# Vuyko Mistral

Adapting LLMs for Low-Resource Dialectal Translation

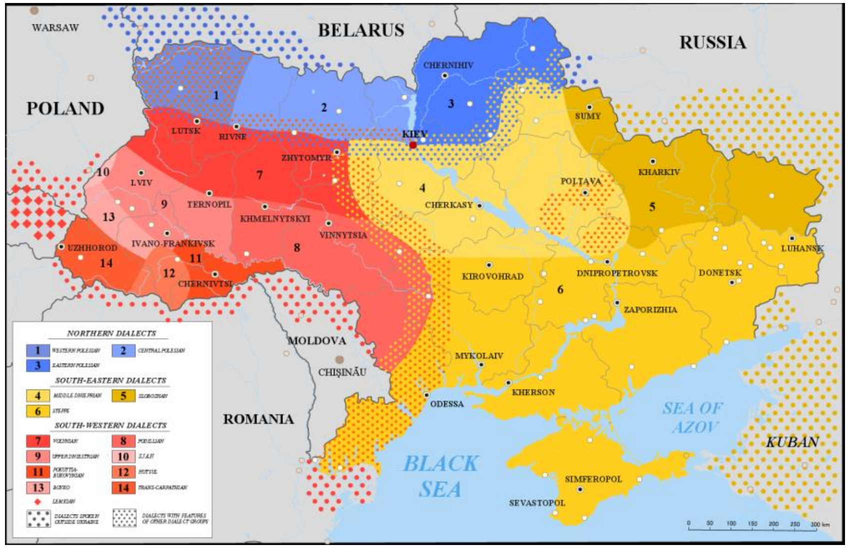
**Roman Kyslyi[1], Yuliia Maksymiuk[2], Ihor Pysmennyi[1]**
[1]National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
[2]Ukrainian Catholic University

Mistral:

Ой, єй, виджу, шо си зібрало богато людів на конфєренцію! Та й файно!

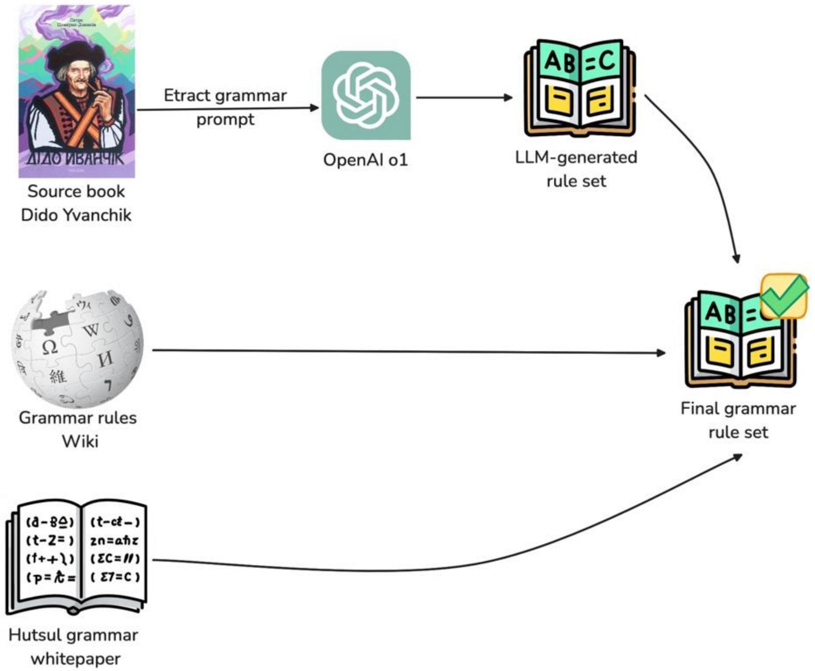
# Goal of study

- Most research is centered on high-resource languages and their standard variants underrepresenting dialects that lack sufficient textual resources and standardized orthographies
- To be preserved they need to be visible to modern LLM toolkit and research, which is challenging due to lack of resources



# Hutsul dialect

- We’ve chosen Hutsul dialect as it is still used in Carpathians (21000 as per 2001 census)
- It has several written sources which can be used for training
- We used hybrid approach to construct grammar rule set



Source book Dido Yvanchik

Etract grammar prompt

OpenAI o1

LLM-generated rule set

Grammar rules Wiki

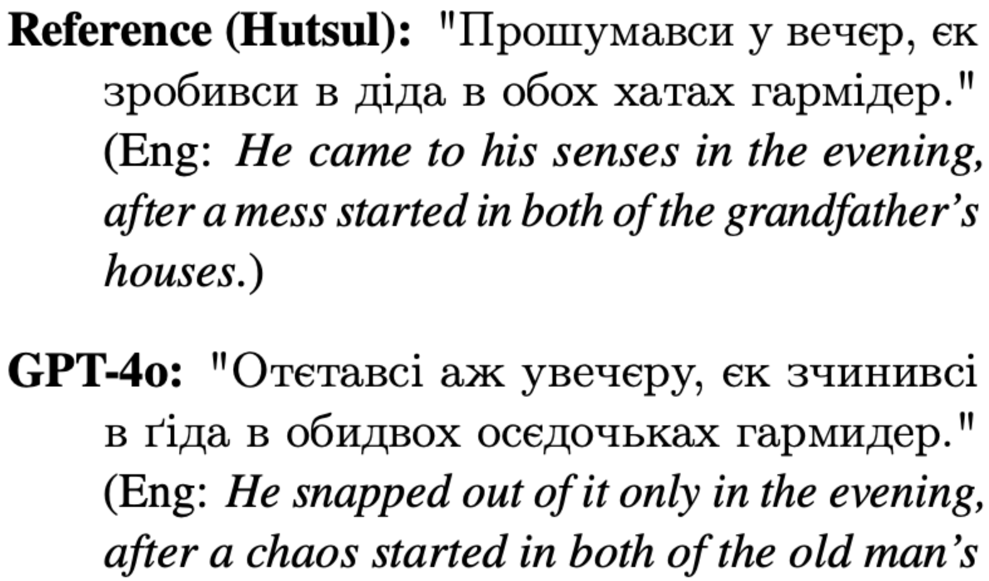
Final grammar rule set

Hutsul grammar whitepaper

**Linguistic characteristics of dialect:**

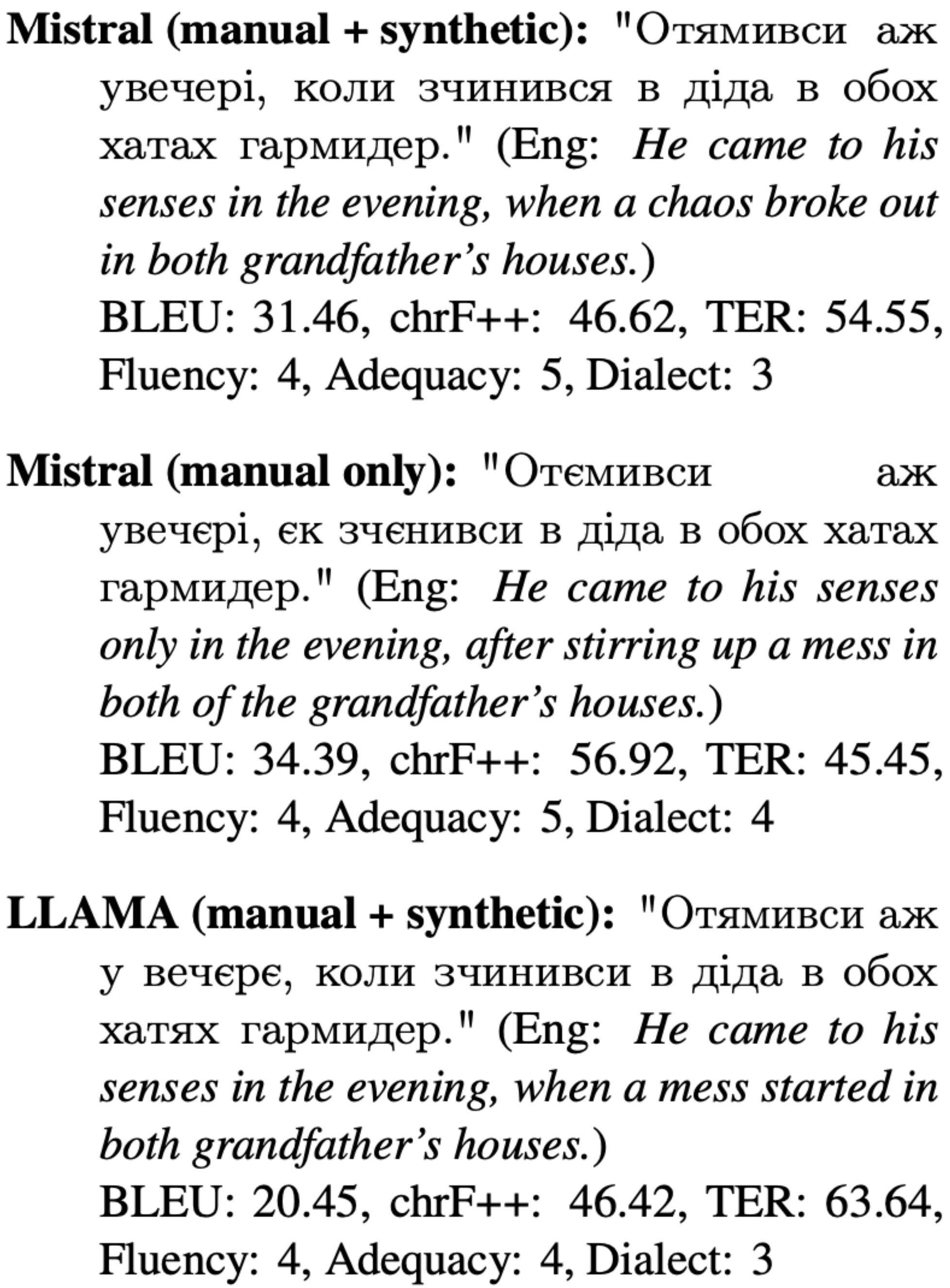
- *Phonetics*: vowel transformations, such aschanging vowels "є" instead of "а", "я"(ya) (example:"як" → "єк", "ягода" → "єгода” (“yak” → “yek”, “yahoda”→ “yehoda”)).
- *Morphology*: unique case endings (-єдь, -сі) (’-yed’, ’-si’) and preserved dual forms дві яблуції (“two apples”, with dual form “yablutsi” instead ofplural “yabluka”).
- *Lexicon*: Romanian, Polish and German borrowings such as "бринза" (cheese) and "шпацірувати" (goforawalk).

# Hutsul dialect



**Reference (Hutsul):** "Прошумавси у вечєр, єк зробивси в діда в обох хатах гармідер." (Eng: *He came to his senses in the evening, after a mess started in both of the grandfather's houses.*)

**GPT-4o:** "Отєтавсі аж увечєру, єк зчинивсі в ґіда в обидвох осєдочьках гармидер." (Eng: *He snapped out of it only in the evening, after a chaos started in both of the old man's*



**Mistral (manual + synthetic):** "Отямивси аж увечері, коли зчинився в діда в обох хатах гармидер." (Eng: *He came to his senses in the evening, when a chaos broke out in both grandfather's houses.*)
BLEU: 31.46, chrF++: 46.62, TER: 54.55, Fluency: 4, Adequacy: 5, Dialect: 3

**Mistral (manual only):** "Отємивси аж увечєрі, єк зчєнивси в діда в обох хатах гармидер." (Eng: *He came to his senses only in the evening, after stirring up a mess in both of the grandfather's houses.*)
BLEU: 34.39, chrF++: 56.92, TER: 45.45, Fluency: 4, Adequacy: 5, Dialect: 4

**LLAMA (manual + synthetic):** "Отямивси аж у вечєрє, коли зчинивси в діда в обох хатях гармидер." (Eng: *He came to his senses in the evening, when a mess started in both grandfather's houses.*)
BLEU: 20.45, chrF++: 46.42, TER: 63.64, Fluency: 4, Adequacy: 4, Dialect: 3

# Fun Examples

*Source (Ukr.):*

Иванчік сказав: “Час і мені йти на толоку”

*Model output:*

Иванчік **вікажу**: “Чєс і мині иди на толоку”

*Real Hutsul:*

Иванчік сказав: “Чєс и мині йти на толоку”

*Source (Ukr.):*

Та й так над тим буком у руках «бога» й «чорта»

*Model output:*

Та й так над тим буком у руках «біга» й «мольфа»

*Real Hutsul:*

Тай так моцно над тим буком у руках «бога» й «чорта»

*Source (Ukr.):*

Аж кришив її: З золотого моря
Вилетіла золота золотариця

*Model output:*

Аж кришив її: «З золотого моря
**Валентина** золота золотариця

*Real Hutsul:*

Аж кришив її: «З золотого моря
Вилетіла золота золотариця

# Our contribution

**First digital resources for the Hutsul dialect**

- a 9 852-pair Hutsul ↔ Standard-Ukrainian parallel corpus and
- a 7 320-entry dialect-to-standard dictionary, compiled from literature, folklore and blogs

**Data-scarcity solution via advanced RAG augmentation** – we designed a rule-guided RAG pipeline that injects phonological and morphological rules into GPT-4o prompts, creating **52 142 high-quality synthetic sentence pairs** further filtered with automatic alignment checks

**Parameter-efficient adaptation of compact open LLMs** – two open-source models (Mistral-7B-Instruct v0.3 and LLaMA-3.1 8B) are fine-tuned with LoRA/QLoRA on the combined manual + synthetic data, making dialect translation feasible on a single consumer GPU

**Comprehensive, dialect-aware evaluation framework** – performance is judged with BLEU, chrF++, TER and a GPT-4o "LLM-judge" scoring Fluency, Adequacy and Dialectal Quality, mitigating the blind spots of standard n-gram metrics for non-standard orthographies

.

**Empirical finding**: small fine-tuned models beat GPT-4o – the best 7 B model surpasses zero-shot GPT-4o across all automatic metrics and in LLM-based human-like scores, demonstrating the value of dialect-specific tuning even with modest model sizes

**Open-source commitment** – all code, data, prompts and trained LoRA weights are released on GitHub to encourage further work on Ukrainian dialects and other low-resource varieties

# Data collection

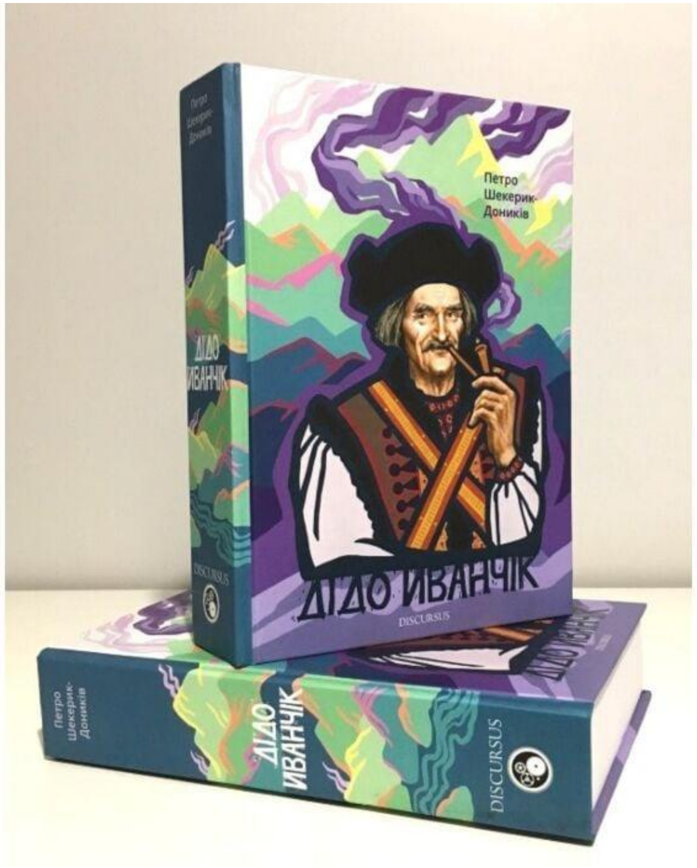
**9 852 manually aligned sentence pairs**

- Hutsul originals from Дідо Иванчік (foundational dialect novel), ethnographic transcripts, folklore websites and dialect blogs
- Standard-Ukrainian references taken from a modern bilingual edition or hand-translated

**7 320-entry Hutsul ↔ Ukrainian dictionary**

- Scraped Дідо Иванчік + five open-access lexicons (“Hutsul Hovir”, “Dictionary of Hutsul Words”, etc.)

*(!) Lexicon is biased toward the vocabulary found in folk-lore, thus lacks diversity in news, science, or politics.*



# Synthetic data generation pipeline

**Grammar Rule Extraction**

- Prompt GPT-4o to distill phonological, morphological, and syntactic rules unique to the Hutsul dialect.
- Output: structured rule set → reusable prompt template.

**RAG Index Creation**

- Embed every sentence of “Дідо Иванчік” with text-embedding-3-large and store in a vector index to act as an authentic dialect reference base.

**Candidate Retrieval**

- Sample Standard-Ukrainian lines from the UberText corpus; for each, retrieve the top-3 semantically closest Hutsul sentences from the index.

**Prompt Assembly**

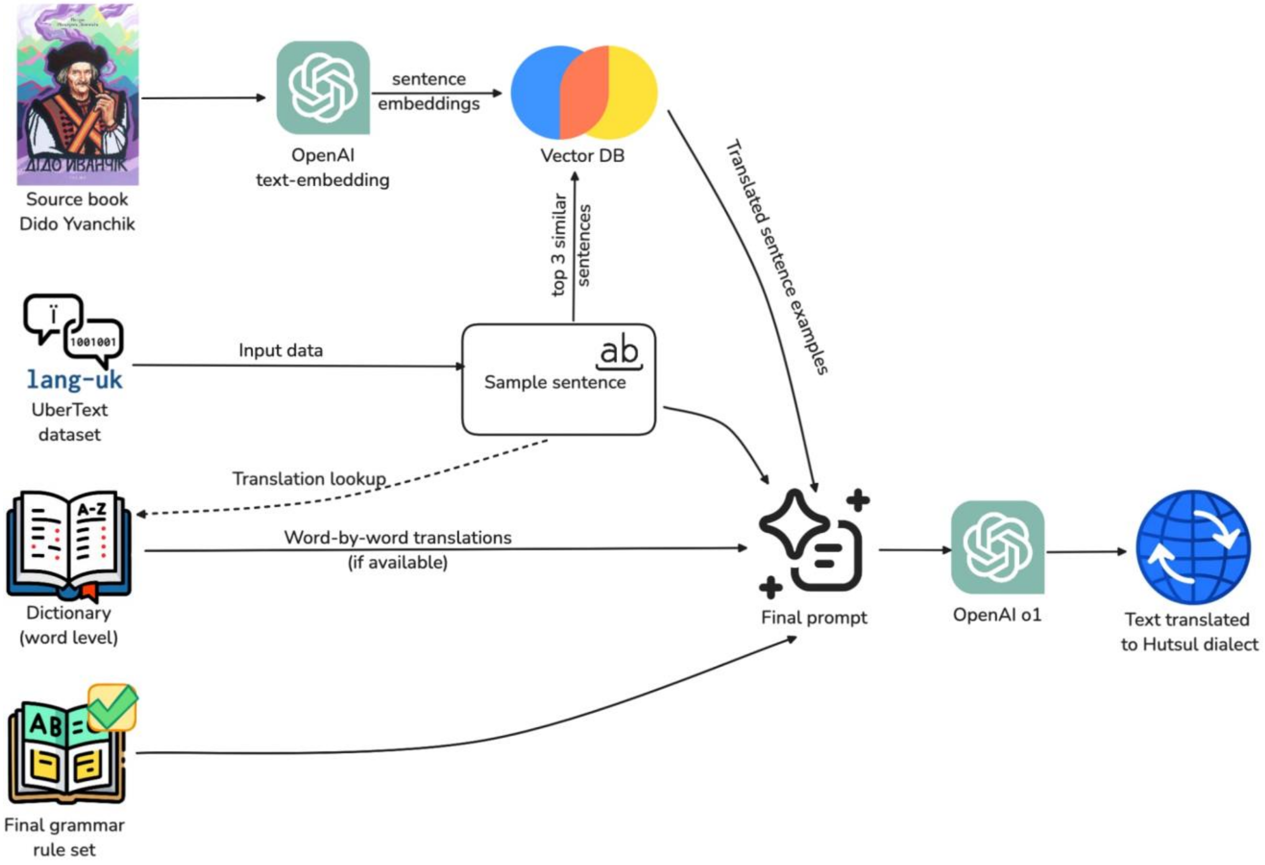
- extracted grammar rules
- three retrieved Hutsul examples
- source Standard-Ukrainian sentence

⇒ one rich prompt ready for generation.

**Dialect Generation & Filtering**

- Instruct GPT-4o to translate the source into Hutsul, adhering to the rules and style cues. Post-process with alignment/character-ratio filters to keep only high-quality synthetic pairs.

**Result: 52 k clean Standard ↔ Hutsul sentence pairs that enlarge the training corpus times**



# Alignment Quality Metrics

**Three metrics** to check if sentence pairs align well:

## 1.U-src (Unaligned source)

- Proportion of source language characters that cannot be aligned to target
- **Lower values = better alignment quality**

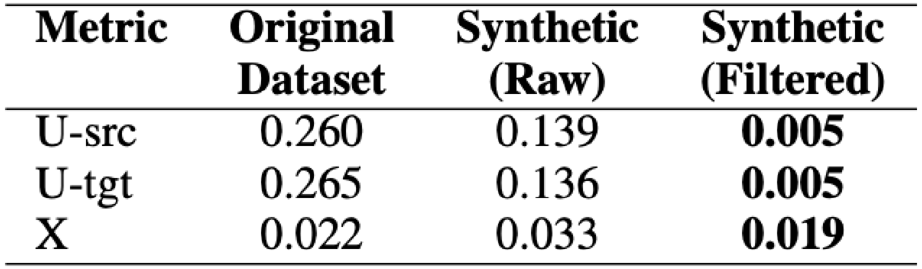
## **2.U-tgt** (Unaligned target)

- Proportion of target language characters that cannot be aligned to source
- **Lower values = better alignment quality**

## **3.X** (Crossing alignments)

- Proportion of word alignment pairs that cross/swap positions
- **Shows structural differences** between source and target

*Calculated using automatic word alignment tools*

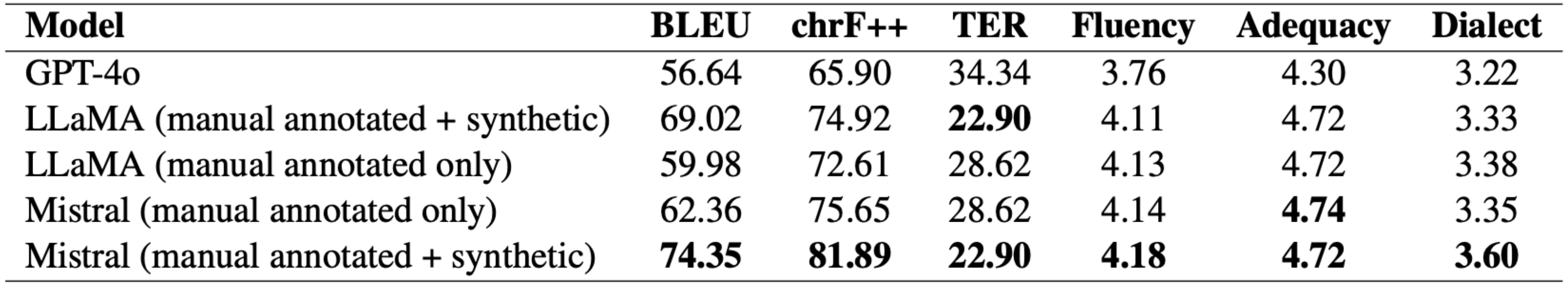


| **Metric** | **Original Dataset** | **Synthetic (Raw)** | **Synthetic (Filtered)** |
|---|---|---|---|
| U-src | 0.260 | 0.139 | **0.005** |
| U-tgt | 0.265 | 0.136 | **0.005** |
| X | 0.022 | 0.033 | **0.019** |

Table 1: Alignment quality metrics comparison between the original dataset, raw synthetic dataset, and synthetic dataset after alignment-based filtering.

# Model fine-tuning and evaluation

- **Mistral-7B-Instruct v0.3** – Chosen for its performance-to-size ratio. It outperforms some larger models on many benchmarks, supports multilingual instructions, and includes explicit support for Ukrainian.
- **LLaMA-3.1 8B Instruct** – The instruction-tuned version of LLaMA 3.1 8B. This model has a strong multilingual support and improved instruction-following ability, making it a good candidate for low-resource translation.
- Each model was trained for 3 epochs using LoRa on two dataset variants:
  - a manually created Hutsul–Ukrainian parallel corpus
  - an extended version that included combined manual and filtered synthetic data.



| Model | BLEU | chrF++ | TER | Fluency | Adequacy | Dialect |
|---|---|---|---|---|---|---|
| GPT-4o | 56.64 | 65.90 | 34.34 | 3.76 | 4.30 | 3.22 |
| LLaMA (manual annotated + synthetic) | 69.02 | 74.92 | **22.90** | 4.11 | 4.72 | 3.33 |
| LLaMA (manual annotated only) | 59.98 | 72.61 | 28.62 | 4.13 | 4.72 | 3.38 |
| Mistral (manual annotated only) | 62.36 | 75.65 | 28.62 | 4.14 | **4.74** | 3.35 |
| Mistral (manual annotated + synthetic) | **74.35** | **81.89** | **22.90** | **4.18** | **4.72** | **3.60** |

Table 2: Automatic and LLM-based evaluation results. BLEU, chrF++, and TER are computed with sacreBLEU. Fluency, adequacy, and dialect quality are rated by GPT-4o (1–5 scale).

# Conclusions and future work

- We have created comprehensive and high-quality Hutsul ↔ Standard-Ukrainian parallel corpus as well as
- Novel method of generating synthetic dataset for low-resource dialects was developed
- Proof-of-concept: small open LLMs can outperform GPT-4o when dialect-tuned

- What's next? **Enrichment**!

Ґєкую за увагу! (Thanks!)