to ward off noxious fumes or avoid the gaze of the Ravenous Bugblatter Beast of Traal (a mind-bogglingly stupid animal, it assumes that if you can't see it, it can't see you – daft as a brush, but very very ravenous); you can wave your towel in emergencies as a distress signal, and of course you can dry yourself off with it if it still seems to be clean enough.

Text Token IDs

UA

Tokens
549

Characters
919

Рушник – це, мабуть, найкорисніша річ, яку може носити з собою міжзоряний мандрівник. **00**астково тому, **000** він ма**00** велику пра**00**тичну цінність. Ви можете обгорнути ним себе для тепла, коли перескаку**00**те через холодні місяці **00**глан Бета; ви можете лежати на ньому на блискучих мармурово-піщаних пляжах Сантра**00**інуса V, вдихаючи п'янки морський випар; ви можете спати під ним під зірками, які так червоно світять на пустельній планеті Какарафун; використовувати **000**ого для плавання на мініплаті вниз за повільною важкою річкою Мот; змочити **000**ого для використання у рукопашному бою; обгорнути ним свою голову, **000**об уникнути шкідливих випарів або погляду **00**адібно**00** Звірюки Бугблаттера з Траала (неймовірно дурне тварина, воно вважа**00**є, **000** якщо воно не бачить вас, то ви не бачите **000**ого – тупе як **00**ітка, але дуже дуже жадібне); ви можете махати сво**000**м рушником у надзвичайних ситуаціях як сигналом біди, **i**, звичайно, ви можете витертися ним.

Text Token IDs

The model sees no characters, it sees tokens.
EN tokens are longer, making the entire text shorter for the model.
That means larger context window and cheaper usage.

Code switching

Want to help me make un castillo?

¡Sí! Let's make a castle grande!



Neither of these kids have DLD. Code-switching is a normal part of being bilingual.

Human 2: 请你为全球军事实力进行国家层面上的排序, 不包括民间组织与其他机构。

(**Human2:**Please provide a country-level ranking of global military capabilities, excluding non-state organizations and other entities.)

Assistant 2: ... 1. 美国: 武器: 美国武器有着世界上最新的轻装甲车、追踪性能最强的潜水艇等等的優れた武器化状况。...

(**Assistant 2:** ... 1. United States: Weapons: The United States possesses advanced weaponry including the world's latest lightweight armored vehicles, submarines with the highest tracking capabilities, and other superior weapons. ...)

In Ukraine bilingualism is extremely common. Although almost never written, one can encounter oral code-switching everywhere. We even had a Prime-Minister Mykola Azarov who was mocked a lot for code-switching all the time.

Example 1:

Azirivka: Твір про колекцію кольорових олівців Василя Голобородька.

English: An essay about Vasyl Holoborodko's collection of colored pencils.

Example 2:

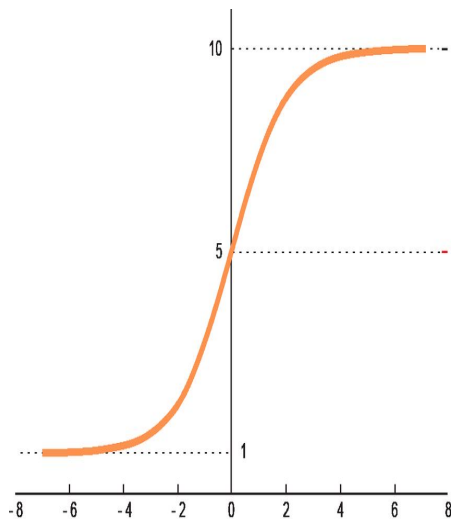
Azirivka: Привітати друзів с одруженням можно множеством способов.

English: You can congratulate friends on their marriage in many ways.



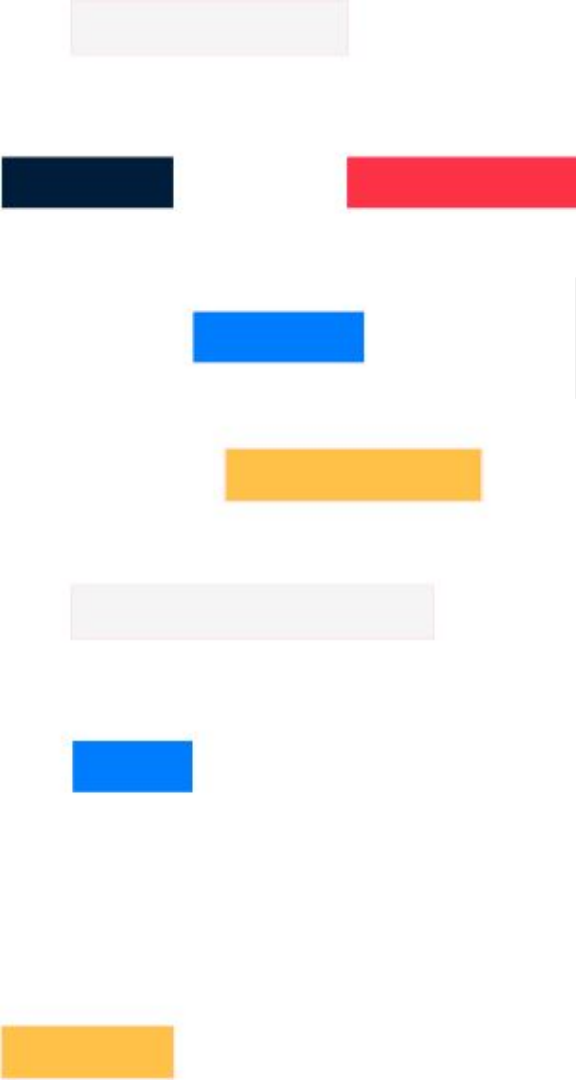
Why not making LLMs truly multilingual?

- Bigger dataset -> hard to get, more expensive training
- Bigger vocabulary -> larger model, more expensive training
- Technical difficulties -> code switching
- Reinforced learning from human feedback (RLHF) -> expensive



Softmax





Datasets and experimental setup

Ukrainian Knowledge and Instruction Dataset (UKID)

(UKID) has been created by our team specifically for this fine-tuning task through the following steps:

- Identify 1064 most visited Wikipedia pages
- Filter out 367 pages that are relevant to Ukrainian context
- Compose 962 Question-Answer-Fact (QAF) triplets using Gemini 1.0 API.

The resulting UKID dataset is made public.

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

```
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:
```

Fine-tuning Gemma and Mistral into Ukrainian

- Datasets: instruction datasets [UAlpaca](#), [SQuAD-uk](#), UKID
- Training setup:
 - Gemma: official fine-tuning guidelines from the [Vertex AI platform](#)
 - Mistral: [axolotl](#) tool, configs can be found in [our repo](#)
- Hardware: 4x Nvidia Tesla A100-80Gb GPU instance on Google Cloud Platform

A decorative graphic in the top-left corner of the slide, consisting of several horizontal bars of different colors and lengths. The colors include light gray, blue, dark blue, red, orange, and white. The bars are arranged in a somewhat scattered pattern, with some overlapping.

Results

Linguistic benchmarking: II parts

Multiple choice questions in two categories:

- Ukrainian history
- Ukrainian language and literature

Open questions with human-in-the-loop evaluation according to CRUG system:

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

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

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

A decorative graphic on the left side of the slide consists of several horizontal bars of different colors and lengths, arranged in a staggered, descending pattern from top-left to bottom-left. The colors include light gray, blue, dark navy, red, blue, orange, light gray, blue, and orange.

Follow-up

Follow-up and work in progress: towards a model with better language capabilities

- Larger corpus training: improving general language skills
- Adding 70k new UA tokens: reducing token/word ratio
- [Smart embedding initialization](#) of new tokens: leverage existing tokens
- Training:
 - Gemma-7b using JAX and model parallelism
- Hardware: 8x Nvidia Tesla A100-40Gb GPU