UNLP 2026

The Fifth Ukrainian Natural Language Processing Conference (UNLP 2026)

Proceedings of the Conference

May 29-30, 2026

The UNLP organizers gratefully acknowledge the support from the following sponsors.

UNLP 2026 Partners:


Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-359-3

Welcome to UNLP 2026

We warmly welcome you to the Fifth Ukrainian Natural Language Processing Conference, held on May 29–30, 2026!

The conference brings together leading professionals from academia and industry who develop language resources, tools, and NLP solutions for the Ukrainian language. UNLP provides a platform for discussion and sharing of ideas, fosters collaboration between different research groups, and improves the visibility of the Ukrainian research community worldwide.

This year, the conference expanded the list of topics to include more aspects of Ukrainian language processing, such as multimodality, speech, optical character recognition, agentic systems, educational tools, and more. We received a record 52 submissions, of which 22 were accepted to be presented at the conference. The paper topics follow the global NLP trends and focus on the customization and application of large language models to a variety of tasks in Ukrainian. Almost half of the papers introduce new datasets for training and benchmarking. We are immensely grateful to the program committee for their careful and thoughtful reviews of the papers submitted this year!

UNLP 2026 will host three keynote speeches. Anna Rogers, Associate Professor at IT University of Copenhagen, will discuss the future of NLP in the age of large language models. Mykola Haltiuk, PhD Student at AGH University of Krakow, will present his work on model-aware tokenizer transfer. Yurii Paniv, PhD Student at Ukrainian Catholic University, will explore approaches to automating research.

The fifth UNLP will feature the Shared Task on Multi-Domain Document Understanding. The shared task challenges AI systems to identify relevant information in collections of domain-specific documents and to generalize across domains. Fifteen teams submitted their solutions, and four shared task papers were accepted for presentation at the conference.

To address the topic of responsible AI, UNLP 2026 will host a panel discussion on ethics of AI usage with Oleksii Molchanovskyi, Chief Innovation Officer at Ukrainian Catholic University, and Olena Andriienko, Chief LegalTech Officer at Publicis Groupe.

We are grateful to the Applied Sciences Faculty of the Ukrainian Catholic University for hosting the conference. We thank MacPaw for its financial support, Preply for labeling the shared task dataset, and Superhuman for providing Grammarly Pro to UNLP authors. Our thanks go to NaUKMA's Faculty of Computer Sciences for technical support. Finally, we thank Kyiv School of Economics, Texty.org.ua, AI House, and all our partners for their promotional and outreach support.

We are looking forward to the conference and anticipate lively discussions on Ukrainian NLP!

Organizers of UNLP 2026,
Roman Kyslyi, Mariana Romanyshyn, Olena Nahorna, Oleksii Ignatenko, and Andrii Hlybovets

Organizing Committee

Workshop Organizing Committee

Roman Kyslyi, Kyiv School of Economics, Ukraine
Mariana Romanyshyn, Superhuman, Ukraine
Olena Nahorna, Preply, Germany
Oleksii Ignatenko, Ukrainian Catholic University, Ukraine
Andrii Hlybovets, National University of Kyiv-Mohyla Academy, Ukraine

Shared Task Organizing Committee

Roman Kyslyi, Kyiv School of Economics, Ukraine
Mariana Romanyshyn, Superhuman, Ukraine
Olena Nahorna, Preply, Germany
Volodymyr Sydorskyi, Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Ukraine
Nataliia Romanyshyn, Ukrainian Catholic University, Texty.org.ua, Ukraine

Program Committee

Program Committee

Andrii Liubonko, EPAM Systems, Ukraine
Andrii Yaroshevskyi, Respeecher, Ukraine
Anna Rogers, IT University of Copenhagen, Denmark
Anton Bazdyrev, Dun&Bradstreet, Ukraine
Artem Chernodub, Zendesk, Poland
Bogdan Babych, Heidelberg University, Germany
Bohdan Didenko, WebSpellChecker LLC, Ukraine; Lviv Polytechnic National University, Ukraine
Hasan Abu-Rasheed, Goethe University Frankfurt, Germany
Igor Samokhin, Adimen, Ukraine
Iuliia Makogon, Newxel, Ukraine
Kostiantyn Omelianchuk, Superhuman, Germany
Maksym Tarnavskyi, Shelf, Poland
Maria Shvedova, National Technical University "Kharkiv Polytechnic Institute", Ukraine; Friedrich Schiller University Jena, Germany
Mark Norris, Superhuman, USA
Mykola Khandoga, AryaXAI, France
Mykola Sazhok, CyberMova, Ukraine; Institute of Information Technologies and Systems of the NAS of Ukraine
Natalia Grabar, CNRS, Université de Lille, France
Natalia Kotsyba, Samsung Research Poland, Poland
Nataliia Cheilytko, Friedrich Schiller University Jena, Germany
Nataliya Polyakovska, SoftServe, USA
Nazarii Drushchak, Ukrainian Catholic University, Ukraine
Oleksandr Marchenko, Taras Shevchenko National University of Kyiv, Ukraine
Oleksandr Skurzhanskyi, Superhuman, Germany
Oleksii Molchanovskyi, Ukrainian Catholic University, Ukraine
Oleksii Turuta, Kharkiv National University of Radio Electronics, Ukraine
Oleksiy Syvokon, Lviv Polytechnic National University, Ukraine
Olha Kanishcheva, Friedrich Schiller University Jena, Germany
Serhii Hamotskyi, Anhalt University of Applied Sciences, Germany
Svitlana Galeshchuk, Université Paris Dauphine, France; West Ukrainian National University, Ukraine
Taras Lehinevych, Amazon, Ireland
Taras Shevchenko, Proxet, Ukraine
Taras Ustyianovych, Lviv Polytechnic National University, Ukraine
Tatjana Scheffler, Ruhr-Universität Bochum, Germany
Thierry Hamon, Université Paris-Saclay, CNRS, LIMSI & Université Sorbonne, France
Uta Seewald-Heeg, Anhalt University of Applied Sciences, Germany
Valentyna Robeiko, Taras Shevchenko National University of Kyiv, Ukraine
Vera Schmitt, Technische Universität Berlin, Germany
Veronika Solopova, Technische Universität Berlin, Germany
Volodymyr Kyrylov, OpenAI, USA
Volodymyr Mudryi, Ukrainian Catholic University, Ukraine
Volodymyr Taranukha, Taras Shevchenko Nationa University of Kyiv, Ukraine
Yevhen Kostiuk, Aarhus University, Denmark
Yevhenii Azarov, Respeecher, Ukraine

Yurii Paniv, Ukrainian Catholic University, Ukraine

Table of Contents

Program

Friday, May 29, 2026

10:00 - 10:15 *Opening Remarks*

10:15 - 11:15 *Morning Session: Translation & Evaluation*

SimIdioms: A Corpus and Benchmark for Ukrainian Idiom Translation
Yaryna Petruniv, Iuliia Makogon and Roman Kyslyi

Semantic Fidelity Versus Literary Quality: A Construct Validity Study of Neural Machine Translation Metrics
Dmytro Chaplynskyi, Ivan Kulynych, Maria Shvedova and Lesia Ivashkevych

Professional Translators Versus Quality Estimation Models: Reliability and Agreement in English-Ukrainian Translation Evaluation
Dmytro Chaplynskyi, Kyrylo Zakharov and Lesia Ivashkevych

11:15 - 11:45 *Morning Coffee Break*

11:45 - 13:05 *Morning Session: Core NLP & Linguistic Resources*

Data-Efficient Adaptation of Multilingual LLMs to Ukrainian
Yurii Paniv, Bohdan Didenko, Mykola Haltiuk, Vladyslav Humennyy, Andrian Kravchenko, Roman Kyslyi, Viktoriia Makovska, Artem Orlovskyi, Bohdan Ruban, Maksym-Yurii Rudko, Anastasiia Senyk, Nazarii Drushchak, Dmytro Chaplynskyi and Mariana Romanyshyn

Dictionary-Based Speculative Decoding for Non-Latin-Script Languages
Oleksiy Syvokon

How Far Can Prompting Go for Minimal-Edit Ukrainian Grammatical Error Correction?
Kateryna Karpo and Artem Chernodub

Entropy of Ukrainian
Anton Lavreniuk, Mykyta Mudryi and Markiian Chaklosh

13:05 - 14:15 *Lunch*

14:15 - 15:15 *Keynote: Anna Rogers, "What's Next for NLP after LLMs?"*

Friday, May 29, 2026 (continued)

15:15 - 16:15 *Afternoon Session: Language Proficiency & Paraphrasing*

Toward a Gold-Standard Benchmark for Evaluating Ukrainian Language Proficiency in LLMs
Svitlana Galeshchuk, Yuliia Maksymiuk, Yuliia Chernobrov, Nina Stankevych, Oleksandra Antoniv, Nataliia Faryna and Oksana Popkova

Automated CEFR-Level Assignment for Ukrainian Texts
Olha Kanishcheva and Mikhail Kopotev

Mining Native Ukrainian Paraphrases: A Multi-Source Comparison
Vladyslav Fesenko, Hanna Dydyk-Meush and Volodymyr Mudryi

16:15 - 16:45 *Afternoon Coffee Break*

16:45 - 18:00 *Keynote: Mykola Haltiuk, "Model-Aware Tokenizer Transfer"*

18:00 - 18:10 *End of Day 1*

Saturday, May 30, 2026

10:00 - 10:10 *Day 2 Welcome*

10:10 - 11:10 *Morning Session: Speech, Multimodality & OCR*

Scaling ASR for Hutsul Dialect: Multi-Speaker Data Collection, Enhanced Transcription and Cross-Speaker Evaluation
Artem Orlovskyi, Zakhar Guzii, Bohdan Onyshchenko, Roman Kyslyi and Pavlo Khomenko

UkrSL: Towards a Ukrainian Continuous Sign Language Dataset
Oleksandr Sobetskyi, Maryna Kosse, Roman Kyslyi and Angelina Savchenko

Digitizing Old Ukrainian Texts: A Prompt-Based OCR Pipeline and Evaluation Dataset
Dmytro Chaplynskyi and Hanna Dydyk-Meush

11:10 - 11:40 *Morning Coffee Break*

11:40 - 13:00 *Morning Session: Applications, Systems & Tools*

Improving Domain-Specific Translation from English into Ukrainian with Retrieval-Augmented Generation
Anton Shpigunov

UAReviews: A Multi-Task Ukrainian Dataset for Emotion and Intent Classification
Roman Kyslyi, Ihor Pysmennyi and Denys Mykhailov

Graph-Based Detection of Disinformation Narrative Diffusion between Russian and Ukrainian Telegram Channels
Yuliia Vistak, Viktoriia Makovska, Vera Schmitt and Veronika Solopova

A Two-Axis Framework for Analyzing Ukrainian Dialogues
Artem Korotenko and Roman Kyslyi

13:00 - 14:15 *Lunch*

14:15 - 14:35 *Afternoon Session: Applications, Systems & Tools (continued)*

Saturday, May 30, 2026 (continued)

Belief Propagation in LLM World Models: Measuring Strategic Information Bias with Prediction Markets
Mykola Khandoga, Yevhen Kostiuk, Anton Polishko, Yurii Filipchuk, Kostiantyn Kozlov, Dmytro Zamriy and Artur Kiulian

14:35 - 15:35 *Keynote: Yurii Paniv, "On Automating Research"*

15:35 - 16:55 *Afternoon Session: Shared Task on Multi-Domain Document Understanding*

The UNLP 2026 Shared Task on Multi-Domain Document Understanding
Volodymyr Sydorskyi, Nataliia Romanyshyn, Roman Kyslyi and Olena Nahorna

RAG Pipeline Strategies for Ukrainian Multi-Domain Document Understanding Task
Mykola Nosenko and Pavlo Kilko

An End-to-End Ukrainian RAG for Local Deployment. Optimized Hybrid Search and Lightweight Generation
Mykola Trokhymovych, Yana Oliinyk and Nazarii Nyzhnyk

Qwen Goes Brrr: Off-the-Shelf RAG for Ukrainian Multi-Domain Document Understanding
Anton Bazdyrev, Oleksandr Kharytonov, Artur Khodakovskyi, Ivan Havlytskyi and Ivan Bashtovyi

17:00 - 18:00 *Panel Discussion: Oleksii Molchanovskyi, Olena Andriienko, and Roman Kyslyi, "Ethics of AI Usage"*

18:00 - 18:15 *Closing Words*

Improving Domain-Specific Translation from English into Ukrainian with Retrieval-Augmented Generation

Anton Shpigunov
Taras Shevchenko National University of Kyiv
14 Taras Shevchenko Blvd., Kyiv 01601, Ukraine
shpigunov@knu.ua

Abstract

Large language models have demonstrated competence as language translators, including for lower-resourced languages like Ukrainian. However, in specialized or novel domains, translation quality can suffer without adequate lexical and stylistic reference material. We present a retrieval-augmented approach to English–Ukrainian machine translation in a narrow domain (a private legal/military bilingual corpus), where semantically similar translation units retrieved via vector embeddings are provided as in-context examples to the LLM. We evaluate three open-weight Gemma 3 models (4B, 12B, 27B) against Gemini 3 Flash as a baseline across five augmentation conditions ($k \in \{0, 3, 5, 10, 25\}$) on a 2,581-pair index/test set split into 2,323 indexed pairs and 258 test pairs. We find that context augmentation yields statistically significant improvements in both ChrF++ and COMET for all models, with the smallest model's COMET improving by $+0.076$ at $k = 3$. However, smaller models exhibit context saturation: the 4B model's performance peaks at $k = 10$ and degrades with additional context, losing 9.72 ChrF++ points and 0.007 COMET between $k = 10$ and $k = 25$, while larger models continue to benefit.

1 Introduction and Related Work

It has been established that large language models appear to be competent language translators (Ye et al., 2025), including for lower-resourced languages (Enis and Hopkins, 2024). However, their capabilities rely on static, parametric knowledge encoded during training. This can lead to poor adaptation to new domains and difficulties translating rare, specialized, or culturally specific terminology. Additionally, frontier models provided by leading labs over an API may not be optimal for translating text in multiple scenarios, including medical, government, defence, and other sensitive settings where there are concerns over privacy, sensitive topics, or national security priorities.

To address limitations of knowledge acquired during the fixed training phase, the text-generation capabilities of LLMs have been extended using a variety of approaches. As one instance, retrieval-augmented generation (RAG) (Lewis et al., 2020) has been integrated into translation tasks to create hybrid architectures that ground models in external, non-parametric knowledge bases.

In human translation and localization, assistance based on retrieval of previously translated segments (translation memory) has been industry standard for decades, although such retrieval is overwhelmingly based on fuzzy string matching. As neural machine translation (NMT) emerged, early efforts sought to integrate this existing TM technology directly into neural workflows; for example, Bulte and Tezcan (2019) demonstrated that augmenting NMT input with fuzzy TM matches significantly boosted translation performance.

Expanding on the value of retrieved examples, subsequent non-LLM approaches moved beyond surface-level fuzzy matching. Zhang et al. (2018) employed a search engine to retrieve previously seen translation examples and incorporate them into the NMT model's decoding process, while Khandelwal et al. (2020) introduced kNN-MT, which augmented NMT decoding with dense vector similarity search over a massive datastore of cached examples to improve domain adaptation at test time without additional training.

Building upon these dense retrieval methods in the era of large language models, Donthi et al. (2024) utilized vector-based cosine similarity to retrieve culturally specific terms such as idioms from external databases, successfully improving LLM translation quality in both high-resource (Chinese) and low-resource (Urdu) languages over direct-translation baselines. Li et al. (2023) used an external knowledge base (IdiomKB) and employed

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 1–11

retrieval augmentation to provide context (figurative meanings of idioms) to smaller LLMs, resulting in better translations according to their evaluation. These approaches demonstrate the value of adding non-parametric, flexible human knowledge to LLMs for translation tasks.

Taking these developments further with semantic similarity search based on vector embeddings and pairing it with open-weight language models, a localized RAG approach emerges as a viable means to improve translation quality and consistency, especially where pre-existing training data is insufficient or otherwise inadequate. In other words, we expect a beneficial in-context learning effect on translation quality (Brown et al., 2020). With respect to all of the above, we specifically hypothesize that:

> *smaller, open-weight models augmented with example sentences retrieved using vector similarity search can improve MT quality and consistency within a specific domain, approaching or even surpassing MT quality provided by larger models without such augmentation.*

We demonstrate and assess this RAG-MT approach, its effect on MT quality as assessed using two automated metrics, and discuss our findings and implications for further research. The code, prompt template, and de-identified evaluation outputs supporting this work are released at `https://github.com/shpigunov/unlp2026-rag-mt`.

2 Data

2.1 Dataset Selection

We have considered two principal sources for translation data:

- WMT-2025 `uk-en` corpus with its multiple sub-components;
- a private en–uk corpus of 2,864 sentences which constitutes a full translation of the US Unified Code of Military Justice into Ukrainian.

Ultimately, we chose the private dataset out of two main considerations. Firstly, there is no way to establish whether WMT data has been used in training of frontier models (i.e., that the LLM does not already contain this data in its training set). Secondly, the WMT dataset is `uk-en`, and reversing the direction would introduce translationese artifacts into the source sentences. Zhang and Toral (2019) demonstrated that translationese in test sets inflates MT evaluation scores and can alter system rankings; Graham et al. (2020) further show that reverse-created test data compounds this problem by reducing the statistical power of human evaluations, potentially masking real quality differences between systems. For these reasons, Graham et al. recommend against using reverse-created test data in MT evaluation.

2.2 "Train"–Test Split

The split between the reference set for populating the vector index (we shall refer to it further as the "train" set for convenience and convention, although strictly speaking, no model was trained on it) has been chosen at 90% train and 10% test. This somewhat skewed split is influenced by the relatively small size of the private dataset. To ensure statistical significance on a smaller test set, we employed significance testing protocol, as explained below.

2.3 Data Preparation

Our initial exploration has shown that the chosen dataset contains repetitive segments such as references to section names, boilerplate legal language, and similar formulaic text. On a first evaluation, this led to what we assessed as leakage of training data into the testing set.

To mitigate this, we implemented a rigorous two-stage deduplication pipeline:

1. **Exact Deduplication:** All translation units first underwent strict text normalization involving HTML unescaping, NFKC Unicode normalization, markup tag removal, downcasing, and aggressive number and punctuation masking to eliminate superficial character and casing variations. We then discarded exact duplicate pairs by comparing hashes of the concatenated, normalized source and reference segments. This step removed 203 sentence pairs from the set.

2. **Near Deduplication (Fuzzy Matching):** To identify and remove near-duplicates differing only by minor syntactic variations or formatting, we applied locality-sensitive hashing (LSH) using MinHash signatures. Operating exclusively on the normalized source text, we

evaluated character-level 5-grams over 64 permutations, discarding any sentence that exhibited a Jaccard similarity of 0.90 or higher against previously indexed segments. This step further removed 80 pairs from the set.

From the 2,581 de-duplicated sentence pairs, we created a strict 90:10 train–test split (2,323 train segments and 258 test segments). To prevent data leakage, identical normalized source segments were grouped together prior to splitting, ensuring that any remaining exact source-text variants were assigned exclusively to a single split partition.

The train subset was encoded using the multilingual embedding model `baai/bge-m3` (Chen et al., 2024), which produced 1024-dimensional dense vectors for the English source segments; these were indexed in `qdrant` for cosine-similarity retrieval.

3 Experiment

3.1 Setup

To validate our hypothesis stating that smaller models augmented with retrieval can approximate or surpass translations generated by frontier models without additional context, we implemented a custom retrieval-augmented machine-translation (RAG-MT) prototype and designed a controlled evaluation framework.

Our training phase consisted of embedding the source sentence from the train set and inserting the sentence pair into the vector index.

For the testing phase, the prototype executed the following sequence for each test pair:

1. embed the source sentence with the same vector embedding model;
2. use the source-segment embeddings to perform a cosine-similarity search in the vector index populated during the training phase, and obtain k similar source-target pairs; no similarity computation over target text is performed;
3. template the source sentence and retrieved similar sentence pairs into the translation prompt, essentially creating a dynamic few-shot prompt for the LLM;
4. run LLM inference on the prompt and extract only the target text;
5. for each condition, apply ChrF++ and COMET metrics to the entire test set.

3.2 Embedding Model

For this experiment we chose `baai/bge-m3` (Chen et al., 2024). The model was chosen due to its published open weights enabling reproducibility, its strong performance on shorter multilingual segments, and the absence of a fine-tuning requirement. For the experiment it was served from Hugging Face's infrastructure, but in a local privacy-preserving setup, local inference is possible.

We initially ran the same experiment using a closed-source embedding model (`text-embedding-3-large` by OpenAI), and found that the overall metrics were slightly lower when using the open-weight embedder, but the overall dynamic described in the Results was the same. This implies that while the influence of the embedding model is not as significant as that of the LLM, the choice of embedder in a RAG-MT pipeline is more than a mere infrastructural consideration.

A conceivable way to strengthen the pipeline is to introduce a re-ranker model to increase the internal variety of retrieved examples and prevent near-duplication, a direction that can be explored in further work. Furthermore, using different retrieval methods altogether (Bouthors et al., 2024) would be a promising ablation to explore, but falls outside the scope of this investigation.

3.3 Large Language Models

Model Family. For the large language models, we selected Google's Gemini/Gemma model family. For the commercial baseline, we use `gemini-3-flash` due to its reasonable inference availability, speed, and cost.

For the target open-weight models, we use Gemma 3 family models (Team et al., 2025) in different sizes with 4B, 12B, and 27B parameters, deployable on consumer hardware. Gemma 3 is also convenient due to its proximity to our baseline in terms of similarities in tokenizer, attention mechanism, and training infrastructure. We do note differences, however, with Gemini models being sparse mixture-of-experts systems and not disclosing their exact number of parameters.

For inference infrastructure, we used OpenRouter while controlling for quantization (excluding quantized models and limiting to `bf16` only to maintain consistency and reproducibility).

To isolate the impact of retrieved context and measure optimal context-size limits, we evaluate

outputs of four models: `gemini-3-flash` as the commercial baseline, and then the open-weight Gemma 3 models `gemma3-4b`, `gemma3-12b`, and `gemma3-27b`.

Completion Parameters. Following the findings of Li et al. (2025), we select a lower completion temperature at `0.1` to increase repeatability and reduce probability of hallucination. The reasoning setting on `gemini-3-flash` was set to `minimal`, and the rest of the parameters were kept to provider defaults.

Each of these four models was evaluated with different context augmentation sizes: $k = 0$ to assess baseline model performance, and then $k \in \{3, 5, 10, 25\}$.

3.4 Choice of Evaluation Metrics

We rely on automated evaluation metrics that best reflect the rich morphology and flexible word order of Ukrainian as a target language (Shpigunov, 2025). Word-level n-gram metrics such as BLEU and edit-distance metrics such as TER over-penalize the morphological variants and word-order shifts that are natural in Ukrainian, treating different inflections of the same lemma as full mismatches and discounting permissible reorderings. Character-level matching with ChrF++ sidesteps the morphology problem by scoring partial overlap across word forms, while neural reference-based metrics such as COMET evaluate semantic equivalence in a multilingual contextual embedding space; both have been shown to correlate more strongly with human judgment than surface-level metrics for Ukrainian-target translation (Shpigunov, 2025). Pairwise differences between the conditions are evaluated segment by segment on the identical test set.

- **Lexical and Character-level:** We report ChrF++ (Popović, 2017) to capture character n-gram matches, which is particularly robust for morphologically rich target languages like Ukrainian.
- **Semantic:** We utilize neural reference-based evaluation via COMET (Rei et al., 2020) (variant used: `unbabel/wmt22-comet-da`) to measure semantic accuracy and fluency improvements over the baseline.

3.5 Tests of Statistical Significance

To evaluate the statistical significance of improvements over the unaugmented baseline ($k = 0$), we employ paired bootstrap resampling on both ChrF++ and COMET scores. To control the family-wise error rate across multiple comparisons, we apply the Holm–Bonferroni correction to the raw p-values.

3.6 Results

The performance of the selected models with the set context conditions on ChrF++ and COMET can be seen in Tables 1 and 2, respectively.

Detailed paired-bootstrap significance results are reported in Table 3 in Appendix A.

4 Discussion

By translating the same 258-sentence test set with retrieval from a 2,323-sentence set on a grid of four models by five conditions ($k \in \{0, 3, 5, 10, 25\}$) and keeping other variables such as prompts and completion temperature unchanged (temperature was kept fixed at `0.1` for all models tested), we have shown in a controlled experiment that context augmentation using retrieved similar sentences leads to a pronounced increase in both character-based and neural metrics (ChrF++ and COMET-da, respectively), with the extent of this increase being dependent on model capacity and the size of the provided context. The aggregate trend is shown in Figure 1.

4.1 Baseline Performance ($k = 0$)

In the zero-shot baseline, larger models have expectedly outperformed smaller ones, with even the smaller commercial cloud-only Gemini variant, `gemini-3-flash`, scoring higher (COMET 0.896 and ChrF++ 62.47) than the largest open-weight Gemma 3 variant, `gemma3-27b` (COMET 0.875 and ChrF++ 60.73).

Within the Gemma 3 family, the largest 27B variant scored the highest at COMET 0.875 and ChrF++ 60.73, followed by 12B (COMET 0.864, ChrF++ 54.85), and finally 4B with a markedly lower performance (COMET 0.812, ChrF++ 46.87), which reiterates the known correlation between model size and machine-translation performance (cf. Tables 1 and 2).

4.2 Impact of Context Augmentation

Augmenting the translation prompt with semantically similar translated sentences from the training portion of the same-domain corpus ($k \in \{3, 5, 10, 25\}$) has consistently and significantly ($p < 0.01$ for all paired-bootstrap configurations)

Model	$k = 0$	$k = 3$	$k = 5$	$k = 10$	$k = 25$
gemini-3-flash	62.47	77.49	77.64	78.62	78.76
gemma3-27b	60.73	77.65	79.14	79.82	79.61
gemma3-12b	54.85	71.19	73.34	74.94	75.23
gemma3-4b	46.87	68.26	68.28	68.81	59.09

Table 1: Summary of evaluation metrics. ChrF++ by model, by condition.

Model	$k = 0$	$k = 3$	$k = 5$	$k = 10$	$k = 25$
gemini-3-flash	0.896	0.931	0.934	0.935	0.936
gemma3-27b	0.875	0.920	0.928	0.927	0.931
gemma3-12b	0.864	0.906	0.914	0.911	0.916
gemma3-4b	0.812	0.888	0.892	0.892	0.885

Table 2: Summary of evaluation metrics. COMET by model, by condition.

improved assessed MT quality for all models considered.

Small Model, Big Gains. The most drastic relative improvements in translation quality were observed in the smallest model assessed, gemma3-4b. Augmenting the prompt with $k = 3$ examples led to a strong increase in both metrics, increasing COMET by +0.076 (from 0.812 to 0.888) and ChrF++ by +21.39 (from 46.87 to 68.26) over the zero-shot baseline, bringing the performance of the smallest model above the $k = 0$ baseline for gemma3-27b by COMET and above gemini-3-flash by ChrF++.

Reaching for the Cloud. Context augmentation has allowed smaller, open-weight models to surpass the MT quality of unaugmented frontier models. Even the smallest gemma3-4b, which can comfortably run on consumer hardware, when augmented with $k = 3$, is able to closely trail the baseline performance of gemini-3-flash. On the high end, gemma3-27b, which may require higher-end consumer hardware to run but can still be used offline, with $k \geq 5$ augmentation was able to surpass gemini-3-flash by ChrF++ and closely approach it by COMET scores. This allows translators to get a meaningful MT assist from offline open-weight models while translating specialized documents, without the need for potentially sensitive information to go to a cloud provider.

Diminishing Returns, Context Saturation and Context Rot. At the same time, we observed a marked divergence in how models of different sizes handle larger context windows ($k \geq 10$). Larger models like gemini-3-flash, gemma3-27b, and gemma3-12b demonstrate gains in both metrics with an increasing number of similar segments, but these gains exhibit sharply diminishing returns. However, the smallest model, gemma3-4b, demonstrates regression in both metrics when given larger contexts. This suggests that smaller models are more susceptible to confusion or hallucination when overwhelmed with excessive reference examples, a phenomenon referred to as context saturation or context rot (Bianchi et al., 2025; Hong et al., 2025).

Another observation with respect to the smallest model, gemma3-4b: descriptive statistics such as the minimum, maximum, mean, and standard deviations on both metrics (see Appendix B, Figures 2 and 3) point to similar conclusions. Overburdened with context, the smallest model appears to produce more inconsistent and low-quality translations, while larger models become more consistent and produce higher-assessed output more often. For larger models, adding more examples appears to increase consistency in both metrics.

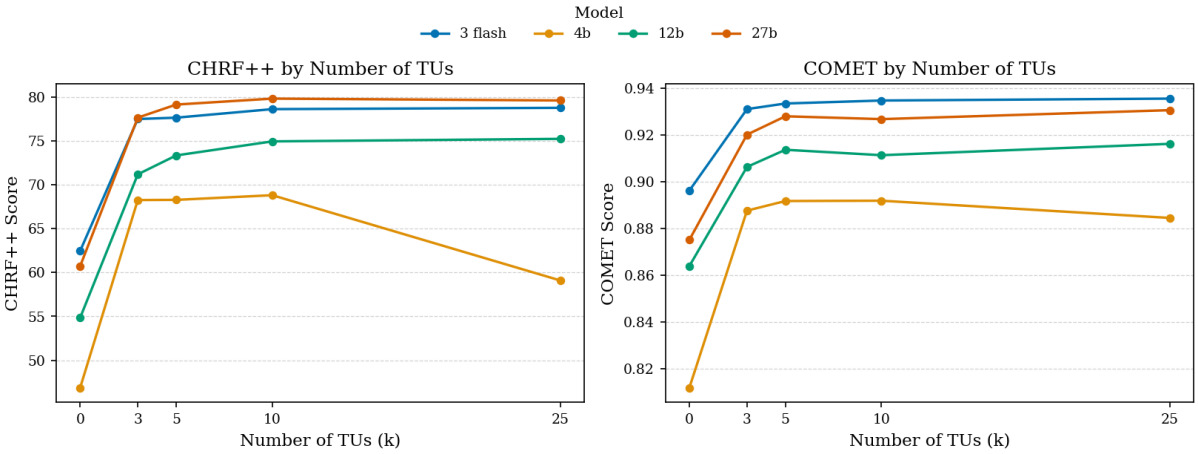


Figure 1: Average ChrF++ and COMET scores by model, by k.

Limitations

The principal limitations of this study mostly concern the data used.

First, at fewer than 3,000 sentence pairs pre-deduplication and 2,581 pairs post-deduplication, the set is rather small. On the other hand, it represents a cohesive legal document translated in the course of a real-world translation project. This translation has not been published previously and is thus guaranteed to be excluded from the training data of the LLMs assessed, as opposed to public bilingual corpora. With this said, **we do not claim or expect the observed improvements to scale to translation of unrelated documents from a different domain**; the improved MT quality can nevertheless be noticeable and helpful to translators working on large legal, technical, or other projects with consistent terminology and stylistic requirements.

Second, the human translations originated from three human translators, with an overwhelming contribution from one of them. This implies that we are evaluating how close an LLM can replicate the style and choice of terminology of an individual translator, not necessarily how universally good the translation is.

Finally, the dataset pertains to the legal/military domain where the terms and definitions are clear and unambiguous. We would not expect the demonstrated gains to scale to domains like literature, art, or colloquial speech. Studying the effects of RAG-MT across domains could be a subject of further investigation.

We also note two further ablations left for future work. First, we have not separated the contribution of semantic relevance of retrieved examples from the lift of merely having additional in-context examples; a controlled comparison against a random-pair retrieval baseline drawn uniformly from the training pool would isolate this effect and is a natural next experiment. Second, comparison against a generic-domain test set would help characterize the extent to which the observed gains are domain-specific rather than a general property of retrieval-augmented translation.

Ethical Considerations

The private testing dataset has been compiled during a volunteer translation project undertaken at the Theory and Practice of Translation from English department of the Institute of Philology, National Taras Shevchenko University of Kyiv. Permission to use translated texts has been obtained from the department and the translators involved in the project.

The source text constitutes a US legal act which is a part of US Code and is published by multiple open sources, thus no permission to use it was necessary.

Use of AI. The author used a frontier AI assistant for preliminary review feedback on drafts, verification of numerical consistency between tables, and language editing. All research design, experimental work, data analysis, and scientific conclusions are entirely the author's own. The author takes full responsibility for the content of this paper.

References

Owen Bianchi, Mathew J. Koretsky, Maya Willey, Chelsea X. Alvarado, Tanay Nayak, Adi Asija, Nicole Kuznetsov, Mike A. Nalls, Faraz Faghri, and Daniel Khashabi. 2025. Hidden in the Haystack: Smaller Needles are More Difficult for LLMs to Find. *arXiv preprint*. ArXiv:2505.18148 [cs].

Maxime Bouthors, Josep Crego, and François Yvon. 2024. Retrieving Examples from Memory for Retrieval Augmented Neural Machine Translation: A Systematic Comparison. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3022–3039, Mexico City, Mexico. Association for Computational Linguistics.

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, and 12 others. 2020. Language Models are Few-Shot Learners. *arXiv preprint*. ArXiv:2005.14165 [cs].

Bram Bulte and Arda Tezcan. 2019. Neural Fuzzy Repair: Integrating Fuzzy Matches into Neural Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1800–1809, Florence, Italy. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Sundesh Donthi, Maximilian Spencer, Om Patel, Joon Doh, Eid Rodan, Kevin Zhu, and Sean O'Brien. 2024. Improving LLM Abilities in Idiomatic Translation. *arXiv preprint*. ArXiv:2407.03518 [cs].

Maxim Enis and Mark Hopkins. 2024. From LLM to NMT: Advancing Low-Resource Machine Translation with Claude. *arXiv preprint*. ArXiv:2404.13813.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical Power and Translationese in Machine Translation Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

K. Hong, A. Troynikov, and J. Huber. 2025. Context rot: How increasing input tokens impacts LLM performance. Chroma Research.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Nearest Neighbor Machine Translation. *arXiv preprint*. ArXiv:2010.00710 [cs].

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Lujun Li, Lama Sleem, Niccolo' Gentile, Geoffrey Nichil, and Radu State. 2025. Exploring the Impact of Temperature on Large Language Models: Hot or Cold? *arXiv preprint*. ArXiv:2506.07295 [cs].

S. Li, J. Chen, S. Yuan, X. Wu, H. Yang, S. Tao, and Y. Xiao. 2023. Translate meanings, not just words: IdiomKB's role in optimizing idiomatic translation with language models. ArXiv:2308.13961.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Anton Shpigunov. 2025. Using automated quality metrics to improve machine translation into Ukrainian. *Humanities Science Current Issues*, 2:282–290.

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

Yanfang Ye, Zheyuan Zhang, Tianyi Ma, Zehong Wang, Yiyang Li, Shifu Hou, Weixiang Sun, Kaiwen Shi, Yijun Ma, Wei Song, Ahmed Abbasi, Ying Cheng, Jane Cleland-Huang, Steven Corcelli, Robert Goulding, Ming Hu, Ting Hua, John Lalor, Fang Liu, and 10 others. 2025. LLMs4All: A Review of Large Language Models Across Academic Disciplines. *arXiv preprint*. ArXiv:2509.19580 [cs].

Jingyi Zhang, Masao Utiyama, Eiichro Sumita, Graham Neubig, and Satoshi Nakamura. 2018. Guiding Neural Machine Translation with Retrieved Translation Pieces. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1325–1335, New Orleans, Louisiana. Association for Computational Linguistics.

Mike Zhang and Antonio Toral. 2019. The Effect of Translationese in Machine Translation Test Sets. In

Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers), pages 73–81, Florence, Italy. Association for Computational Linguistics.

A Statistical Significance Results

Model	Candidate	Δ ChrF++	p-val (ChrF)	Δ COMET	p-val (COMET)
gemini-3-flash	$k = 3$	15.03	< 0.01	0.035	< 0.01
gemini-3-flash	$k = 5$	15.18	< 0.01	0.037	< 0.01
gemini-3-flash	$k = 10$	16.15	< 0.01	0.039	< 0.01
gemini-3-flash	$k = 25$	16.30	< 0.01	0.039	< 0.01
gemma3-27b	$k = 3$	16.92	< 0.01	0.045	< 0.01
gemma3-27b	$k = 5$	18.41	< 0.01	0.053	< 0.01
gemma3-27b	$k = 10$	19.09	< 0.01	0.052	< 0.01
gemma3-27b	$k = 25$	18.88	< 0.01	0.056	< 0.01
gemma3-12b	$k = 3$	16.34	< 0.01	0.042	< 0.01
gemma3-12b	$k = 5$	18.49	< 0.01	0.050	< 0.01
gemma3-12b	$k = 10$	20.09	< 0.01	0.047	< 0.01
gemma3-12b	$k = 25$	20.38	< 0.01	0.052	< 0.01
gemma3-4b	$k = 3$	21.38	< 0.01	0.076	< 0.01
gemma3-4b	$k = 5$	21.41	< 0.01	0.080	< 0.01
gemma3-4b	$k = 10$	21.94	< 0.01	0.080	< 0.01
gemma3-4b	$k = 25$	12.22	< 0.01	0.073	< 0.01

Table 3: Results of statistical significance testing (paired bootstrap, Holm-corrected vs. $k = 0$).

B Score Distributions

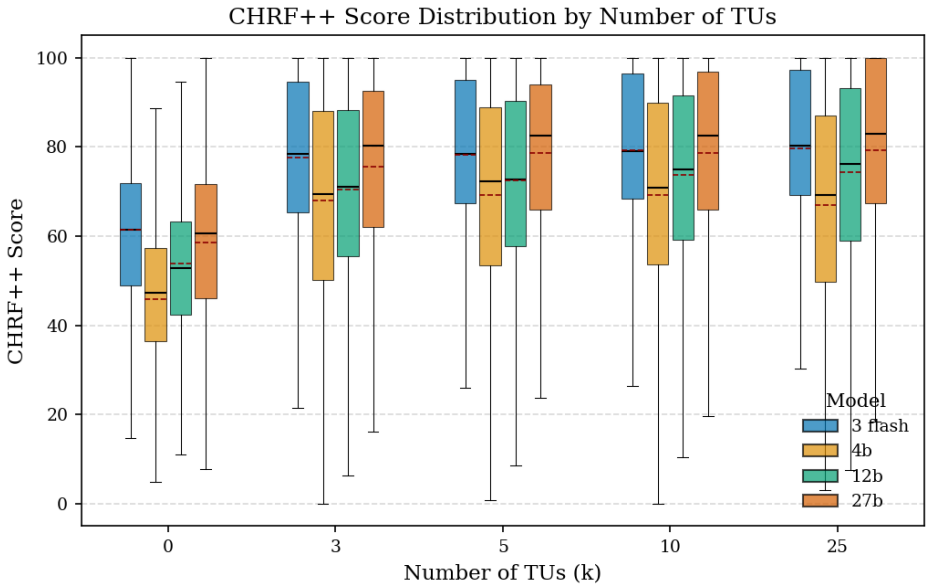


Figure 2: Distribution of ChrF++ scores across context sizes.

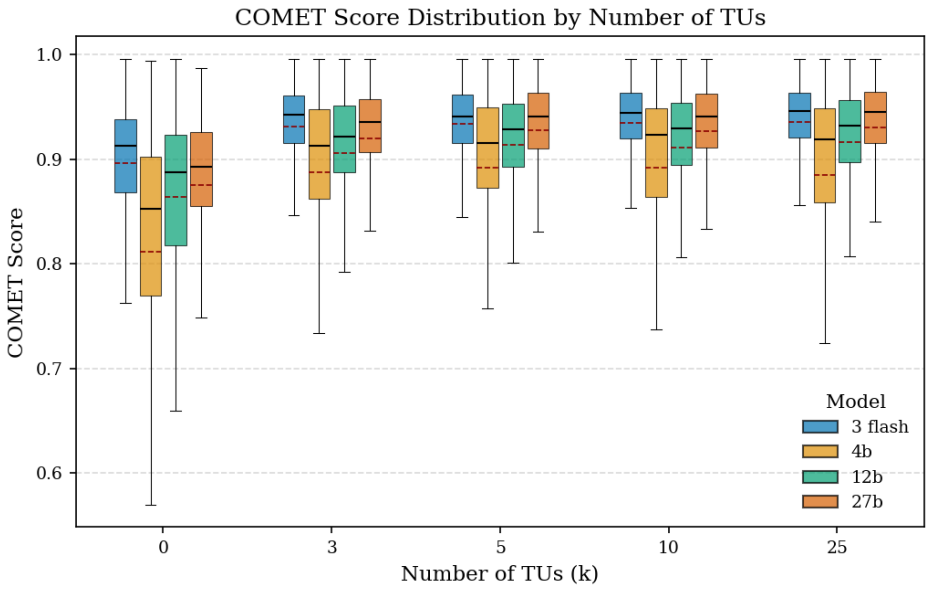


Figure 3: Distribution of COMET scores across context sizes.

C Translation Prompt

```
<Role>You are an expert language translator.</
Role>

<Instructions>

<Instruction>Please translate a text from {{
source_lang }} to {{ target_lang }} with utmost
accuracy and fluency.</Instruction>

<Instruction>

The source text may contain "tags" or "
placeables", such as `<tg123>...</tg123>` or `<
tg456>...</tg456>`. These must be rendered
verbatim, without any changes or alterations. If
 there are any tag attributes, they must not be
translated. If there is an opening tag and no
closing tags, omit the opening tag.

</Instruction>

<Instruction>Ensure that the translation
maintains original formatting, punctuation, and
special characters exactly as in the source text,
 including whitespace and especially soft-return
/soft-break characters.</Instruction>

</Instructions>

<References>

<SimilarTranslations>

<ReferenceDescription>These phrases and
sentences are semantically similar to the text
to be translated. Please refer to them for
vocabulary, grammar, style, and context.</
ReferenceDescription>

{% for tu in tus %}

<TranslationUnit>

<SourceSegment>{{ tu.source_segment }}</
SourceSegment>

<TargetSegment>{{ tu.ref_segment }}</
TargetSegment>

</TranslationUnit>

{% endfor %}

</SimilarTranslations>

</References>

<SourceText>{{ source_text }}</SourceText>

<FinalInstruction>Return ONLY the translation,
no other commentary or additional text:</
FinalInstruction>
```

UAReviews: A Multi-Task Ukrainian Dataset for Emotion and Intent Classification

Roman Kyslyi **Ihor Pysmennyi** **Denys Mykhailov**
Kyiv School of Economics, Ukraine
(rkyslyi,ipysmennyi,dmykhailov)@kse.org.ua

Abstract

We introduce UAReviews, a multi-task Ukrainian-language dataset for emotion and intent classification comprising 11,580 annotated texts. The dataset combines two sources: citizen reviews of government digital services provided by the Ministry of Digital Transformation of Ukraine and Ukrainian-language Telegram posts drawn from the COSMUS corpus. Each text is annotated with both an emotion label following the Ekman taxonomy (seven classes) and an intent label (five classes), making it the first publicly available Ukrainian resource for joint emotion and intent analysis. Annotation was performed by students at the Kyiv School of Economics, with a gold standard subset (20%) validated by three independent annotators achieving Krippendorff's $\alpha = 0.93$. We establish baselines using single-task and multi-task fine-tuned XLM-RoBERTa models and analyze emotion to intent correlation. Both the dataset and the baseline models are publicly available.[1]

1 Introduction

Emotion and intent detection in text are fundamental tasks in natural language processing (NLP) with wide-ranging applications, from customer feedback analysis to conversational AI systems. While substantial resources exist for high-resource languages such as English (Demszky et al., 2020; Mohammad et al., 2018), many languages still lack task-specific benchmarks. Ukrainian is no longer considered a low-resource language: thanks to growing community efforts, pre-trained models and general-purpose corpora are now available (Romanyshyn, 2023). However, the Ukrainian benchmark landscape for affective NLP is still sparse: the only existing emotion benchmark, EmoBench-UA (Dementieva et al., 2025), is multi-label and limited to social-media-style text, and to our knowledge no publicly available intent classification benchmark exists for Ukrainian. UAReviews complements EmoBench-UA by providing single-label emotion annotations together with intent labels on a different domain (government service reviews and Telegram posts).

The need for Ukrainian NLP resources has become particularly pressing in recent years. Government digitalization efforts in Ukraine, led by the Ministry of Digital Transformation, have generated large volumes of citizen feedback in Ukrainian. Understanding the emotions and intents expressed in this feedback is crucial for improving public services, yet no dedicated datasets or models existed for this purpose.

In this paper, we present UAReviews, a Ukrainian-language dataset annotated for both emotion and intent classification. Our dataset combines two complementary data sources: (1) citizen reviews of government digital services (e.g., the Diia platform) provided by the Ministry of Digital Transformation of Ukraine, and (2) Ukrainian-language posts from the COSMUS corpus (Shynkarov et al., 2025), a collection of Telegram channel messages. The inclusion of the COSMUS data is intended to add domain diversity and broaden linguistic register beyond formal government service reviews; we do not claim it balances the domain distribution, as the Telegram subset is much smaller than the government review subset (Section 3).

Our contributions are as follows:

- We release UAReviews, a publicly available multi-task Ukrainian dataset of 11,580 texts annotated for both emotion and intent classification, combining government service reviews and social media posts, with a standard train/dev/test split.
- We describe a rigorous annotation protocol

[1]Dataset and models: https://huggingface.co/KSE-RESEARCH-Group

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 12–23

using Argilla (Vila-Suero and Aranda, 2023) tool hosted on Hugging Face Spaces, with quality control through a gold standard subset achieving Krippendorff's $\alpha = 0.93$ (Krippendorff, 2011).

- We provide comprehensive baselines, including single-task and multi-task models, with per-class evaluation metrics, confusion matrices, and multi-seed variance analysis.

- We quantify the correlation between emotion and intent labels using Cramér's V and normalized mutual information, providing empirical support for joint modeling.

2 Related Work

Emotion Detection Datasets. Early work in emotion detection introduced datasets based on news headlines (Strapparava and Mihalcea, 2007) and social media (Mohammad et al., 2018). Demszky et al. (2020) released GoEmotions, a large-scale English dataset with 27 emotion categories derived from Reddit comments. These resources have driven progress in emotion detection for English, but comparable resources for most other languages remain scarce.

Emotion taxonomies in NLP typically follow either Ekman (1992), who proposed six basic emotions (anger, disgust, fear, happiness, sadness, surprise), or Plutchik (1980), who extended this to eight primary emotions organized in a wheel. Our annotation scheme follows the Ekman taxonomy, adopting the six basic emotions supplemented with a *Neutral* category.

For Ukrainian specifically, EmoBench-UA (Dementieva et al., 2025) is the only other emotion detection benchmark. It was created using crowdsourcing on Toloka and formulated as a multi-label task, with the best-performing model (DeepSeek-V3) reaching a macro F1 of 0.65. UAReviews differs in two key aspects: (i) it targets single-label classification, reflecting the dominant emotion per text, and (ii) it additionally provides intent annotations, enabling joint emotion-intent research.

Intent Classification. Intent classification has been studied primarily in the context of task-oriented dialogue systems (Casanueva et al., 2020). While intent detection datasets exist for English and a few other languages, to our knowledge no publicly available intent classification dataset exists for Ukrainian.

Ukrainian NLP Resources. Ukrainian NLP has seen growing attention in recent years (Romanyshyn, 2023), and the language can no longer be considered low-resource in terms of pre-trained models. Prior work has addressed sentiment analysis, with Shynkarov et al. (2025) introducing the COSMUS corpus—a 12,224-text collection from Telegram channels and product-review sites—for Ukrainian social media sentiment analysis in code-switching contexts. EmoBench-UA (Dementieva et al., 2025) provides multi-label emotion annotations for Ukrainian. UAReviews differs in its focus on single-label classification, the addition of intent annotations, and a different domain (government service reviews), making the two datasets complementary rather than redundant.

3 Dataset

3.1 Data Sources

The UAReviews dataset is constructed from two complementary sources:

Government Service Reviews. The primary data source consists of citizen reviews of government digital services, provided by the Ministry of Digital Transformation of Ukraine. These reviews cover a range of public services delivered through digital platforms (e.g., Diia) and include feedback on service quality, usability, and citizen satisfaction. The texts are exclusively in Ukrainian and represent semi-formal register. Each review is accompanied by a numerical rating.

COSMUS Telegram Posts. To increase domain diversity and balance the dataset, we incorporate Ukrainian-language posts from the COSMUS dataset (Shynkarov et al., 2025). COSMUS is a corpus of Telegram channel messages originally compiled for sentiment analysis research, containing texts in Ukrainian, Russian, and code-switched varieties. We selected only the Ukrainian-language subset of COSMUS. These texts represent a more informal, social-media register and cover a broader range of topics than the government service or institutions reviews.

3.2 Annotation Scheme

Each text in UAReviews is annotated along two dimensions:

Emotion Labels. We adopt an emotion taxonomy grounded in the established psychological framework of Ekman (1992). Each text is assigned one of seven emotion labels: *Happiness*, *Anger*, *Sadness*, *Fear*, *Surprise*, *Disgust*, and *Neutral* (no dominant emotion). The label set was chosen to balance granularity with practical annotator reliability. The *Neutral* category captures texts that express factual content or mixed emotions without a clearly dominant affective signal.

Intent Labels. Each text is additionally annotated with an intent label (stored as `final_category` in the dataset) capturing the communicative purpose of the text. Intent categories were designed to reflect the pragmatic functions common in citizen feedback and social media discourse. The five intent categories are: *Gratitude / Positive Feedback*, *Complaint / Dissatisfaction*, *Question / Request for Help*, *Neutral Comment*, and *Suggestion / Idea*.

3.3 Annotation Process

Annotation was carried out by students at the Kyiv School of Economics using Argilla (Vila-Suero and Aranda, 2023), an open-source data annotation platform hosted on Hugging Face Spaces.

We report inter-annotator agreement using Krippendorff's α (Krippendorff, 2011), a chance-corrected reliability measure suitable for multiple annotators and nominal data (Artstein and Poesio, 2008). Because the dataset has two annotation dimensions, we compute α separately for emotion and for intent and report the average of the two; the values are tightly clustered around the reported figure for both dimensions.

Gold Standard Set. Approximately 20% of the dataset (∼2,316 texts) was designated as the gold standard set. Each text in this subset was independently annotated by three human annotators. The inter-annotator agreement on the gold set was Krippendorff's $\alpha = 0.93$, indicating near-perfect agreement and high annotation quality.

Main Annotation. The remaining 80% of the dataset (∼9,264 texts) was annotated by two human annotators with assistance from Gemini (Gemini Team, 2023), a large language model used as a third annotator. The inter-annotator agreement on this portion, including the LLM-assisted annotations, was $\alpha = 0.87$. Final labels were determined by majority vote across the three annotations (two human + one LLM). Crucially, the gold standard set (20%, fully human-annotated with $\alpha = 0.93$) serves as a quality control benchmark: the small gap between $\alpha = 0.93$ (gold) and $\alpha = 0.87$ (LLM-assisted) provides evidence that the LLM-assisted portion does not substantially degrade annotation quality. The gold set also enables future work to directly compare model performance on fully human-annotated vs. LLM-assisted subsets, should per-annotator labels be released.

Statistic	Value
Total texts	11,580
Gold standard subset (20%)	∼2,316
Main subset (80%)	∼9,264
Language	Ukrainian
Annotation dimensions	emotion, intent
Annotators (gold set)	3 humans
Annotators (main set)	2 humans + LLM
Krippendorff's α (gold set)	0.93
Krippendorff's α (main set)	0.87

Table 1: Overview of the UAReviews dataset.

Discussion of Agreement Levels. We acknowledge that $\alpha = 0.93$ is unusually high for subjective emotion annotation. We attribute this to several factors: (1) the review domain often produces texts with clear emotional signals (e.g., explicit gratitude or complaints), (2) the seven-class taxonomy based on Ekman's well-established framework is relatively coarse-grained, and (3) the annotator training and clear guidelines (Appendix D) promoted consistent interpretations. The dominant classes (*Happiness* at 65.3%, *Gratitude / Positive Feedback* at 64.2%) are typically unambiguous, which likely contributes to the high overall agreement.

3.4 Dataset Statistics

Table 1 summarizes the key statistics of the UAReviews dataset.

Table 2 shows the emotion label distribution. The dataset exhibits a pronounced class imbalance: *Happiness* dominates at 65.3%, and *Gratitude / Positive Feedback* at 64.2%. We emphasize that this distribution is not an artifact of sampling bias - it reflects the real-world distribution of citizen feedback on government digital services. Satisfied users naturally leave positive reviews more

Emotion	Count	%
Happiness	7,557	65.3
Anger	2,264	19.5
Neutral	1,117	9.6
Sadness	424	3.7
Disgust	106	0.9
Surprise	57	0.5
Fear	55	0.5
Total	**11,580**	**100.0**

Table 2: Emotion label distribution in UAReviews, sorted by frequency.

Intent	Count	%
Gratitude / Positive Feedback	7,440	64.2
Complaint / Dissatisfaction	2,730	23.6
Question / Request for Help	615	5.3
Neutral Comment	418	3.6
Suggestion / Idea	377	3.3
Total	**11,580**	**100.0**

Table 3: Intent label distribution in UAReviews, sorted by frequency.

frequently, a well-documented phenomenon in review and customer feedback datasets (Hu and Liu, 2004; McAuley and Leskovec, 2013). Similar positive skews appear in app store reviews (where 5-star ratings routinely exceed 60%), customer satisfaction surveys, and product feedback platforms. Artificially rebalancing the dataset (e.g., by oversampling negative reviews or discarding positive ones) would misrepresent the true distribution that deployed systems must handle.

The inclusion of COSMUS Telegram posts was motivated by the need to diversify the emotional and intent distribution: the COSMUS subset contains 67.1% *Neutral* and 18.6% *Sadness* (compared to 9.3% and 3.6% respectively in government reviews), providing a complementary signal at the per-class level. We caution, however, that with only 170 COSMUS texts versus 11,410 government reviews, this addition adds register and topical diversity rather than meaningfully balancing the overall domain distribution. Expanding the COSMUS component is a priority for future dataset releases and would also enable proper cross-domain evaluation between the two sources. The tail classes - *Sadness* (3.7%), *Disgust* (0.9%), *Surprise* (0.5%), and *Fear* (0.5%) - are substantially underrepresented, posing challenges for classification. To mitigate this, we employ class-weighted cross-entropy loss (Section 4) and report both macro and weighted F1 to make the impact of imbalance transparent.

Figure 1 visualizes the class distributions for both annotation dimensions.

Each record in the released dataset contains the following fields: a unique identifier (`id`), a numerical `rating` (where applicable), the text `content`, a `source` indicator (government reviews or COSMUS), the `final_emotion` label, the `final_category` (intent) label, the text `length`, and the `split` assignment (train, dev, or test).

3.5 Emotion-Intent Correlation

To quantify the relationship between emotion and intent labels, we compute Cramér's V and normalized mutual information (NMI) on the full dataset. The cross-tabulation (Figure 2, Appendix E) reveals strong alignment: *Happiness* maps almost exclusively to *Gratitude / Positive Feedback*, while *Anger* aligns with *Complaint / Dissatisfaction*. The overall association is strong (Cramér's $V = \mathbf{0.63}$, bias-corrected $V = 0.63$; NMI $= 0.70$; $\chi^2 = 18{,}603.84$, $p < 0.001$), confirming that the two annotation dimensions are statistically dependent but not redundant - tail emotion classes such as *Fear*, *Surprise*, and *Disgust* distribute across multiple intent categories.

4 Experiments

4.1 Data Splits

We partition the dataset into *train* (70%; 8,106 texts), *dev* (15%; 1,737 texts), and *test* (15%; 1,737 texts) using stratified sampling based on the joint emotion-intent label. All hyperparameter and model selection decisions were made on the dev set; the held-out test set was used only for the final evaluations reported in this section. We report mean and standard deviation across five random seeds on this same held-out test set; no model selection was performed on test data. The split preserves class proportions across all three subsets (see Appendix B for per-class breakdowns).

4.2 Model

For all experiments, we use *XLM-RoBERTa-base-uk*, a Ukrainian-adapted variant of XLM-

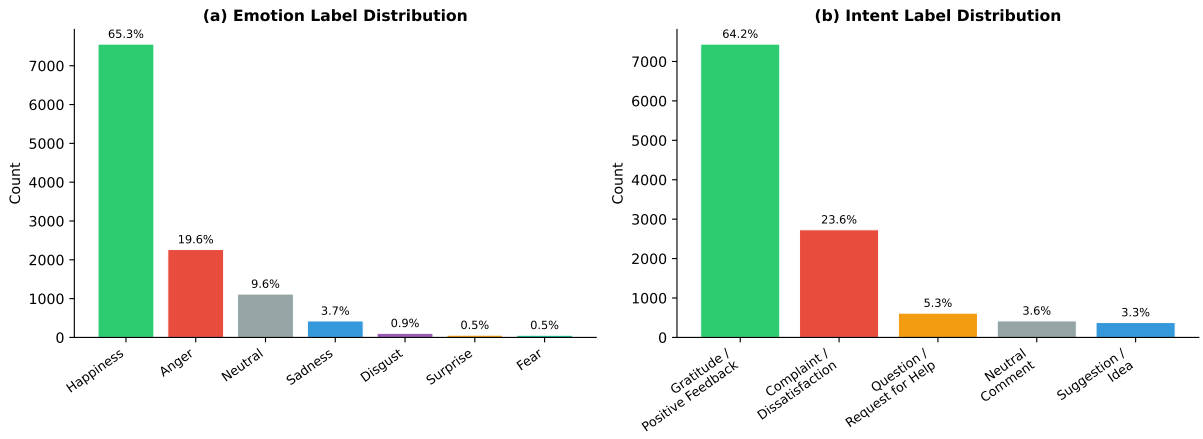


Figure 1: Class distribution for (a) emotion and (b) intent labels in UAReviews. Both dimensions exhibit a long-tail distribution dominated by positive classes, which is characteristic of review and feedback data.

RoBERTa (Conneau et al., 2020). Unlike the original multilingual model (which contains 470M parameters with embeddings for 100 languages), this checkpoint retains only Ukrainian and English vocabulary embeddings, reducing the model to 110M parameters while preserving the encoder's representational capacity for Ukrainian.

4.3 Single-Task Baselines

We train separate models for emotion classification (7 classes) and intent classification (5 classes). A standard `XLMRobertaForSequenceClassification` head is added on top of the pre-trained encoder. Both models share the same hyperparameter configuration (Table 4), fine-tuned using PyTorch Lightning with the Hugging Face Transformers library (Wolf et al., 2020).

We employ separate learning rates for the classification head (5e-5) and the pre-trained embeddings (9e-5), and class-weighted cross-entropy loss to mitigate class imbalance. Weights are computed as inverse frequency raised to a power of 0.2, then normalized to sum to 1.

4.4 Multi-Task Baseline

To test whether joint modeling improves over single-task training, we implement a shared-encoder multi-task model: a single XLM-RoBERTa encoder feeds into two separate classification heads (one for emotion, one for intent), each consisting of a dense layer with tanh activation followed by a linear projection. The total loss is an equally weighted sum of the two task losses ($\lambda_{\text{emo}} = \lambda_{\text{int}} = 0.5$). We report the average macro F1 across both tasks as the primary metric.

Hyperparameter	Value
Pre-trained model	xlm-roberta-base-uk
Classifier LR	5e-5
Embedding LR	9e-5
Optimizer	AdamW
Weight decay	0.005
Effective batch size	128
Max epochs	10
Early stopping patience	5 (on dev F1)
Scheduler	Cosine with warmup
Warmup ratio	0.1
Precision	16-bit mixed
Gradient clipping	1.0 (norm)
Class weight power	0.2

Table 4: Shared hyperparameters for both tasks.

4.5 Results

Table 5 presents the main results. We report macro F1 (mean $\pm$ standard deviation across 5 random seeds) on the held-out test set; per-seed numbers are provided in Appendix A. The single-task models achieve strong performance, with intent classification (macro F1 = 0.93 $\pm$ 0.05) outperforming emotion classification (0.81 $\pm$ 0.15). The high variance on emotion reflects the difficulty of rare-class prediction: across the five seeds, emotion macro F1 ranges from 0.51 (worst seed) to 0.92 (best seed), with the worst seed driven down by tail-class collapse. The per-class results in Tables 6 and 7 are reported for seed 42, which produced near-mean performance (macro F1 $\approx$ 0.82

Model	Emo $F1_M$	Int $F1_M$
Single-task	**0.81 ± 0.15**	**0.93 ± 0.05**
Multi-task (shared)	0.55 ± –	0.81 ± –

Table 5: Test-set macro F1 (mean ± std over 5 seeds) for emotion and intent classification. $F1_M$ denotes macro-averaged F1 (Opitz and Burst, 2021).

Emotion	P	R	F1
Happiness	98.1	99.0	98.6
Anger	96.6	94.9	95.8
Neutral	88.7	91.1	89.9
Sadness	83.1	81.2	82.1
Disgust	72.7	76.2	74.4
Surprise	100.0	63.6	77.8
Fear	100.0	36.4	53.3
Macro avg	91.3	77.5	81.7
Weighted avg	96.1	96.1	96.0

Table 6: Per-class precision (P), recall (R), and F1 for **emotion** classification on the test set (single-task model, seed 42, used here as a representative near-mean run).

on emotion and 0.81 on intent); we use this seed as a representative single-run breakdown rather than as the worst- or best-case run. The multi-task model underperforms single-task baselines, particularly for emotion (0.55 vs. 0.81 macro F1), suggesting that the shared encoder may not effectively balance both objectives with equal loss weighting. This is an area for future work.

Tables 6 and 7 present per-class precision, recall, and F1 for both tasks. As expected, the high-frequency classes (*Happiness*, *Gratitude / Positive Feedback*) achieve the strongest F1 scores, while the tail classes (*Fear*, *Surprise*, *Disgust*) show lower performance. Fear achieves the lowest F1 (53.3) with perfect precision but only 36.4% recall, indicating the model identifies Fear when it predicts it but misses most instances. Surprise similarly suffers from low recall (63.6%). The confusion matrix confirms that tail-class errors are not simply majority-class predictions: Fear is most often confused with Sadness and Neutral, while Disgust errors distribute across Anger and Fear.

Figures 3 and 4 (Appendix E) show the confusion matrices. The emotion confusion matrix reveals that errors are concentrated among tail classes, which tend to be confused with the dominant *Happiness* class. The intent confusion matrix shows a cleaner diagonal, consistent with the higher macro F1 for this task.

Intent	P	R	F1
Gratitude / Pos. Fb.	96.3	97.6	96.9
Complaint / Dissat.	95.0	89.6	92.2
Question / Help	86.7	90.2	88.4
Neutral Comment	57.0	63.1	59.9
Suggestion / Idea	70.8	68.0	69.4
Macro avg	81.2	81.7	81.4
Weighted avg	93.2	93.1	93.1

Table 7: Per-class precision (P), recall (R), and F1 for **intent** classification on the test set (single-task model, seed 42, used here as a representative near-mean run). Note that the macro F1 of 81.4 on this individual seed is consistent with the 5-seed mean of 0.93 reported in Table 5: most seeds reach ∼0.96 macro F1, with seed 42 being a lower-end run.

5 Analysis

Minority Class Performance. The per-class results (Tables 6 and 7) allow direct assessment of whether the model's aggregate F1 is driven primarily by head classes. While the gap between macro and weighted F1 quantifies this effect, the tail-class F1 scores are the critical metric. Despite extreme imbalance (e.g., *Fear*: 55 examples, 0.5%), the class-weighted loss with power 0.2 provides meaningful correction.

Class Weight Ablation. We chose a class weight power of 0.2 based on preliminary experiments. A power of 0 (uniform weighting) leads to majority-class bias; powers above 0.5 can overfit to rare classes. We acknowledge that we do not present a systematic ablation over this hyperparameter; future work could compare with focal loss (Lin et al., 2017) or resampling strategies.

Multi-Task Performance Collapse. The multi-task model exhibits a severe performance drop on emotion classification (macro F1: 0.81 → 0.55), while intent performance remains comparable to the single-task baseline (0.81 vs. 0.93). This asymmetric degradation - where one task collapses while the other is preserved - needs careful analysis. We identify several plausible contributing factors:

(1) Task difficulty asymmetry and head dominance. The intent task (5 classes, less imbalanced,

macro F1 = 0.93 single-task) is substantially easier than the emotion task (7 classes, extreme imbalance, macro F1 = 0.81). With equal loss weighting ($\lambda = 0.5$), the shared encoder may converge toward representations that favor the easier intent task, as intent gradients dominate early training and shape the shared representation space before the harder emotion task can establish useful features. This is a well-known failure mode in multi-task learning (Chen et al., 2018).

(2) Strong label correlation as a double-edged sword. The high correlation between emotion and intent (Cramér's $V = 0.63$, NMI $= 0.70$) means that intent labels are partially predictive of emotion. Rather than providing complementary signal, the shared encoder may learn to rely on intent-discriminative features that are sufficient for the coarse emotion–intent mapping (e.g., *Happiness* $\leftrightarrow$ *Gratitude*) but insufficient for distinguishing tail emotion classes that span multiple intents (e.g., *Fear* appears in both *Complaint* and *Question*).

(3) Representation conflict on tail classes. The emotion tail classes (*Fear*, *Surprise*, *Disgust*: collectively 1.9% of data) require fine-grained distinctions that the shared encoder may sacrifice to improve performance on the majority classes shared across both tasks. The per-class results confirm this: multi-task emotion F1 for *Disgust* drops to 22.2 (from 74.4 single-task) and *Fear* to 30.0 (from 53.3), while *Happiness* remains high at 95.9.

(4) Loss imbalance. Although both task losses use class weights, the equal $\lambda = 0.5$ weighting does not account for the difference in task difficulty or convergence rate. The intent head converges faster and continues to dominate gradient updates throughout training.

We did not ablate over λ values, encoder freezing schedules, or hierarchical architectures (e.g., predicting intent first and conditioning emotion on intent) in this work. These directions represent important areas for future work.

We present the multi-task result as a baseline, and we note that its performance itself is informative: it demonstrates that naive shared-encoder multi-task learning is insufficient for this dataset and that the strong emotion-intent correlation does not guarantee multi-task gains.

Annotation Quality: Per-Class Considerations. While the overall $\alpha = 0.93$ on the gold set indicates high agreement, this aggregate metric may mask lower agreement on inherently subjective categories. Emotions like *Neutral* (which serves as a residual category) and *Sadness* (which can overlap with *Anger* in complaint contexts) are likely harder to annotate consistently. Future releases of UAReviews will include per-class α values. We also note that the high proportion of unambiguous positive reviews (65.3% *Happiness*) contributes to the high aggregate agreement.

6 Discussion

UAReviews addresses a significant gap in Ukrainian NLP resources by providing the first publicly available dataset for joint emotion and intent classification. Several design decisions merit further discussion.

Hybrid Annotation Strategy. Our use of an LLM (Gemini) as a third annotator alongside two human annotators for 80% of the dataset is a design choice to balance annotation cost with quality. The key safeguard is the gold standard set: 20% of the dataset was annotated entirely by three human annotators ($\alpha = 0.93$), providing an LLM-free reference point. The marginal drop to $\alpha = 0.87$ on the LLM-assisted portion suggests that Gemini's annotations are largely consistent with human judgments in this domain. Importantly, because final labels are determined by majority vote (two humans + one LLM), Gemini can only influence the label when the two human annotators disagree - it cannot override a human consensus. This architecture means the LLM acts as a tiebreaker rather than a primary annotator. We acknowledge that without per-annotator labels, we cannot quantify potential LLM biases (e.g., toward majority classes) and plan to release per-annotator labels in a future version to enable such analysis.

Ukrainian-Adapted Pre-training. Our choice of `ukr-models/xlm-roberta-base-uk` - a vocabulary-trimmed variant of XLM-RoBERTa that retains only Ukrainian and English embeddings - over the full multilingual checkpoint reflects a practical trade-off: the trimmed model is significantly smaller (110M vs. 470M parameters) while preserving the encoder's capacity for Ukrainian.

Joint Modeling Potential. The strong statistical dependence between emotion and intent (Section 3.5) motivates multi-task learning, yet our

shared-encoder baseline demonstrates that naive approaches fail (Section 5). The performance collapse on emotion suggests that effective joint modeling requires architectural innovations: task-specific loss weighting ($\lambda_{\text{emo}} > \lambda_{\text{int}}$) to compensate for difficulty asymmetry, gradient balancing methods (Chen et al., 2018; Kendall et al., 2018), hierarchical prediction (e.g., predicting intent first and conditioning emotion on intent), or cross-attention mechanisms between task heads. The dataset's dual annotations make it a valuable testbed for such research.

7 Conclusion

We have presented UAReviews, a multi-task Ukrainian-language dataset of 11,580 texts annotated for emotion classification (7 classes) and intent classification (5 classes). The dataset combines citizen reviews of government services with Ukrainian Telegram posts from the COSMUS corpus, annotated with a rigorous quality control protocol achieving Krippendorff's $\alpha = 0.93$ on the gold standard subset. We provide comprehensive baselines - single-task and multi-task, with per-class metrics, confusion matrices, and multi-seed variance analysis, including a detailed analysis of multi-task performance collapse. Both the dataset and the models are publicly released to support Ukrainian NLP research.

UAReviews fills a gap in Ukrainian NLP by providing the first jointly annotated emotion–intent dataset, enabling research at the intersection of affective computing and pragmatic analysis for an underserved language. Our experiments reveal two key findings. First, single-task fine-tuning of a vocabulary-trimmed XLM-RoBERTa achieves strong intent classification (macro F1 = 0.93) but exposes the difficulty of tail-class emotion detection under extreme imbalance, where seed sensitivity alone can swing macro F1 by 0.4. Second, the failure of naive shared-encoder multi-task learning—despite strong statistical dependence between labels—demonstrates that label correlation is a necessary but insufficient condition for multi-task gains.

Limitations

The UAReviews dataset has several limitations. First, the dataset exhibits significant class imbalance, which affects tail-class performance. Although class-weighted loss partially addresses this, we do not present ablations comparing it with alternatives such as focal loss or oversampling. Second, the dataset is drawn from only two domains (government service reviews and Telegram posts), with the Telegram subset (170 texts) being too small for reliable cross-domain evaluation. Expanding this component would improve the dataset's utility for domain adaptation research. Third, the high α values may partly reflect the dominance of unambiguous positive reviews rather than genuine ease of annotation across all categories. Fourth, we report results from a single model architecture, exploring larger models, ensembles, or additional pre-training may be beneficial.

Ethics Statement

The government service reviews used in this dataset were provided by the Ministry of Digital Transformation of Ukraine for research purposes. All texts were anonymized prior to annotation to remove personally identifiable information. The Telegram posts were sourced from the publicly available COSMUS dataset (Shynkarov et al., 2025). Annotators were students at the Kyiv School of Economics who participated voluntarily under a formal arrangement with the institution; the annotation activity was not unpaid labor. The annotation process was designed to avoid exposing annotators to harmful content. We release the dataset under a CC-BY-4.0 licence.

Acknowledgements

We thank the Ministry of Digital Transformation of Ukraine for providing access to the citizen review data, and the student annotators at the Kyiv School of Economics for their careful annotation work. We also thank the anonymous reviewers of UNLP 2026 for their constructive feedback, which helped improve this paper.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. 2018. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning*, pages 794–803.

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

Daryna Dementieva, Nikolay Babakov, and Alexander Fraser. 2025. EmoBench-UA: A benchmark dataset for emotion detection in Ukrainian. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2025–2048.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7482–7491.

Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability. *Departmental Papers (ASC)*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988.

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 165–172.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17.

Jürgen Opitz and Sebastian Burst. 2021. Macro F1 and macro F1. *arXiv preprint arXiv:1911.03347*.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman, editors, *Theories of Emotion*, pages 3–33. Academic Press.

Mariana Romanyshyn. 2023. Proceedings of the second Ukrainian natural language processing workshop (UNLP). In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*.

Yurii Shynkarov, Veronika Solopova, and Vera Schmitt. 2025. Improving sentiment analysis for Ukrainian social media code-switching data. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 179–193.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74.

Daniel Vila-Suero and Francisco Aranda. 2023. Argilla: Open-source framework for data-centric NLP.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moe, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacin Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

A Training Details

All models were trained using PyTorch Lightning with class-weighted cross-entropy loss. Class weights were computed from inverse frequency weighting (power = 0.2, normalized). Tables 8 and 9 list the weights for each task.

Emotion Class	Weight
Happiness	0.076
Anger	0.096
Neutral	0.111
Sadness	0.135
Disgust	0.178
Surprise	0.201
Fear	0.203

Table 8: Class weights for emotion classification.

Intent Class	Weight
Gratitude / Positive Feedback	0.134
Complaint / Dissatisfaction	0.164
Question / Request for Help	0.221
Neutral Comment	0.238
Suggestion / Idea	0.243

Table 9: Class weights for intent classification.

The training split preserves the original class distribution via stratified sampling. All models were trained for a maximum of 10 epochs with early stopping (patience = 5) based on dev macro F1 (Opitz and Burst, 2021). Results are reported on the held-out test set. Multi-seed experiments use seeds $\{42, 123, 456, 789, 2024\}$.

B Data Splits

The dataset is split into train (70%), dev (15%), and test (15%) via stratified sampling on the joint emotion–intent label. Table 10 shows the sizes.

	Train	Dev	Test
Texts	8,106	1,737	1,737
%	70.0	15.0	15.0

Table 10: Dataset split sizes.

C Data Fields

Each record in the UAReviews dataset contains the following fields:

- `id`: Unique identifier for the text.
- `rating`: Numerical rating (applicable to government service reviews).
- `content`: The raw text of the review or post.
- `source`: Indicator of the data source (government reviews or COSMUS).
- `final_emotion`: The consensus emotion label (one of seven classes).
- `final_category`: The consensus intent/category label.
- `length`: The character length of the text.
- `split`: Train/dev/test split assignment.

D Annotation Guidelines

The following guidelines were provided to annotators via the Argilla platform. Annotators were instructed to read each text carefully and assign exactly one emotion label and one intent label. When a text conveyed multiple emotions or intents, annotators were asked to select the dominant one - the emotion or intent that most strongly characterized the text as a whole. In cases of genuine ambiguity, annotators were encouraged to use the *Neutral* emotion label or the *Neutral Comment* intent label.

D.1 Emotion Label Definitions

The emotion taxonomy follows Ekman (1992), with the addition of a *Neutral* category. Definitions and representative examples are provided below in the original Ukrainian with English translations.

Happiness. The text expresses positive feelings such as joy, satisfaction, gratitude, delight, or contentment. The author conveys a favorable experience or outcome.

Example: « Все працює ідеально, дякую! Отримав документ за 5 хвилин без жодної черги. » ("Everything works perfectly, thank you! Got my document in 5 minutes without any queue.")

Anger. The text expresses irritation, frustration, outrage, or hostility. The author is displeased and may direct negativity at a service, person, or situation.

Example: « Жахливе місце. Жахлива черга. Ніхто нічого підказати не може. » ("Terrible place. Terrible queue. Nobody can help with anything.")

Sadness. The text expresses grief, disappointment, sorrow, or melancholy. The tone is somber or dejected rather than angry.

Example: « На жаль, я не зміг отримати потрібну послугу. Прикро, що нічого не змінилося. » ("Unfortunately, I couldn't get the service I needed. It's disappointing that things haven't improved.")

Fear. The text expresses worry, anxiety, concern, or apprehension about a potential negative outcome or threat.

Example: « Хвилююся, що мої персональні дані можуть бути в небезпеці » ("I'm worried

that my personal data might be at risk")

Surprise. The text expresses astonishment or unexpectedness, either positive or negative. The author did not anticipate the experience described.

Example: «Зовсім не очікував цього - процес був неймовірно швидким, я був у шоці!» ("I didn't expect this at all - the process was incredibly fast, I was shocked!")

Disgust. The text expresses revulsion, contempt, or strong aversion. The author finds the situation deeply unpleasant or morally objectionable.

Example: «Умови в офісі огидні - брудно, переповнено, а персонал грубий.» ("The conditions in the office were revolting - dirty, overcrowded, and the staff were rude.")

Neutral. The text does not express a clearly dominant emotion. It may be purely factual, informational, or contain mixed emotions that do not resolve to a single category.

Example: «Офіс працює з 9:00 до 18:00 у будні. Потрібно мати при собі паспорт та ідентифікаційний код.» ("The office is open from 9:00 to 18:00 on weekdays. You need to bring your passport and ID code.")

D.2 Intent Label Definitions

The intent labels capture the communicative purpose of the text. Definitions and representative examples are provided below.

Gratitude / Positive Feedback. The author's primary purpose is to express thanks, appreciation, or provide a positive evaluation of a service, product, or experience.

Example: «Дуже дякую! Працівники були дуже привітні і все зробили швидко.» ("Thank you so much! The staff were very helpful and everything was done quickly.")

Complaint / Dissatisfaction. The author's primary purpose is to express dissatisfaction, report a problem, or criticize a service, product, or experience.

Example: «Чекаю вже три години і досі не обслужили. Це неприпустимо.» ("I've been waiting for three hours and still haven't been served. This is unacceptable.")

Question / Request for Help. The author's primary purpose is to seek information, ask a question, or request assistance with a specific issue.

Example: «Підкажіть, які документи потрібно взяти для заміни посвідчення особи?» ("Can someone tell me which documents I need to bring for the ID card replacement?")

Neutral Comment. The author's primary purpose is to share factual information, make an observation, or provide a comment without a clear positive, negative, or help-seeking intent.

Example: «Нове відділення відкрилося минулого тижня на Хрещатику.» ("The new branch opened last week on Khreshchatyk Street.")

Suggestion / Idea. The author's primary purpose is to propose an improvement, offer constructive feedback, or share an idea for how a service or process could be enhanced.

Example: «Було б чудово, якби додали можливість записуватися онлайн, щоб уникнути черг.» ("It would be great if you added an option to schedule appointments online to avoid the queues.")

D.3 Special Instructions

Annotators were provided with the following additional guidance:

- **Sarcasm and irony:** If a text uses sarcasm (e.g., «Чудовий сервіс, лише чотири години почекав!» — "Great service, only had to wait four hours!"), annotate based on the *intended* meaning, not the literal surface form. In this case, the emotion would be *Anger* and the intent *Complaint / Dissatisfaction*.
- **Mixed signals:** If a text contains both positive and negative elements (e.g., «Працівники привітні, але чекати занадто довго.» — "The staff were friendly but the wait was too long"), select the emotion and intent that dominate the overall tone.
- **Short texts:** For very short texts (e.g., single words or emoji-only), annotate based on the available signal. If no clear emotion can be inferred, use *Neutral*.
- **Code-switching:** Some texts from the COSMUS subset may contain Russian words or phrases. Annotate based on the overall meaning regardless of the language used within the text.

E Additional Figures

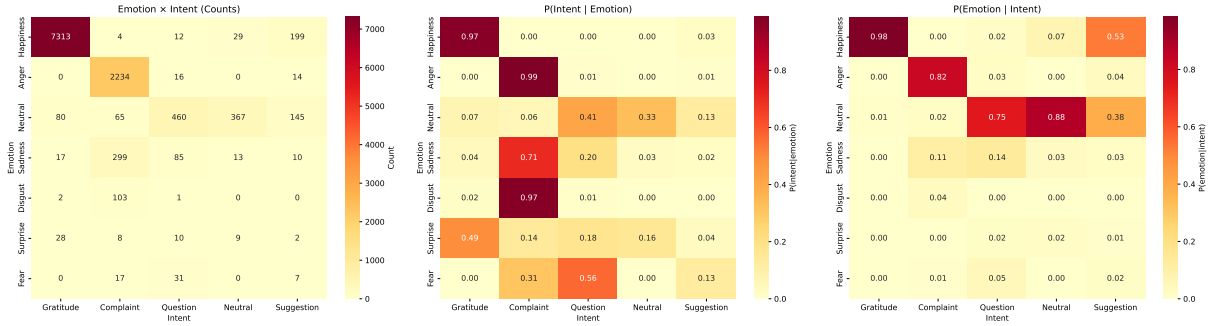


Figure 2: Emotion–intent cross-tabulation heatmap showing raw counts (left), P(intent|emotion) (center), and P(emotion|intent) (right).

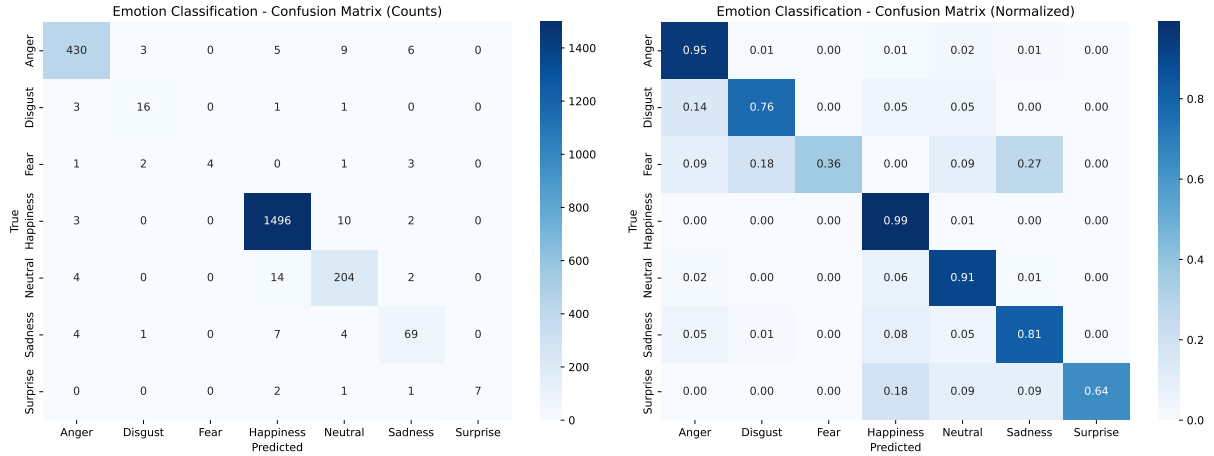


Figure 3: Confusion matrix for emotion classification (normalized by true class). The model achieves strong performance on head classes but struggles to distinguish tail classes from *Happiness* and *Neutral*.

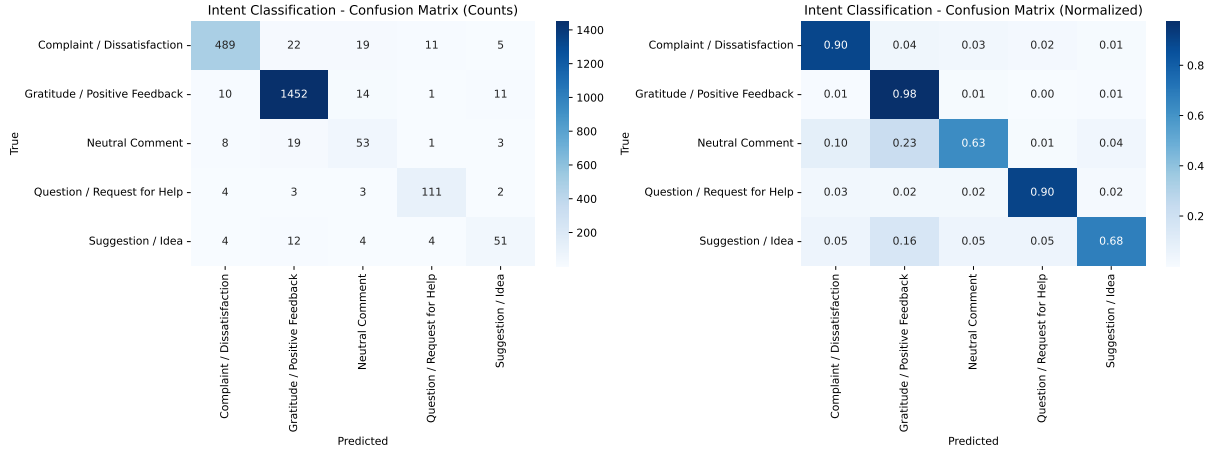


Figure 4: Confusion matrix for intent classification (normalized by true class).

A Two-Axis Framework for Analyzing Ukrainian Dialogues

Artem Korotenko
Kyiv School of Economics
akorotenko@kse.org.ua

Roman Kyslyi
Kyiv School of Economics
rkyslyi@kse.org.ua

Abstract

Online discussions increasingly serve as a major venue for exchanging information and evaluating competing viewpoints. Yet most computational approaches to discourse quality focus on detecting harmful language or predicting engagement, providing limited insight into whether interactions actually improve collective understanding.

We introduce an initial two-dimensional framework for modeling dialogic constructiveness, distinguishing between substantive contribution (SC) and relational conduct (RC). Using expert-annotated Ukrainian-language discussions, we find preliminary evidence that collapsing rubric-level labels into these axes improves inter-annotator agreement, which is consistent with constructiveness being captured more reliably as a multidimensional judgment.

We further compare nominal, regression, and ordinal prediction approaches and find that explicitly modeling constructiveness as an ordinal task yields higher agreement with expert annotations under quadratic weighted kappa (QWK) on our gold test set. These results are consistent with dialogic constructiveness being more effectively modeled as an ordered interactional judgment than as a binary label or continuous score.

1 Introduction

Online discussions increasingly shape how people understand political, social, and scientific issues. Comment threads on news sites, forums, and messaging platforms often serve as a primary space for exchanging information and evaluating competing viewpoints. However, while some discussions help participants clarify arguments or reconsider positions, others produce confusion, repetition, or conflict without contributing to shared understanding.

Current computational tools can detect harmful language or predict engagement, but they provide limited insight into whether a discussion is actually productive. In practice, conversations that generate new insight and those that merely generate noise may appear equally civil or equally active. This makes it difficult to assess whether online interaction supports collective reasoning or simply amplifies disagreement.

The concept of *constructiveness* attempts to capture this positive dimension of discourse quality. In ordinary language, *constructive* means "serving a useful purpose" or "helping to improve something rather than damage it" (Oxford University Press, 2026). In interactional terms, a constructive comment is one that helps move a discussion forward instead of obstructing it. Kolhatkar and Taboada (2017) define constructive comments as follows:

> "Constructive comments intend to create a civil dialogue through remarks that are relevant to the article and not intended to merely provoke an emotional response. They are typically targeted to specific points and supported by appropriate evidence."

However, in most computational work, constructiveness is treated as a single binary label (Napoles et al., 2017; Kolhatkar et al., 2020; Nguyen et al., 2021). A comment is classified as either constructive or non-constructive. Empirically, such labels show only moderate inter-annotator agreement, suggesting that annotators compress multiple dimensions of discourse quality into one scalar judgment. Conceptually, this formulation conflates interpersonal conduct with argumentative substance.

This binary approach leaves a vast *"grey zone"* of interaction unaddressed: the cognitively demanding but relationally rough disagreements where participants may express strong opposing views without resorting to *ad hominem* attacks. To bridge this gap, we propose a shift from binary classification to a dual-axis architecture.

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 24–32

We define constructiveness as a multidimensional property composed of two independent components:

1. Relational Conduct (RC) — the interpersonal quality of interaction, including respect, engagement, and conflict handling (Gibb, 1961; Gottman, 1994)

2. Substantive Contribution (SC) — the epistemic value of the message, including reasoning quality, justification, informativeness, and coherence within the dialogue (Habermas, 1984; Grice, 1975; Steenbergen et al., 2003).

These dimensions are orthogonal. A contribution may be respectful yet add little substance, or it may present rigorous reasoning while escalating interpersonal tension. Collapsing both into a single label obscures this structure and limits analytical precision.

For this reason, we move from binary classification to a two-axis framework that disentangles Relational Conduct from Substantive Contribution, allowing a more granular analysis of conversational health.

We demonstrate the utility of this framework by applying it to a novel dataset of Ukrainian political dialogues sourced from **Telegram** and **Ukrayinska Pravda Forum**. By utilizing a 10-message context window, we capture pragmatic signals—such as sarcasm and relational repair—that are frequently lost in single-comment analysis.

We present this work as an exploratory study intended to motivate further investigation. Our contributions are as follows:

1. We introduce a two-axis theoretically grounded framework for conversational health with six specific subdimensions.

2. We show that collapsing rubric-level labels into these axes improves inter-annotator agreement, suggesting more stable latent dimensions of discourse quality.

3. We demonstrate that modeling constructiveness as an ordinal prediction task yields higher agreement with expert judgments than nominal or continuous formulations.

2 Related Work

The automated analysis of online discourse has produced several foundational resources, yet existing frameworks remains largely fragmented across single-label or outcome-based metrics.

1. **The Yahoo News Annotated Comments Corpus** (YNACC) attempted to identify "constructive" comments but found the category too broad for consistent human judgment, resulting in a Krippendorff's alpha of only 0.48–0.63 (Napoles et al., 2017)

2. **The Constructive Comments Corpus** (C3) (Kolhatkar et al., 2020) labeled comments in isolation, a method that often fails to detect pragmatic signals like sarcasm or "repair attempts" that only emerge across multiple conversational turns;

3. **The ChangeMyView** (CMV) dataset measures constructiveness through the lens of persuasion (Tan et al., 2016), focusing on the outcome (successful change of mind) rather than the deliberative process itself.

4. **The Stanford Politeness Corpus** utilizes computational markers (e.g., hedges, gratitude) to score etiquette (Danescu-Niculescu-Mizil et al., 2013), which can lead to the "Kind but Empty" trap where civil but substantively vacuous content is amplified

5. Recent work by Zhou et al. (2024) utilizes Large Language Models (LLMs) to extract dataset-independent linguistic features—such as dispute tactics and collaboration markers—to predict constructiveness outcomes . While this framework improves model interpretability, it highlights a persistent issue: the omission of a formal, *a priori* definition of constructiveness . By defining "constructiveness" purely through downstream proxies—such as the avoidance of mediation (escalation) or self-reported open-mindedness scores—these models risk learning robust prediction rules for specific datasets while failing to generalize a universal standard for conversational health. Without a top-down theoretical architecture, bottom-up feature extraction remains bound to the artifacts of the target variable it seeks to predict

3 Proposed Framework

The guiding hypothesis of this framework is that constructiveness may not be adequately captured

by a single score, because it appears to conflate two conceptually distinct dimensions. **Relational Conduct (RC)** measures the interpersonal quality of treatment among participants, rooted in Gottman's and Gibb's theories. **Substantive Contribution (SC)** measures the rational-deliberative quality of the content, grounded in the Deliberative Quality Index and deliberative systems theory.

3.1 Scoring Mechanics: The 5-Point Gradient

Every subdimension is evaluated on a centered ordinal scale from -2 to $+2$. A five-point scale was selected as a compromise between expressiveness and annotation reliability. Simpler three-level schemes cannot capture meaningful variation in conversational quality, while more granular scales introduce annotation noise due to small perceptual differences between adjacent categories. The five-point structure provides a clear neutral anchor (0) for factual but shallow contributions while allowing two degrees of positive and negative constructiveness that better match human perception of conversational intensity.

Score	Description
+2	Exemplary: perspective-taking, structured reasoning, proactive de-escalation
+1	Constructive: polite engagement and clear reasoning
0	Neutral: factual and professional but shallow
-1	Poor: dismissiveness, passive-aggressiveness, or topical drift
-2	Anti-constructive: insults, mockery, or escalation

The initial rubric contained eight subdimensions—four relational and four substantive—designed to capture a broad range of dialogic behaviors. In practice, however, annotation revealed substantial overlap between closely related criteria. Real-world comments often exhibited signals that were difficult to attribute to a single dimension, and annotators struggled to consistently separate concepts such as procedural fairness from reasoning quality or responsiveness from discussion progress.

Following several calibration rounds, we simplified the rubric by merging conceptually adjacent dimensions. The resulting framework preserves the two-axis structure but reduces each axis to three core subdimensions, yielding a six-dimensional scheme. This revision maintains theoretical coverage while improving annotation clarity and inter-annotator consistency.

3.2 Relational Conduct (RC): The Interpersonal Layer

Relational Conduct refers to how participants treat one another within an exchange. It captures whether interlocutors recognize each other as legitimate participants in dialogue and whether disagreement is handled in a way that sustains, rather than damages, the interaction.

This dimension is grounded in established communication research. Gibb's theory of defensive versus supportive communication distinguishes interactional climates that encourage openness from those that trigger defensiveness (Gibb, 1961). Gottman's work on conflict behavior identifies contempt and hostility as reliable indicators of relational breakdown (Gottman, 1994). Politeness theory further conceptualizes interaction in terms of managing face threats and acknowledging the social standing of the interlocutor (Brown and Levinson, 1987). Across these traditions, the central concern is not agreement but the preservation of interactional conditions necessary for continued dialogue.

Relational Conduct therefore evaluates whether a message maintains the possibility of discussion. It does not measure emotional neutrality or consensus. Strong disagreement may still exhibit high relational quality if expressed without personal attack or delegitimization. Conversely, even factually correct statements may score low if delivered in a way that undermines the interaction itself.

In our framework, Relational Conduct is decomposed into three components: tone and respect (RC1), engagement with the interlocutor (RC2), and conflict management (RC3). Together, these capture whether the relational layer of the conversation remains intact.

3.3 Substantive Contribution (SC): The Deliberative Layer

The Substantive axis assesses the intellectual capital and "logical payload" added to the conversation.

The framework is rooted in Jürgen Habermas's concept of an "ideal speech situation," where communication is undistorted by power or manipulation and participants are swayed only by the "unforced force of the better argument" (Habermas, 1984)

These Habermasian ethics is operationalized through the lens of the Discourse Quality Index (DQI) (Steenbergen et al., 2003), an empirical methodology originally designed to measure the

Axis	Subdimension	Definition
Relational Conduct (RC)	RC1. Respect / Tone	Focuses on emotional attitude toward others. Measures politeness, hostility, and warmth of language, without judging reasoning or topic relevance.
	RC2. Engagement	Captures social awareness and participation in dialogue. Evaluates whether the speaker actively engages with others and shows interest in continuing the exchange.
	RC3. Conflict Management	Reflects how disagreement is handled. Assesses whether the message escalates tension, stays neutral, or seeks understanding and compromise.
Substantive Contribution (SC)	SC1. Reasoning Quality	Assesses logical structure and fairness of argumentation. Focuses on how claims are supported by reasoning and whether evidence or opposing views are treated honestly.
	SC2. Informativeness	Measures content value. Evaluates how much new, relevant, or specific information the comment adds, separate from tone or reasoning.
	SC3. Dialog Continuity	Captures how well a comment fits into the ongoing dialogue. Evaluates whether it logically follows from what was said before and contributes to the development of the discussion instead of breaking or diverting its flow.

Table 1: The Two-Axis Framework Scoring Rubric

quality of parliamentary deliberation. The DQI focuses heavily on the "level of justification," providing a hierarchy that distinguishes between claims made with no support, inferior justification involving tenuous or fallacious links, and sophisticated justification where multiple reasons are examined in depth.

While the DQI offers a robust foundation, its parliamentary origin assumes a structured environment with a predefined behavioral "floor". In the "wild" digital sphere, this scale lacks the granularity to detect passive-aggressive condescension or the "repair attempts" essential for de-escalation. Moreover, the DQI's three-to-four level justification scale is too coarse to distinguish "polite" vacuity from "rough" but substantively deep contributions.

To resolve these limitations, the Substantive Contribution (SC) axis decomposes value into three granular sub-dimensions. SC1 (Reasoning Quality and Integrity) shifts focus from subjective agreement to objective deliberative rigor and epistemic honesty. SC2 (Informativeness) operationalizes Gricean maxims (Grice, 1975) by rewarding original insights that transform the thread while penalizing "empty" content. Finally, SC3 (Dialog Continuity) maintains systemic health by evaluating semantic coherence and penalizing irrelevant non-sequiturs that stall the deliberative flow (Mansbridge and Parkinson, 2012).

4 Dataset Collection and Annotation

4.1 Sourcing: Telegram and Ukrayinska Pravda Forum

We curated a novel dataset of Ukrainian political discourse [1] by scraping two distinct environments that represent different eras and modes of online interaction:

- **Telegram (TG):** We targeted ShrikeChat[2], one of the most prominent Ukrainian political discussion hubs.
- **Ukrayinska Pravda (UP) Forum[3]:** A legacy platform with over 20 years of history, capturing the long-term evolution of the Ukrainian digital public sphere.

Unlike previous datasets that group comments into loose threads, our methodology preserves the **direct-response chain**.

1. **Queue Extraction:** Linear sequences where each message is an explicit reply to the preceding one.
2. **9-Message Context Window:** For every target message, the nine preceding direct responses were preserved. Recent findings in

[1] https://huggingface.co/datasets/KSE-RESEARCH-Group/ukrainian-dialogs-constructiveness

[2] https://t.me/shrikechat

[3] https://forum.pravda.com.ua/index.php?board=2.0

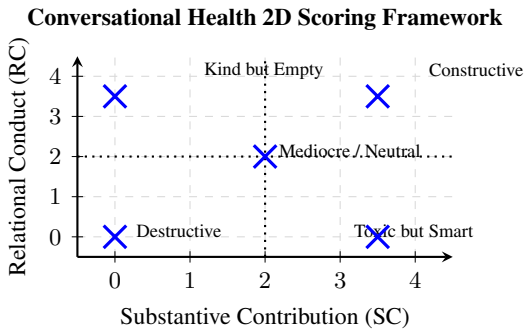


Figure 1: Two-axis conversational health scoring framework.

> NLP show the importance of context to grasp the nuances of conversation, specifically regarding how perceived toxicity can shift when the preceding dialogue is considered (Xenos et al., 2022)

This approach enables distinction between unprovoked aggression and retaliatory low-RC scores, as well as identification of substantive repair.

4.2 Labeling and Annotation

On the first stage, each dialogue was pre-labeled using the Gemini 2.5 Flash model (Google DeepMind, 2025). Manual annotation was subsequently performed using the Argilla framework (Vila-Suero and Aranda). Due to technical constraints, the theoretical -2 to $+2$ scale was mapped to a 0 to 4 integer scale.[4]

Before constructing the final evaluation sets, calibration rounds were conducted to align human perception with the framework. Items with directional disagreement were discussed until consensus or documented divergence was reached.

The Gold Set was sampled from the pre-labeled corpus using a stratified approach.

Sampling prioritized dimensional coverage, bucket balancing, and stratum diversity, including rare cases such as "Toxic but Smart" (low RC, high SC).

5 Data Analysis

To validate the reliability of our multi-dimensional annotation framework, we conducted an extensive Inter-Annotator Agreement (IAA) analysis. The dataset consists of 1,371 annotated items divided into two subsets. The gold subset was annotated independently by the same three calibrated expert annotators, providing a consistent expert reference set used for evaluation and reliability analysis. The main annotation set was labeled by two annotators per item, with a subset receiving a third annotation when additional verification was required. This design allows the main dataset to increase training coverage while maintaining a stable expert-validated subset for benchmarking and agreement analysis.

We assess reliability through three complementary perspectives: agreement at the level of individual rubric axes, agreement after collapsing axes into the two higher-level meta-dimensions (Relational Constructiveness and Substantive Constructiveness), and the empirical independence of the relational and substantive rubrics.

5.1 Inter-Annotator Agreement (IAA)

Table 2: Inter-Annotator Agreement (IAA) for the calibrated expert annotators ($N = 300$). Meta-dimensions collapse granular axes into a 2D framework, showing significantly higher conceptual alignment.

Dimension	QWK	α	Ex.%	Cl.%
RC1: Respect/Tone	0.650	0.647	68.0	96.8
RC2: Engagement	0.471	0.445	49.8	90.8
RC3: Conflict Mgmt	0.555	0.562	64.6	95.7
SC1: Reasoning	0.611	0.604	56.8	92.9
SC2: Informativeness	0.682	0.674	54.2	93.4
SC3: Dialog Continuity	0.521	0.512	58.8	93.0
Content (Y)	**0.761**	**0.762**	–	–
Relationship (X)	**0.666**	**0.663**	–	–

We report four agreement metrics per dimension. **Quadratic weighted Cohen's** κ (QWK) and **Krippendorff's** α are chance-corrected coefficients that penalize disagreements by squared distance on the ordinal scale; both range from -1 to 1, with higher values indicating stronger chance-corrected agreement. **Ex.%** denotes exact agreement and **Cl.%** close agreement (within ± 1).

5.2 2D Quality Framework and Orthogonality

One observation in our analysis is the transition from 6 granular axes to a 2D Quality Map. Collapsing the axes into meta-dimensions—Relationship ($X = \sum RC_{1-3}$) and Content ($Y = \sum SC_{1-3}$)—yields higher signal-to-noise ratios. The Content Meta score reached $\alpha = 0.801$, a level of agreement consistent with—though not by itself establishing—a more stable two-dimensional latent structure. The average Euclidean distance between expert points in

[4] A score of 2 represents the Neutral (0) anchor.

this 12×12 space is 1.66 units, indicating close agreement.

Furthermore, Pearson correlation analysis supports the orthogonality of these dimensions. As illustrated in the correlation matrix (Table 3), while internal clusters show moderate correlation ($r \approx 0.5$), cross-cluster correlation is significantly lower. Specifically, the correlation between *Respect/Tone* and *Informativeness* is weak ($r = 0.14$), justifying the treatment of model "politeness" and "accuracy" as independent variables.

Table 3: Pearson Correlation Matrix between annotation axes ($N = 300$). Shading indicates correlation strength: Darker blue denotes $r \geq 0.5$; lighter shades denote $r < 0.3$.

	RC1	RC2	RC3	SC1	SC2	SC3
RC1	1.00	0.58	0.68	0.23	0.16	0.24
RC2	0.58	1.00	0.60	0.29	0.14	0.38
RC3	0.68	0.60	1.00	0.28	0.24	0.29
SC1	0.23	0.29	0.28	1.00	0.51	0.49
SC2	0.16	0.14	0.24	0.51	1.00	0.60
SC3	0.24	0.38	0.29	0.49	0.60	1.00

If constructiveness were a single scalar construct, relational and substantive indicators would collapse into one highly correlated cluster. Instead, we observe two moderately coherent but separable clusters, supporting the interpretation of constructiveness as a two-dimensional state.

6 Baseline Modeling

To examine whether the proposed constructiveness framework can be predicted automatically, we evaluate several baselines under different assumptions about label structure. Each subdimension (RC1–RC3, SC1–SC3) is modeled independently on a five-point ordinal scale (0–4) derived from the original -2 to $+2$ rubric. Performance is measured with Quadratic Weighted Kappa (QWK), which accounts for ordinal disagreement by penalizing larger errors heavily than near misses.

We use a fixed split protocol. The gold set (N=300) is divided into train and test (210/90). The gold training portion is merged with the main dataset, from which 15% is used for validation and 85% for training. Final evaluation is conducted exclusively on the gold test set (N=90).

All neural baselines share the same multilingual encoder, XLM-RoBERTa (Conneau et al., 2020), fine-tuned for constructiveness prediction. The nominal, regression, and ordinal models therefore differ only in output formulation and decoding strategy, making their comparison directly controlled.

Because QWK is sensitive to prediction variance, trivial strategies can achieve moderate proximity to the gold labels while still yielding near-zero agreement. We therefore calibrate all non-trivial baselines on the gold validation split before final evaluation.

The first baseline is a distribution-aware non-text model that samples labels from the empirical score distribution of the gold training subset for each axis. This captures dataset priors without using textual information.

We then train a nominal softmax classifier with cross-entropy loss, treating each axis as a five-class classification problem and ignoring the ordered structure of the scale. At inference time, class probabilities are converted to ordinal predictions using expectation decoding followed by rounding.

Next, we consider a regression baseline trained with mean squared error (MSE) to predict continuous constructiveness scores. Per-axis decision thresholds are tuned on the validation split to maximize QWK before testing. This allows the model to partially adapt to ordinal structure, although continuous predictions remain sensitive to boundary effects.

Finally, we evaluate an ordinal classifier based on CORAL (COnsistent RAnk Logits) (Niu et al., 2016). CORAL decomposes prediction into a sequence of threshold exceedance decisions and enforces monotonic consistency across class boundaries. As with regression, thresholds are calibrated on the validation split. By modeling ordered decision boundaries directly, CORAL is better suited to the ranked nature of the constructiveness scale.

6.1 Baseline Results

Table 4 summarizes the predictive performance of the evaluated baseline approaches on the held-out test split. In addition to Quadratic Weighted Kappa (QWK), we report the proportion of predictions within one ordinal step of the expert annotation (Near$\pm$1), which reflects coarse agreement with annotator judgments under the five-point constructiveness scale.

The distribution-aware baseline, which samples predictions from the empirical training distribution without using textual information, achieves relatively high Near$\pm$1 agreement despite near-zero QWK. This reflects the strong concentration of annotations within a narrow central range of the

Table 4: Baseline performance across six constructiveness dimensions. All models share an identical XLM-RoBERTa encoder and differ only in output formulation and decoding strategy.

Model	Macro QWK	Near±1	RC1	RC2	RC3	SC1	SC2	SC3
Distribution (no text)	−0.016	81.11%	0.019	0.018	-0.072	0.127	0.013	-0.116
Softmax (CE)	0.006	92.22%	0.008	0.000	0.030	0.000	0.000	0.000
Regression (MSE)	0.074	76.83%	0.122	0.059	-0.006	0.052	0.123	0.091
Ordinal (CORAL)	**0.3891**	**86.67%**	0.309	0.435	0.241	0.452	0.576	0.328

ordinal scale, where coarse proximity can be obtained without meaningful alignment with expert judgments.

The nominal softmax classifier substantially improves numerical proximity to the gold annotations, as reflected by lower MAE and higher Near±1 agreement. However, this formulation yields only marginal gains in QWK, indicating that while the model captures overall constructiveness polarity, it fails to reliably place predictions across the ordinal boundaries used by annotators.

The regression baseline partially recovers latent ordering information, resulting in modest improvements in ordinal agreement. Nevertheless, without calibrated decision thresholds, continuous predictions frequently cross class boundaries in ways that degrade alignment with the discrete annotation scale.

In contrast, the ordinal CORAL formulation achieves higher QWK across all dimensions while maintaining similar Near±1 agreement. This is consistent with the improvement being driven by explicit modeling of ordered structure rather than by changes in text representation. The ordinal model appears to better approximate the decision boundaries reflected in expert annotations within the proposed two-axis framework

7 Conclusion

We introduced a two-dimensional framework for modeling dialogic constructiveness in online discussions, separating substantive contribution (SC) from relational conduct (RC). Unlike prior approaches that treat constructiveness as a binary property of individual messages, the proposed formulation captures both the informational and interactional contributions of a message to the progression of a discussion.

Empirical analysis of expert annotations suggests that constructiveness judgments behave as multidimensional and ordered in our data. Collapsing the six rubric dimensions into the two proposed axes increases inter-annotator agreement, which is consistent with—though not conclusive of—RC and SC capturing more stable aspects of expert judgment than the individual annotation criteria.

Our modeling experiments further show that approaches ignoring ordinal structure struggle to align with expert decisions under agreement-based metrics. Nominal classifiers achieve high proximity to the gold labels but produce weak ordinal agreement, while regression models only partially recover the ordering of constructiveness scores. In contrast, explicitly modeling constructiveness as an ordinal prediction task yields higher alignment with expert judgments on our gold test set.

Overall, the results support modeling dialogic constructiveness as a bounded ordinal judgment expressed along multiple interactional dimensions, though further validation is needed. Future work should explore architectures that better exploit ordinal structure and extend the framework to additional languages, platforms, and forms of online discussion.

Ethics

The dataset was constructed from publicly available Telegram channels via the official Telegram API and from publicly accessible discussion threads on the Ukrainska Pravda Forum. No private groups, direct messages, or restricted-access content were accessed. Data were processed for research purposes under legitimate interest and scientific research provisions. Identifiers were pseudonymized at the time of data collection, and only content necessary for the analytical objectives was retained. Messages containing information that could reasonably enable re-identification of individuals were removed from the dataset prior to analysis.

Annotation and labeling were conducted by student researchers who receive educational grants that include a requirement to contribute a defined number of working hours to university research activities. Their participation in the annotation

process formed part of these structured research duties. Annotators were granted access exclusively to anonymized data and were instructed on confidentiality and data protection requirements prior to participation.

Limitations

Several limitations of this study should be noted.

First, the framework and models were developed using Ukrainian-language online discussions, primarily from political and civic contexts. Norms of disagreement and cooperation may vary across languages and communities, meaning that constructiveness thresholds learned in this setting may not directly transfer to other discourse environments.

Second, the current approach assumes an explicit reply-based conversational structure, evaluating each message in relation to its immediate dialogic context. In less structured discussions involving indirect or topic-level responses, constructiveness may depend on broader interaction patterns not captured by pairwise message adjacency.

Third, inter-annotator agreement remains moderate, particularly along relational dimensions. This reflects the subjective nature of tone and intent interpretation, and suggests that some learned decision boundaries may be influenced by annotation uncertainty.

Fourth, items were pre-labeled with Gemini 2.5 Flash before human review. Despite calibration rounds and independent expert annotation, we cannot rule out anchoring effects that may inflate observed agreement, particularly along the collapsed two-axis dimensions. Comparison with fully blind annotation is needed to quantify this.

Finally, the expert-annotated gold subset is relatively limited in size (N = 300, with N = 90 held out for test). Triple expert annotation across six ordinal dimensions is labor-intensive, which limited the feasible scale within our resources. While the larger main split (N = 1,071) supports broader training coverage, statistical power for the QWK comparisons reported on the gold test set remains limited; differences between baselines should be interpreted accordingly.

References

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, and et al. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL*.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. *ACL Anthology*.

Jack R. Gibb. 1961. Defensive communication. *Journal of Communication*, 11(3):141–148.

Google DeepMind. 2025. Gemini 2.5 flash. Accessed: 2026-02-10.

John M. Gottman. 1994. *Why Marriages Succeed or Fail*. Simon & Schuster.

Herbert P. Grice. 1975. Logic and conversation. In *Speech Acts*, pages 41–58. Brill.

Jürgen Habermas. 1984. *The Theory of Communicative Action*. Beacon Press.

Varada Kolhatkar and Maite Taboada. 2017. Constructive language in news comments. In *Proceedings of the First Workshop on Abusive Language Online*, pages 11–17, Vancouver, BC, Canada. Association for Computational Linguistics.

Varada Kolhatkar, Nanyun Wu, Luca Cavasso, Emmanuel Francis, and Maite Taboada. 2020. Classifying constructive comments. *arXiv preprint arXiv:2004.05476*.

Jane Mansbridge and John Parkinson. 2012. *Deliberative Systems: Deliberative Democracy at the Large Scale*. Cambridge University Press.

Courtney Napoles, Joel Tetreault, Aasish Pappu, Enrica Rosendahl, and Madira Thampiratnam. 2017. Finding good conversations online: The yahoo news annotated comments corpus. *Proceedings of the 11th Linguistic Annotation Workshop*, pages 13–23.

Luan Thanh Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. *Constructive and Toxic Speech Detection for Open-Domain Social Media Comments in Vietnamese*, page 572–583. Springer International Publishing.

Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. 2016. Ordinal regression with multiple output cnn for age estimation. In *CVPR*, pages 4920–4928.

Oxford University Press. 2026. Constructive. https://www.oxfordlearnersdictionaries.com/definition/english/constructive. Accessed: 2026-02-10.

Marco Steenbergen, André Bächtiger, Markus Spörndli, and Jürg Steiner. 2003. Measuring political deliberation: A discourse quality index. *Comparative European Politics*, 1:21–48.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in Good-faith online discussions. *arXiv preprint arXiv:1602.01103*.

Daniel Vila-Suero and Francisco Aranda. Argilla – open-source framework for data-centric NLP.

Alexandros Xenos, John Pavlopoulos, Ion Androutsopoulos, Lucas Dixon, Jeffrey Sorensen, and Léo Laugier. 2022. Toxicity detection sensitive to conversational context. *First Monday*, 27(5).

Lexin Zhou, Youmna Farag, and Andreas Vlachos. 2024. An LLM feature-based framework for dialogue constructiveness assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5389–5409. Association for Computational Linguistics.

A Annotation Examples from the Gold Set

A.1 Toxic but Smart

Original

"відомо про співвідношення переданих/відпущених хамасівців? чи просто чергове узагальнення" (vidomo pro spivvidnoshennia peredanykh/vidpushchenykh hamasivtsiv? chy prosto cherhove uzahalnennia)

Translation

Is there known data on the ratio of Hamas prisoners exchanged or released, or is this just another generalization?

Scores RC1 = -1, RC2 = 0, RC3 = -1; SC1 = 1, SC2 = 1, SC3 = 1

A.2 Kind but Empty

Original

"Дякую, а то я вже почала писати Краще не скажеш." (Diakuiu, a to ya vzhe pochala pysaty. Krashche ne skazhesh.)

Translation

Thanks, I had already started writing it myself. Couldn't have said it better.

Scores RC1 = 2, RC2 = 2, RC3 = 0; SC1 = 0, SC2 = -1, SC3 = 0

A.3 Constructive

Original

"Я це розумію, але я не розумію, чому не можна навести порядок в генштабі." (Ya tse rozumiiu, ale ya ne rozumiiu, chomu ne mozhna navesty poriadok v henshtabi.)

Translation

I understand that, but I do not understand why it is not possible to bring order to the General Staff.

Scores RC1 = 1, RC2 = 1, RC3 = 1; SC1 = 1, SC2 = 1, SC3 = 1

A.4 De-escalation

Previous message

"Та це ж очевидно: нормальні громадяни не голосують за Поршенка."

Response

"хто такі ці нормальні громадяни і чому невибір Поршенка зробив по вашому 73 % українців — ненормальними, та ще й 'прихильники певної політсили'? може досить бавитись в гівнокидання? питання дійсно серйозні." (khto taki tsi normalni hromadiany i chomu nevybir Porshenka zrobyv po vashomu 73% ukraintsiv nenormalnymy? mozhe dosyt bavytys v hivnokydannia? pytannia diisno seriozni.)

Translation

Who exactly are these "normal citizens," and why does not voting for Poroshenko make 73% of Ukrainians "abnormal" in your view? Maybe it is time to stop the mud-slinging; these are serious questions.

Scores RC1 = -1, RC2 = 1, RC3 = 2; SC1 = 1, SC2 = 1, SC3 = 1

A.5 Critical but Substantive

Original

"Тобто статистика показує, що Юкі був абсолютно неправий." (Tobto statystyka pokazue, shcho Yuki buv absoliutno nepravyi.)

Translation

So the statistics show that Yuki was completely wrong.

Scores RC1 = 0, RC2 = 0, RC3 = 0; SC1 = 1, SC2 = -1, SC3 = 1

A.6 Destructive

Original

"И ты такая же." (I ty takaya zhe.)

Translation

And you are the same.

Scores RC1 = -2, RC2 = -2, RC3 = -2; SC1 = -1, SC2 = -2, SC3 = -2

A.7 Neutral Inquiry

Original

"Це було до чи після початку повномасштабної війни?" (Tse bulo do chy pislia pochatku povnomasshtabnoi viiny?)

Translation

Was this before or after the start of the full-scale war?

Scores RC1 = 0, RC2 = 0, RC3 = 0; SC1 = 0, SC2 = 0, SC3 = 0

Entropy of Ukrainian

Anton Lavreniuk[1] **and Mykyta Mudryi**[1,2] **and Markiian Chaklosh**[1,3]
[1]ARIMLABS.AI [2]Polish-Japanese Academy of Information Technology
[3]University of the National Education Commission in Kraków
{alavreniuk, mmudryi, mchaklosh}@arimlabs.ai

Abstract

In natural language processing, the entropy of a language is a measure of its unpredictability and complexity. The first study on this subject was conducted by Claude Shannon in 1951. By having participants predict the next character in a sentence, he was able to approximate the entropy of the English language. Several follow-up studies by other authors have since been conducted for English, and one for Hebrew. However, to date, Shannon's experiment has never been conducted for Ukrainian. In this paper, we perform this experiment for Ukrainian by recruiting 184 volunteers using social media channels. We rely on techniques used for English to approximate the entropy value of Ukrainian. The final result is an upper bound of $H_{upper} \approx 1.201$ bits per character. We compare this to the performance of current Large Language Models. The methods and code used are also documented and published, along with a discussion of the main challenges encountered.

1 Introduction

In information theory, entropy is a measure of the information content or surprise associated with an event.

For a natural language, entropy represents the unpredictability of its characters. Entropy is a fundamental property of the language itself. Measuring this value has broad implications, including but not limited to cross-linguistic information density comparisons, data compression, and language modeling.

This unpredictability cannot be calculated directly, as that would require knowing the true probability distribution of the language - which words follow which, which topics relate to each other, all for arbitrarily long contexts and across all speakers of that language.

Several methods for approximating the entropy of a language have been used.

Statistical N-gram-based models produce exact values for their corpora but are greatly limited by data sparsity as N grows. One such study was conducted for Ukrainian (Babenko and Sushko, 2012), estimating the entropy value for N values up to 9, with the result $H_9 = 1.25$–1.40 bpc.

Another method is compression-based analysis. Takahira et al. (2016) attempted to use PPM compression on massive corpora for entropy estimation, with the smallest bound for English being $H \approx 1.4$ bpc. However, this method consistently overestimates entropy due to the limited predictive capacity and restrictive context windows of compression algorithms.

Some neural-network approaches have been used to attempt to measure language complexity. They have the advantage of being able to access the model's internal probabilities and calculate entropy directly, avoiding bounds-based calculations. Takahashi and Tanaka-Ishii (2018) measured the complexity of English with neural networks, resulting in $H \approx 1.12$ bpc. However, these methods rely on minimizing the cross-entropy loss for a training dataset

$$D_{training} \subset D_{language}$$

, which means that the resulting entropy estimate reflects the chosen training corpus instead of the entire language. This makes overfitting to a small or unrepresentative corpus a significant concern. Humans are the source of ground truth for the natural language they speak, and as such are much less prone to overfitting. Modern Transformer-based language models perform significantly better than humans, but their training corpora might not represent their language fully, leaving gaps or overfocusing on some aspects of a language.

A different approach, using human knowledge, was proposed in 1951 by Claude Shannon (Shannon, 1951).

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 33–40

By making humans guess the next character in a sentence, it is possible to closely approximate a lower and upper bound on the true entropy of a natural language using the number of guesses it takes a human to guess the next character given N characters of context. Note that calculating the exact number would require knowing the exact probabilities that a human assigns to guesses in their head, which is impossible. One study (Cover and King, 1978) asked humans to output the entire probability distribution of their guesses. However, while theoretically improving the results, this severely strains human participants who are forced to set a percentage probability for each of the 27 characters. Additionally, humans' internal probabilities are likely different from what a human can actually output, as humans are not consciously aware of the exact computations performed by their brains.

As such, later studies used the method originally described in Shannon (1951), and aimed at increasing the number of participants to combat bias.

After considering the different methods available, we chose Shannon's experiment for estimating the entropy of Ukrainian - the first time this experiment has been performed for this language.

A section is also dedicated to using LLMs for entropy estimations.

2 Methodology

We conducted Shannon's experiment, based on the methodology described in Ren et al. (2019). Due to the volunteer-based nature of the experiment, extensive modifications were made to the experimental setting to ensure participant engagement and minimize frustration and early quitting. Namely, the volunteers were provided with 70 characters of initial context as opposed to starting from 0, and the dataset was filtered to sentences from 120 to 200 characters long. The user interface included a small delay after each attempt to prevent button mashing. Users could complete as many sessions as they wanted.

2.1 Participant Recruitment

Volunteers were primarily recruited through the Telegram messenger and word-of-mouth, with the pool consisting mostly of young adults and adults willing to take part in the experiment. The recruitment was conducted in Ukrainian, with participants expected to be native Ukrainian speakers, though language background was not formally assessed.

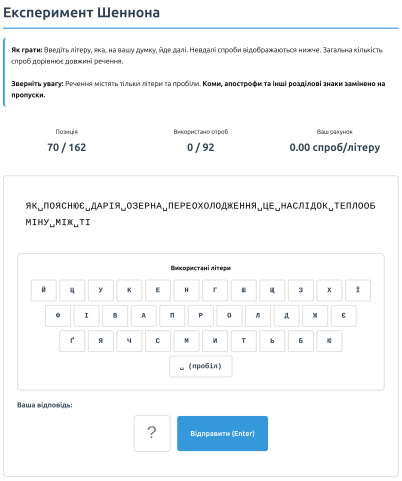


Figure 1: User interface used during the experiment

A giveaway was held to encourage participation, with victory odds being directly proportional to the number of sessions completed, and at least 2 completed sessions required to be eligible. The total experiment cost was ∼200 USD, split between giveaway prizes and website hosting. A total of 323 people started the registration process and 256 confirmed their intention to participate. Out of those, 184 started at least one session and 131 completed at least two sessions. The total number of completed sessions was 501, and data from 192 more incomplete sessions were also used in the final calculations. Overall participant engagement was moderate, with the median volunteer completing 2 sessions.

Following the methodology of Ren et al. (2019), who used news articles, our dataset consisted of 136 sentences collected from 5 news articles on different topics from Ukrainska Pravda, published from August 8, 2025 to January 23, 2026. The pre-processing of articles consisted of splitting into sentences on ".!?" and filtering out sentences that contained Latin characters or digits. All characters that are not whitespace or one of the 33 letters of the Ukrainian alphabet were replaced with whitespace. Lastly, all neighboring whitespaces were replaced with a single whitespace, and sentences shorter than 120 or longer than 200 characters were discarded.

For each session, a volunteer is presented with one sentence from the total sentence pool, with the initial 70 characters revealed, and prompted to guess the next character. They keep guessing that character until a correct one is entered, which constitutes a single observation - the number of attempts needed to guess a character at

position N. For example, a participant seeing the sentence "ЯК_ ПОЯСНЮЄ_ ДАРІЯ_ ОЗЕРНА_ ПЕРЕОХОЛОДЖЕННЯ_ ЦЕ_ НАСЛІДОК_ ТЕПЛООБМІНУ_ МІЖ_ ТІ"(*YAK POIASNIUIE DARIIA OZERNA PEREOKHYLODZHENNIA TSE NASLIDOK TEPLOOBMINU MIZH TI* 'as Dariia Ozerna explains, hypothermia is a consequence of heat exchange between [the] bo-') might guess the character "Н"(*N*) incorrectly, having it highlighted in red, then guess the character "Л"(*L*)correctly, resulting in gathering an observation of "2 guess attempts until correct at 70 characters of context", and the participant proceeding to guessing the next character.

The length of a session is limited by the number of guesses available. For each sentence, the number of total guess attempts available is equal to its total length minus the 70 revealed characters. As such, completing the sentence would require guessing each character correctly on the first try, and this was not expected to occur in practice. A session ends when all the attempts are exhausted.

After gathering a sufficient number of observations, the entropy bounds are calculated as follows.

Given K as the number of possible guesses (alphabet size of the target language + whitespace), and q_i as the fraction of characters guessed correctly on the i-th try, the true entropy H is approximated by

$$H_{lower} \leq H \leq H_{upper}$$

, where the upper bound is the entropy of the distribution of guess counts

$$H_{upper} = -\sum_{i=1}^{K} q_i \log_2 q_i$$

and the lower bound is the entropy of the "least surprising" distribution - one where each character guessed on the i-th try had a probability of $\frac{1}{i}$

$$H_{lower} = \sum_{i=1}^{K} q_i \log_2 i$$

(Shannon, 1951)

3 Result Analysis

In this analysis, we follow methods used in Ren et al. (2019).

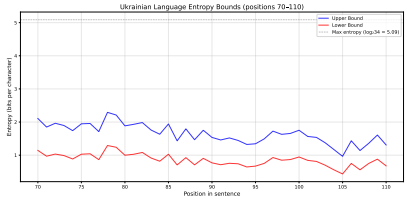


Figure 2: Lower and Upper bounds of entropy per position for positions 70-110

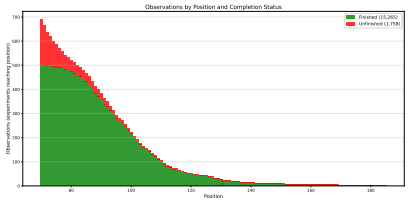


Figure 3: Number of observations gathered per position

Data collection took place from January 24 to January 31, 2026. During this period, 853 total sessions were started, out of which 501 were fully completed (ran out of guesses) and 352 were abandoned early. Note that guesses from abandoned sessions are also included in the final calculations. The completed sessions contributed 38,977 character guesses and the unfinished sessions contributed 5788 guesses for a total of 44,765 character guesses. Out of those, 17,023 guesses (38%) were correct. These 17,023 correct guesses correspond to 17,023 observations - ∼10% of Ren et al.'s scale of 172,954 observations. The number of observations is highest for positions 70-90, and drops off sharply after 100 as participants exhaust their guesses. We discard positions after 110 due to insufficient data.

The per-position entropy values exhibit considerable variance. A pooling approach, together with filtering, is used to mitigate this.

3.1 Constraining Positions

Ren et al. (2019) relied on the application of an ansatz function to fit the existing data and extrapolate the entropy value to infinite context length. This approach requires data starting from position 0, which was not gathered due to participant engagement concerns. Alternative curve-fitting approaches were attempted but were unreliable due to noise in the collected data. Instead, we use the fact that context beyond $N = 70$ only marginally improves guess accuracy. The value of the ansatz function from Ren et al. (2019) at $N = 70$ differs from the extrapolated value by approximately 1%.

While this ansatz value may not perfectly model English or Ukrainian, calculating an estimate using only positions 70-110, without extrapolation, is highly unlikely to introduce more error than other sources of variance in this experiment. Additionally, guess attempt limits bias the data by filtering out poor performers early and populating the larger N values only by observations from skilled guessers.

3.2 Data Trimming

For the purposes of calculating an upper bound on natural language entropy, we are interested in using only the results from top performers, as their results are an indication of low natural language entropy, and poor performance by a participant might indicate a lack of commitment or interest as opposed to language complexity. As such, we need to trim the existing data.

Ren et al. (2019) trimmed the 50% of the worst-performing sessions, dropping the final estimate by about 0.2 bpc. The goal is to discard poor performers without removing the natural variance in responses. However, this makes the final estimate more sensitive to cheating by top performers in online experimental settings.

We perform additional outlier detection beforehand to discard improbably good results. By running a binomial analysis and discarding sessions which had a $< 1\%$ chance of arising given the mean accuracy score, we discard 30 sessions (4.3%), marked as suspicious. This accounts for possible random guessing, as well as possible noise or cheating, as some participants reported that they were able to look up the full sentences online.

Additionally, we argue that volunteer participants are not as engaged as paid MTurk workers. We revise the trimming percentage to remove 65% of the worst performing humans. This trim is chosen due to having one of the smallest CI widths, and due to the insufficiency of the original 50% trim for volunteer participants.

Note that the choice of trim matters greatly for the final result. In Table 1, the effect of trim percentage on the final result is shown. Choosing a trim percentage is a balance of excluding underperforming participants and not removing true performers in favor of lucky guesses. More precise results can be achieved by using controlled environments and unpublished datasets (removing cheating), and additional participant encouragement, such as direct monetary compensation, participant filtering, and

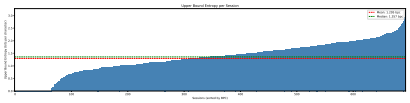


Figure 4: BPC per session for all sessions, ordered by performance.

directly rewarding better results (removing underperformance), as well as larger sample sizes.

3.3 Entropy Calculation

For a more robust entropy calculation, observations from positions 70-110 are pooled. This brings the overall number of observations used for this calculation post-trimming to 4,869.

The final upper entropy bound of the collected observations is derived from a mean of all upper entropy bounds in range 70-110, weighted by the number of observations for that position.

The formula used is

$$H_{upper} = \sum_{n=n_1}^{n_2} w_n \cdot H_{upper}(n), \quad w_n = \frac{N_n}{\sum_k N_k}$$

, where $n_1 = 70, n_2 = 110$, and N_n is the number of observations for position n, and $H_{upper}(n)$ is the upper bound calculated for position n.

The resulting upper bound is $H_{upper} = 1.201$ bpc.

The lower entropy bound, calculated in the same way for $H_{lower}(n)$, is $H_{lower} = 0.5987$ bpc.

3.4 Bootstrap Analysis

To quantify the uncertainty in this result, we perform a bootstrap analysis on the existing data. We resample sessions with replacement 2000 times to construct a bootstrap distribution of the upper bound estimate. The 95% CI is [1.10, 1.19], with a width of 0.090. The downward bias of the bootstrap median compared to the initial point estimate is consistent with small sample sizes (Ren et al., 2019). Note that the final calculations used 233 different sessions post-trim, which is approximately 23% of the 1000 sessions per N used for calculations in (Ren et al., 2019). Increasing sample size is expected to reduce confidence intervals further. Direct comparisons to (Ren et al., 2019) are not possible due to only pre-trimming 90% CIs being published by the authors.

3.5 Final Result

We report an upper bound on entropy for Ukrainian of 1.201 bpc, a redundancy of 76.4%.

Table 1: Sensitivity of the upper bound to bottom-trim level after binomial outlier removal. Session-level bootstrap, 2,000 iterations.

Bottom trim	Pool	Point est. (bpc)	Bootstrap median	95% CI	Width
0%	663	1.830	1.790	[1.731, 1.849]	0.118
10%	597	1.756	1.714	[1.656, 1.764]	0.109
20%	531	1.671	1.630	[1.582, 1.682]	0.099
30%	465	1.581	1.536	[1.485, 1.584]	0.099
40%	398	1.493	1.443	[1.389, 1.492]	0.103
50%	332	1.382	1.331	[1.279, 1.373]	0.094
55%	299	1.327	1.280	[1.233, 1.322]	0.088
60%	266	1.255	1.204	[1.154, 1.246]	0.092
65%	**233**	**1.201**	**1.149**	**[1.102, 1.192]**	**0.090**
70%	199	1.136	1.083	[1.029, 1.128]	0.098
80%	133	0.896	0.839	[0.753, 0.887]	0.134
90%	67	0.327	0.290	[0.211, 0.352]	0.141

Per Babenko and Sushko (2012), conditional n-gram entropy of Ukrainian is $1.25 - 1.40$ bpc for $N = 9$. This represents a more direct computation limited to short context windows. Conditional calculations for longer N values are impossible due to the frequency of n-grams diminishing rapidly. Our result using human participants confirms that the trend of lesser entropy with larger context windows continues beyond N=9 for Ukrainian.

3.6 Comparison to English

Comparing to the results of Ren et al. (2019) for English, we see very similar patterns.

Their experiment consisted of predicting 27 characters, 26 from the English alphabet and whitespace, with the max entropy of $\log_2(27) \approx 4.75bpc$. For Ukrainian, we use 34 characters, 33 letters of the alphabet + whitespace, with the max entropy of $\log_2(34) \approx 5.09bpc$.

Despite this, the entropy and redundancy values of $\approx$ $1.20bpc$ + 76.4% for Ukrainian and $\approx$ $1.22bpc$ + 74.3% for English are similar. It is likely that a more refined measurement for either language will result in lower estimates. Despite Ukrainian having 7 more letters, some are seldom used, contributing less to increasing the overall entropy value.

Both experiments suffered from an online setting, having some shared biases in the form of cheating and inconsistent participant engagement. Additionally, our trimming was more aggressive due to the use of volunteer participants.

In general, there are no substantial differences in the entropy estimates between the two languages.

4 Large Language Models as Predictors of Ukrainian

We compare human results to the performance of recent Transformer-based Large Language Models.

As outlined in Section 1, neural-network based methods provide lower entropy estimates than other approaches. The core concern that disallows using those estimates as true language entropy bounds is overfitting the test domain. Nevertheless, we provide results from LLM entropy measurements on the same data corpus, effectively comparing LLM performance to human results.

Given a language model M. For a string of N characters long text, a token-level language model tokenizes it into T tokens. Given T, BPC is equal to:

$$\text{BPC} = -\frac{1}{N}\sum_{i=1}^{T} \log_2 P_M(t_i \mid t_1, t_2, \ldots, t_{i-1})$$

, where t_i is the i-th token in the sequence, and $P_M(t_i \mid t_1, t_2, \ldots, t_{i-1})$ is the probability assigned to it by the model given i-1 preceding tokens. Dividing by N normalizes the value and removes tokenizer differences, allowing for direct comparisons.

In practice, language models add an additional cost from imperfect language modeling. For a language model, the BPC is equal to BPC $= H(p) + D_{\text{KL}}(p\|q) \geq H(p)$, where p is the true probability distribution of a language, q is the learned distribution, $H(p)$ is the entropy of the language, and $D_{\text{KL}}(p\|q)$ is the Kullback–Leibler divergence between the two distributions, which is

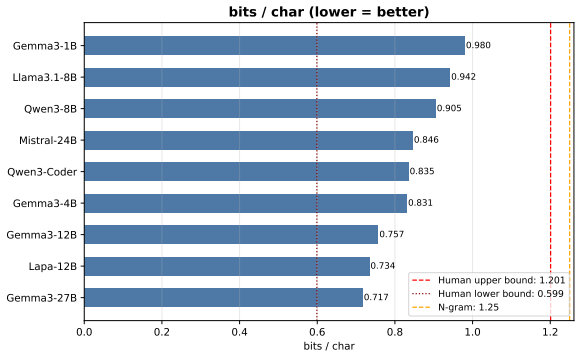


Figure 5: LLM Results for Shannon's experiment

the result of imperfect modeling. KL divergence is the number that is reducible by better training, and is the main metric that we are interested in when comparing models.

For measuring language knowledge, we use pre-trained model checkpoints before instruction tuning takes place. Fewer pre-processing steps are performed - characters are not capitalized and punctuation is not removed. We only count loss on tokens that start after position 70.

To prevent data contamination, the news dataset spans August 2025 - January 2026, after the training cutoffs for LLMs being benchmarked.

LLM results show substantial improvements over other measurement types. Llama 3.1, which does not officially support Ukrainian, scores worse than other models which do support Ukrainian. Gemma 27B and Lapa (Paniv et al., 2025), a Gemma 12B finetune on Ukrainian, show the best results, suggesting low KL divergence. Lapa has the highest fertility due to its updated tokenizer, however, its good results are not directly attributable to this. Many models get close to the human lower bound, suggesting that overfitting to the news domain is present. Note that those results were measured with sentence-level context windows of 120-200 characters, mirroring the main experiment. Gemma-3 scales cleanly across parameter counts (1B→4B→12B→27B), with BPC improving from 0.980 to 0.717

The relationship of fertility to BPC is not conclusive - Qwen models show decent results on Ukrainian despite fertility of $\approx$ 2 characters per token. Establishing a relationship of fertility to bpc requires follow-up investigation. All the obtained results are well below (Takahashi and Tanaka-Ishii, 2018)'s 1.12 BPC result, due to using much newer and larger transformer-based models.

The best results being in the $\approx$ 0.7BPC range imply that the real language entropy lies below our human experiment results. However, results being very close to the human lower bound suggest that overfitting is present. Quantifying the degree of overfitting is not trivial and requires follow-up investigation, however, modern language models might be a pathway to measuring entropy of natural language.

5 Conclusion

This paper presented the results of a replication of Shannon's experiment for the Ukrainian language. Relying on methods from Ren et al. (2019), we recruited volunteers and gathered 17,023 observations from 184 individuals. This was the first time this experiment was performed for the Ukrainian language. After extensive analysis and data filtering due to the online environment of the experiment, we calculate the upper bound of Ukrainian given at least 70 but no more than 110 characters of context to be $\approx 1.20bpc$, and, based on the ansatz function from Ren et al. (2019), argue that this value is close to the entropy bound given infinite context. This improves upon the only previous estimate for Ukraine by Babenko and Sushko (2012), whose n-gram-based approach yielded $> 1.25bpc$. The final results are compared to English, showing similar information densities, and to LLM performance, showing a greater predictive performance in language models. We describe the limitations of this approach and the difficulties in using volunteer participants, and point out directions for future work. We also publish code and data used in the experiment.

Limitations

The primary limitation of this experiment is the volume of data gathered.

184 participants contributed a total of 17,023 total and 4,869 post-filtering observations, which proved to be insufficient for a precise result. Comparing to Ren et al. (2019), a volunteer on average gathered 93 observations, compared to 253 observations for a paid MTurk worker. A volunteer is naturally less engaged than a paid worker, leading to lower accuracy and, consequently, higher estimates. Additionally, our experiment started from position 70 to ensure participant engagement. This makes some analysis, such as fitting an ansatz function or plotting the increase in precision with more

Table 2: LLM results on the experiment corpus, sorted by BPC.

Model	Params	Fertility (Chars/Token)	BPC
Gemma-3	27B	3.33	0.717
Lapa	12B	5.67	0.734
Gemma-3	12B	3.33	0.757
Gemma-3	4B	3.33	0.831
Qwen3-Coder-Next	80B	2.03	0.835
Mistral-Small	24B	2.95	0.846
Qwen3	8B	2.03	0.905
Llama-3.1	8B	3.24	0.942
Gemma-3	1B	3.33	0.980
Human upper bound	–	–	1.201
Human lower bound	–	–	0.599

context, impossible.

Additionally, the choice of a dataset influences the results greatly. Some parts of a language, such as informal speech, regional variations, and personal accounts of unknown events will naturally have much higher unpredictability compared to widely-known idioms, proverbs and works of art, which can be not only predicted but recited by most native speakers. We follow Shannon (1951) and Ren et al. (2019), which used datasets composed of news articles, with 100 and 225 sentences respectively. Our dataset consisted of 136 sentences sampled from 5 recent articles, which likely introduced bias due to limited topical and stylistic variety.

Future Work

For future work, we recommend recruiting at least 500 participants, and gathering at least 100,000 observations, with context lengths ranging from zero to several hundred characters, allowing for extrapolation via fitting of an ansatz function. Additional participant encouragement is also highly advised - we suggest more direct forms of compensation such as a monetary payment. Rewarding participants for better guesses should also be considered to improve the data quality. However, with more rewards for better results, cheating should be taken into account.

For dataset creation, it is advised to use fully private datasets of over 200 sentences, created specifically for this purpose. This eliminates cheating by lookup, and leaves only possible cheating by using a language model, a method that is likely too complicated for the average participant. If this is infeasible, news sampling should consist of at least 20 articles over wider time periods, resulting in 200-300 sentences.

For LLM usage, the core concerns to be addressed are modeling representative training and test sets. This raises the question of what constitutes a language, and what parts of a language are more important. This is left for future work.

Finally, the published experiment results enable some additional analysis that is out of scope for this work, such as examining how guessing accuracy varies by position within a word or by character.

Ethical Considerations

Use of Volunteers

Participants for the experiment were recruited using word-of-mouth and social media channels. Each participant was informed about the nature of the experiment prior to providing consent and no personal data was collected other than an optional request to be named in acknowledgments.

Use of AI-based Writing Assistance

During this study, large language models of the Claude family were used extensively for research, code generation, and proofreading purposes. All of the text was written by the authors.

Acknowledgements

We thank the following volunteers for contributing their time and effort to advancing Ukrainian linguistics: 007morf, Alina Kryvobokova, Anastasiia Loban, Andrii, AnimeGame, Artem Dzhemesiuk, Artem Sokolik, BeHappy1337, Bob, Bohdan

Rubakha, BrokenByteOfCode, DARIA, Danylo Vorvul, DitriX, Dmytro Potapov, Kateryna Hordiienko, KroZen_Dev, Maksbid, Mamontovozik, Mariia Chepeliuk, Sofiia Kuplevatska, Viktor Kauk, Vladis002, Yan Krasnyi, grantuseyes, lime, marmazon, moonlightqrl, popo, rawrshah, virusebe, Yevhen Domeretskyi, Yevhen Movsesov, Yevhen Tuturov, Yevhenii, Yevhenii, Yevhenii, Yevhenii Batiuta, Yehor, Yelyzaveta, Ivan, Ivan, Ivan, Ivan Turchyn, Ihor, Ihor, Ihor Braichenko, Illia Repin, Inna Kaliafitska, Iryna, Aleks, Aleks, Alinochka, Anastasiia, Anastasiia Pysmak, Anatolii, Anatolii, Anhelion, Andrii, Andrii B, Andrii Voinarovskyi, Andrii L., Andrii Liashchuk, Anna, Anton, Artem, Artem, Artem, Artem Olimpiiev, Artur, Artur Kornilov, Bairaktar, Vadym, Valeriia, Vlad Milka, Vladyslav, Vladyslav, Vladyslav Diachenko, Vladyslav Zamotailo, Volodymyr, Volodymyr Yatsynych, Vik Lisovyi, Viktoriia, Viktoriia, Vitalii, Vitalii, Havrylov Illia, Herman, Herman, Hladka Marharyta, Davyd4, Dmytro, Dmytro, Dmytro, Dmytro S, Dmytro Ch, Zhanna, Zlata, Kartoplianchyk, Kateryna, Kateryna Tatarchuk, Kolesnyk Vlad, Kostiantyn, Kisilov Kyrylo, Kit, Larysa, Liubov, Maksym, Maksym, Marharyta, Markian, Mariia, Mykyta, MykytaYe, Mykola, Mykola, Mykola Usik, Mykolai, Mykhailo, Mykhailo, Mohyla Stanislav, Miroshnykov Illia, Nazarii, Nataliia, Nataliia Hrechushkina, Nataliia Puniak, NeuroKit, Nikita, Oleh, Oleh, Oleksandr, Oleksandr, Oleksandr Bohun, Oleksandr D, Oleksandr and Iryna, Oleksandra, Oleksii Maryshchyn, Olena, Olha, Olha, Pavlo, Pavlo Chuiko, Roman P., Roman Serdiuk, Rina, Rinat Marinad, Riia Kovalenko, Svizhyi, Serhii, Serhii O., Sonia, Stesi, Tamila Krashtan, Taras, Yuliia, Yurii, Yurii, Yurii Prokopenko, Yaroslav, Yaroslav, and an additional 25 people who have decided to stay anonymous.

We thank Yurii Paniv and Mariana Romanyshyn for valuable discussion that helped shape this work, and NeuroKit for invaluable assistance with volunteer recruitment.

References

Tetiana V. Babenko and Svitlana O. Sushko. 2012. About an entropy of Ukrainian language. *Ukrainian Information Security Research Journal*, 14(3):104–107.

Thomas M. Cover and Roger C. King. 1978. A convergent gambling estimate of the entropy of English. *IEEE Transactions on Information Theory*, IT-24.

Yurii Paniv, Bohdan Didenko, Mykola Haltiuk, Vladyslav Humennyy, Andrian Kravchenko, Roman Kyslyi, Viktoriia Makovska, Artem Orlovskyi, Bohdan Ruban, Maksym-Yurii Rudko, Anastasiia Senyk, Nazarii Drushchak, Dmytro Chaplynskyi, and Mariana Romanyshyn. 2025. Lapa LLM v0.1.2 — the most efficient Ukrainian open-source language model.

Guangyu Ren, Shuntaro Takahashi, and Kumiko Tanaka-Ishii. 2019. Entropy rate estimation for English via a large cognitive experiment using Mechanical Turk. *Entropy*, 21(12):1201.

Claude Elwood Shannon. 1951. Prediction and entropy of printed English. *Bell System Technical Journal*.

Shuntaro Takahashi and Kumiko Tanaka-Ishii. 2018. Cross entropy of neural language models at infinity—a new bound of the entropy rate. *Entropy*, 20(11):839.

Ryosuke Takahira, Kumiko Tanaka-Ishii, and Łukasz Dębowski. 2016. Entropy rate estimates for natural language—a new extrapolation of compressed large-scale corpora. *Entropy*, 18(10):364.

A Published Data

We publish the code and data used in this work, as well as the raw data gathered from volunteers. The results of the experiment are anonymized.

Experiment and processing code: https://github.com/arimlabs/ule

Dataset: https://huggingface.co/datasets/a-l-o/shortnews

Data gathered: https://huggingface.co/datasets/a-l-o/ule_results

SimIdioms: A Corpus and Benchmark for Ukrainian Idiom Translation

Yaryna Petruniv
Ukrainian Catholic University
petruniv@ucu.edu.ua

Iuliia Makogon
Independent Researcher
juogon@gmail.com

Roman Kyslyi
Kyiv School of Economics
rkyslyi@kse.org.ua

Abstract

We present a corpus of aligned Ukrainian–English idiomatic expressions and a comprehensive evaluation of six large language models on the task of translating sentences containing idioms. The corpus is constructed by linking entries across multiple phraseological dictionaries and the MIDAS corpus using vector similarity search, enriched with figurative meanings, contextual sentences from the UberText fiction corpus, and semantic transparency scores. We evaluate Gemini 2.5 Flash, Claude Haiku 4.5, Gemma 3 12B, Qwen3-30B-A3B, LapaLM, and Tiny Aya Global in both Ukrainian-to-English and English-to-Ukrainian directions under default and context-augmented prompting. Our evaluation of 65,723 translations reveals a pronounced direction asymmetry, with all models performing substantially worse when translating into Ukrainian. Identifying the source idiom in the prompt improves quality for most models in Ukrainian-to-English but has limited effect in the reverse direction, suggesting that the bottleneck there is morphological generation rather than idiom recognition. We additionally show that semantic transparency of idioms is only weakly correlated with translation quality. We release the corpus[1] and evaluation framework[2] to support research on idiomatic translation for mid-resource languages.

1 Introduction

Idiomatic expressions pose a fundamental challenge for machine translation: their meaning cannot be derived compositionally from constituent words, requiring models to recognize figurative usage and retrieve appropriate equivalents in the target language. While recent large language models have demonstrated impressive translation capabilities on standard benchmarks (Paniv, 2025), their ability to handle phraseological units remains underexplored, particularly for mid-resource languages such as Ukrainian.

This work addresses this gap with three contributions. First, we construct a corpus of aligned Ukrainian–English idiom pairs enriched with figurative meanings, contextual sentences, and semantic transparency scores (Section 3). Second, we evaluate six LLMs in both translation directions under default and context-augmented prompting, using an LLM-as-judge evaluation paradigm (Section 4). Third, we provide a detailed analysis of the factors influencing translation quality, including direction asymmetry, context utilization, semantic transparency, and model-specific failure modes (Sections 5–6).

2 Related Work

2.1 Multi-Word Expression and Idiom Corpora

Idioms are a subclass of multi-word expressions (MWEs) — conventionalized sequences of words whose meaning is often not fully derivable from their parts. The construction of idiom resources for computational research has received growing attention, evolving significantly in their creation methodologies from manual curation to AI-assisted generation. A prominent example of datasets that heavily rely on manual or crowdsourced annotations applied to automatically extracted text is the MAGPIE corpus (Haagsma et al., 2020), a single-language English resource containing tens of thousands of instances. To accelerate the development of such resources and handle their inherent complexity, researchers have increasingly adopted model-in-the-loop and LLM-generated approaches. The examples are FLUTE GPT-3 assisted dataset (Chakrabarty et al., 2022) or the recent MIDAS corpus (Kim et al., 2025) which employs LLMs for the initial refinement of idiomatic meanings followed

[1] https://huggingface.co/datasets/KSE-RESEARCH-Group/sim-idioms

[2] https://github.com/petrunivyaryna/sim-idioms

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 41–52

by native-speaker annotation.

These resources also differ fundamentally in their approach to multilinguality. While corpora like MAGPIE and FLUTE are strictly monolingual, the MIDAS corpus provides separate, non-parallel data for six typologically diverse languages. In contrast, datasets such as LIdioms (Moussallem et al., 2018) are explicitly designed as parallel corpora where data for different languages are manually synchronized and linked based on semantic equivalence. Flor et al. (2025) provide a comprehensive survey of these idiom datasets across psycholinguistic and computational paradigms, highlighting the persistent scarcity of high-quality resources for non-English languages. Furthermore, there remains a critical linguistic blind spot: none of these established benchmarks currently targets or supports the Ukrainian language.

2.2 Psycholinguistic Characteristics of Idioms

The psycholinguistic literature (Nunberg et al., 1994) classifies idioms as normally decomposable, abnormally decomposable, or semantically non-decomposable. Transparency is defined as the ease with which the structural motivation of an idiomatic expression can be deduced from its literal analysis. In our work, we operationalize this construct as the cosine similarity between idiom phrase embeddings and figurative meaning embeddings, following a distributional approach to transparency scoring. Computational assessment of semantic transparency was previously utilized by Gao et al. (2025) for Chinese idiom datasets. Kim et al. (2025) investigated whether LLMs rely on memorization or genuine reasoning when processing idiomatic expressions, finding that model performance degrades substantially on novel or low-frequency idioms. This motivates our evaluation design, which tests models on a diverse set of Ukrainian idioms with varying degrees of transparency.

2.3 Metrics for Idiom Translation

The inadequacy of surface-level metrics such as BLEU and METEOR for evaluating figurative language translation has been widely noted. Yang et al. (2025) demonstrated that reference-based metrics systematically underestimate the quality of idiomatic translations that use valid but lexically different target-language equivalents. Li et al. (2023) proposed augmenting translation systems with an idiom knowledge base (IdiomKB) and showed that standard metrics fail to capture improvements in figurative adequacy. These findings motivate our adoption of an LLM-as-judge evaluation paradigm (Zheng et al. 2023, Donthi et al. 2025), which can assess both meaning preservation and idiomatic adequacy without being constrained to exact lexical matches.

2.4 LLM Translation Benchmarking for Ukrainian

Several works address the need to evaluate LLM performance on tasks tailored to Ukrainian that may require cultural understanding. Recognizing limitations in translation capacity is especially important in adaptations of metrics constructed for high-resource languages with different grammar and morphology, e.g., Kravchenko et al. (2025), where translated cross-lingual pairs were used to assess LLM moral and cultural alignment. Researchers encounter difficulties applying traditional translation metrics to LLM-generated texts: for example, Paniv et al. (2025) report SacreBLEU on Multi30K-UK (Saichyshyna et al., 2023), a Ukrainian extension of the Multi30K multimodal benchmark. The recent LLM benchmark by Paniv (2025) includes translation evaluation on FLORES and LongFLORES for English–Ukrainian pairs with the BLEU metric. However, as demonstrated by Yang et al. (2025), such metrics do not suit phraseologically rich texts well, motivating the need for evaluation approaches specifically designed for figurative language.

3 Dataset Creation

3.1 Data Sources

We constructed the corpus of Ukrainian idioms and their English equivalents using multiple sources and matching strategies. We are grateful to the Condor Publishing House for granting permission to use the "Ukrainian-English and English-Ukrainian Phraseological Dictionary" (Horot et al., 2024) for academic purposes. Our dataset is primarily based on this dictionary, which provides Ukrainian and English idioms with corresponding translations but lacks consistent definitions or usage examples. To enrich the corpus with figurative meanings and contextual sentences, we incorporated the "Dictionary of Phraseological Units of the Ukrainian Language" by Bilonozhenko et al. 2003, which provides Ukrainian idioms with figurative interpretations and usage examples, as well as the English portion of the MIDAS corpus (Kim et al.,

Pattern	Count
VERB + obl	631
VERB + obj	558
NOUN + amod	251
NOUN + nmod	155
NOUN + case	145
VERB + obj + obl	115

Table 1: Most common syntactic patterns (root POS + dependency labels).

2025), which contains English idioms with figurative meanings and example sentences. Together, these resources allow the final corpus to capture not only cross-lingual idiom alignments, but also meaning and context on both sides.

We applied OCR to scanned dictionary pages that were not available in digital form.

3.2 Context for idioms retrieval

To address the shortage of contextual sentences for idioms, we used the UberText corpus (Chaplynskyi, 2023), focusing on its Fiction subset. This corpus contains modern Ukrainian texts segmented into sentences. The goal of this step was to automatically identify sentences in which a target idiom appears, including inflected forms and variations in word order within the idiom.

Each idiom can be parsed into a root token and its immediate dependency relations. We used this dependency representation to build reusable matching templates for idiom retrieval. For each idiom, we extracted the POS tag of the root token and the dependency labels of its direct children. We then grouped idioms into clusters that share the same root POS and the same set of dependency relations. This clustering step allowed us to define one dependency pattern per cluster, rather than writing a separate pattern for every idiom. Table 1 summarizes the most common syntactic patterns in our idiom set.

For each cluster, we constructed a dependency pattern using spaCy's dependency matcher [3]. The pattern requires a root token with a given POS tag and child tokens attached to the root with the required dependency labels. In parallel, we built a lookup map for each pattern that maps a tuple of lemmas (the root lemma and the required child lemmas) to the corresponding idiom.

[3] https://github.com/explosion/spacy-models/releases/tag/uk_core_news_md-3.8.0

We then applied the dependency matcher to sentences from the UberText Fiction subset. Each sentence was parsed and checked against the set of dependency patterns. When a pattern matched, we formed the same lemma tuple used in the idiom lookup map (the root lemma plus the lemmas of the required dependency children) and used it to retrieve the corresponding idiom for that pattern.

To address false positives after dependency matching, we added a validation step using Gemini 2.5 flash. For each idiom-sentence pair, the model verified the phrase presence and labelled the usage as idiomatic or literal.

This approach helped us obtain additional context for 1,639 Ukrainian idioms.

3.3 Vector Indexing in Qdrant

Because the three main resources provide different types of information, they are not aligned at the entry level (see Table 2). Condor provides Ukrainian–English pairs, whereas Bilonozhenko and MIDAS provide figurative meanings and example sentences for Ukrainian and English idioms, respectively. To merge these sources into a single corpus, we needed a way to identify corresponding idiom entries across resources. To this end, we used Qdrant, a vector database for similarity search, to index idiom texts as embeddings and retrieve semantically similar candidate entries across sources.

We first transformed all sources into a unified set of records. Each record corresponds to a text fragment taken from a dictionary entry. Depending on the source, this fragment is an idiom string, a translation equivalent, a figurative meaning, or an example sentence. Each record is stored together with metadata: a stable idiom identifier, the source name, and the language.

We embedded the text of each record using the multilingual sentence embedding model Multilingual-E5-large (Wang et al., 2024) and stored the resulting vectors in a Qdrant collection configured with cosine similarity. The embedding vector is stored in the index, while all metadata fields are stored as payload.

3.4 Building the Final Aligned Corpus

We treated corpus construction as a linkage problem across sources: entries that refer to the same idiom should be connected, and the final corpus should consist of consistent cross-source groups.

Source	# UK	# EN	# Meanings	# Examples	Notes
Condor (UK–EN)	4,565	12,159	✗	✗	EN are translation equivalents (not necessarily idioms)
Condor (EN–UK)	14,342	5,638	✗	✗	UK are translation equivalents (not necessarily idioms)
Bilonozhenko	3,304	✗	5,176	8,441	UA idioms
MIDAS (EN)	✗	12,662	11,806	19,410	EN idioms

Table 2: Summary statistics for the main sources used to construct the corpus.

First, we defined which source–field subsets were allowed to be linked. We generated links only for the pairings shown in Table 8.

The matching pipeline consists of several steps:

1. **Candidate retrieval:** For each pair, we used Qdrant to retrieve the most similar candidates from the target subset for every entry in the source subset.

2. **Threshold filtering:** We kept a candidate link only if its similarity score exceeded a predefined threshold. In our implementation, the threshold depends on the syntactic type of the idiom root in the source subset: 0.90 for common POS types and 0.95 for all others.

3. **Edge storage:** All accepted matches were stored in a separate Qdrant collection as edge records. Each edge stores the two connected idiom IDs, along with metadata such as the matching direction and the reason for the link.

After the matching step, we built an undirected graph in which the nodes correspond to the idiom identifiers and the edges correspond to the links from the link collection. We defined the final clusters as the connected components of this graph. Each connected component was assigned a unique cluster identifier (e.g. SIM-i) and represents a set of entries that are linked directly or via intermediate nodes.

Field	EN	UK
example	2,885	2,737
figurative meaning	1,550	363
figurative meaning translated	173	830
idiom	3,595	2,751
translation	3,427	3,026

Table 3: Item counts per field in the final corpus of 2,262 clusters.

To reduce manual workload, we assigned each cluster an initial decision label: KEEP or ANNOTATE. The label was determined based on source coverage. All ANNOTATE-labeled clusters were reviewed manually. The counts of unique entities grouped into 2,262 clusters are presented in Table 3.

3.5 Semantic Transparency Scoring

To characterize the difficulty of each idiom cluster, we computed semantic transparency scores that quantify how close an idiom's literal wording is to its figurative meaning. For each cluster, we embedded both the idiom phrase and its figurative meaning using Multilingual-E5-large (Wang et al., 2024) and computed the cosine similarity between the resulting vectors. Scores were computed separately for the Ukrainian and English sides, and an aggregate score was derived.

Out of 526 clusters with computed transparency scores, 77 were flagged as outliers, cases where the semantic transparency score was anomalously high, suggesting that the expression is more compositional than idiomatic. The remaining 449 clusters constitute the core evaluation set with genuine figurative expressions.

4 Experimental Setup

4.1 Models

We evaluated six large language models spanning different model families and sizes: Gemini 2.5 Flash (Google DeepMind, 2025), Claude Haiku 4.5 (Anthropic, 2025), Gemma 3 12B (Gemma Team, Google DeepMind, 2025), Qwen3-30B-A3B (Qwen Team, 2025), LapaLM (open large language model based on Gemma-3-12B adapted for Ukrainian language processing Team 2025), and Tiny Aya Global (Salamanca et al., 2026). The selection includes both proprietary API-based

models (Gemini, Claude) and open-weight models (Gemma, Qwen, LapaLM, Tiny Aya Global), allowing us to compare translation quality across different accessibility tiers. Notably, LapaLM is a Ukrainian-adapted model based on the Gemma 3 architecture, making the LapaLM–Gemma comparison a direct test of whether Ukrainian-specific fine-tuning improves idiomatic translation.

4.2 Translation Directions and Prompt Types

Each model was tested in two translation directions — Ukrainian-to-English (UK→EN) and English-to-Ukrainian (EN→UK) — under two prompt conditions:

- **Default:** The model receives only the source sentence and a translation instruction.
- **With context:** The prompt additionally identifies the source idiom by name. No figurative meaning or candidate equivalent is provided — only the idiom string itself, sufficient to signal that a non-literal interpretation is required.

This design isolates the contribution of idiom *recognition* from the contribution of richer semantic scaffolding such as figurative meanings or candidate target-language equivalents (Li et al., 2023; Donthi et al., 2025). It allows us to measure both baseline translation ability and the marginal benefit of explicit idiom identification.

4.3 Evaluation

We employed an LLM-as-judge evaluation paradigm (Zheng et al., 2023) using Gemini 2.5 Flash as the judge model. Each translation was evaluated on a 3-point scale:

- **Score 3:** The figurative meaning is preserved *and* the translation uses an idiomatic expression in the target language.
- **Score 2:** The figurative meaning is preserved but the translation is literal (no target-language idiom).
- **Score 1:** The figurative meaning is lost or severely distorted.

We frame Score 3 as a measure of *idiomatic retrieval competence*: the model's ability to recall a target-language idiom and produce it in context, rather than translation correctness in an absolute sense. A meaning-preserving but literal translation (Score 2) is not a translation failure; it is a distinct outcome that we analyze separately. We adopt this framing in both directions, including English-to-Ukrainian, because a model that reliably produces idiomatic output where appropriate demonstrates deeper phraseological knowledge of the target language than one that only produces literal renderings, even when both convey the same meaning. This is the capability we aim to assess.

In addition to the numeric score, the judge annotated each translation with binary labels for *meaning preservation* and *target idiom usage*, together with free-text reasoning. The total evaluation corpus comprises 65,723 annotated translations: 32,158 for UK→EN and 33,565 for EN→UK.

To validate the reliability of the LLM judge, we conducted a manual evaluation on a random sample of 100 sentences from the data set. These examples were independently annotated by a human evaluator using the same scoring criteria. The annotations were then compared with the scores produced by the LLM judge.

The comparison showed that the judge achieved an accuracy of 87% for the translation of the UK→EN and 78% for the translation of the EN→UK, indicating that the LLM-based evaluation is reasonably aligned with human judgment for the task of assessing idiom translation.

A potential limitation of this approach is the judge model's knowledge of Ukrainian idioms. Determining idiomaticity requires recognizing whether a phrase is an established Ukrainian idiom, which large language models may fail to do. To mitigate this issue, prompts include golden examples of acceptable translations and use separate instructions for each translation direction, with stricter idiomaticity guidance in the EN→UK setup.

5 Results

5.1 Overall Model Comparison

Table 4 and Figure 1 present the mean translation scores across all conditions. A clear three-tier structure emerges: Gemini leads with a mean score of 2.54, followed by Claude (2.32) and Gemma (2.29) in a middle tier, while LapaLM (2.15), Qwen (2.08), and Tiny Aya (1.89) form the lower tier.

5.2 Translation Direction Asymmetry

All models perform substantially worse on EN→UK than on UK→EN, as visualized in Figure 2. The largest directional gap is observed for Qwen (−0.57), while Gemini shows the most sta-

Model	Overall	UK→EN	EN→UK	Δ
Gemini	2.54	2.66	2.42	−0.24
Claude	2.32	2.56	2.09	−0.47
Gemma	2.29	2.49	2.09	−0.40
LapaLM	2.15	2.34	1.98	−0.37
Qwen	2.08	2.37	1.80	−0.57
Tiny Aya	1.89	2.11	1.68	−0.43

Table 4: Mean translation quality scores (1–3 scale). Δ shows the drop from UK→EN to EN→UK.

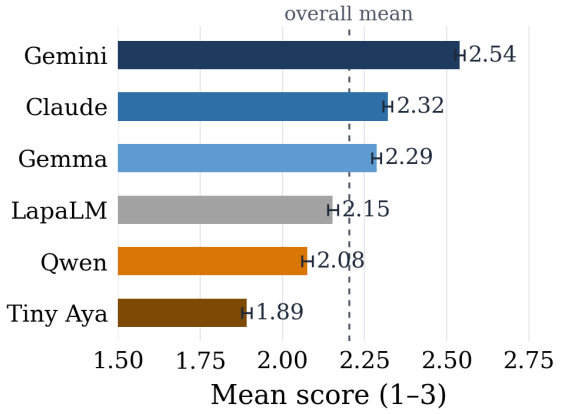


Figure 1: Mean translation quality scores by model across all conditions. Error bars show 95% bootstrap confidence intervals.

ble performance across directions (−0.24). This asymmetry is consistent with the morphological richness of Ukrainian and the greater challenge of generating idiomatic expressions in a morphologically complex target language.

In the UK→EN direction, the percentage of perfect translations (score 3) ranges from 40.4% (Tiny Aya) to 74.0% (Gemini). In the EN→UK direction, these rates drop substantially: from 17.0% (Tiny Aya) to 54.1% (Gemini). Meaning preservation rates remain comparatively stable across directions for the top models (Gemini: 91.8% vs. 93.7%), but degrade notably for lower-tier models (Tiny Aya: 70.6% vs. 61.2%).

5.3 Effect of Context

Table 5 and Figure 3 summarize the impact of identifying the source idiom in the prompt. In the UK→EN direction, context yields meaningful improvements for five out of six models, with the largest gain for Gemma (+0.233) and the smallest for LapaLM (+0.033).

In the EN→UK direction, context benefits are dramatically reduced. Only Gemini (+0.161) and Gemma (+0.133) show notable improvements. LapaLM, Qwen, and Tiny Aya show near-zero or negative deltas. This suggests that generating morphologically correct idiomatic Ukrainian is a bottleneck that additional semantic context alone cannot overcome.

LapaLM stands out as uniquely unable to leverage context in either direction ($\Delta_{\text{ctx}} = +0.033$ and $+0.000$), suggesting fundamental limitations in incorporating auxiliary information during generation.

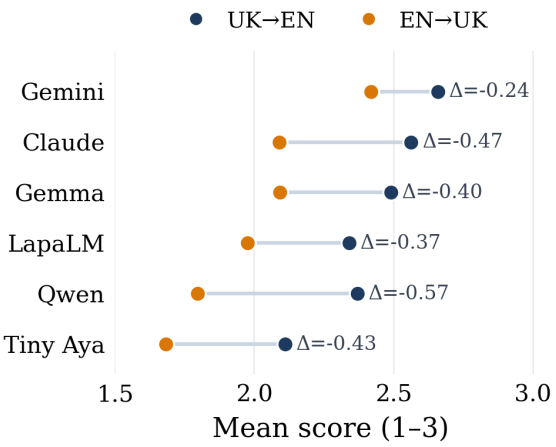


Figure 2: Direction asymmetry: performance drop from UK→EN to EN→UK per model.

	UK→EN		EN→UK	
Model	Default	Δ_{ctx}	Default	Δ_{ctx}
Gemini	2.587	+0.142	2.339	+0.161
Claude	2.454	+0.218	2.043	+0.094
Gemma	2.374	+0.233	2.026	+0.133
LapaLM	2.324	+0.033	1.975	+0.000
Qwen	2.278	+0.184	1.784	+0.024
Tiny Aya	2.062	+0.097	1.688	−0.009

Table 5: Default scores and context deltas (Δ_{ctx}) by direction. Positive values indicate improvement from context.

5.4 Score Distribution and Failure Modes

Table 6 details the score distribution for each model and direction. In UK→EN, Gemini achieves 74.0% perfect translations, whereas Tiny Aya reaches only 40.4%. In EN→UK, the proportion of complete failures (score 1) increases substantially for all models: Tiny Aya and Qwen produce over one-third score-1 translations.

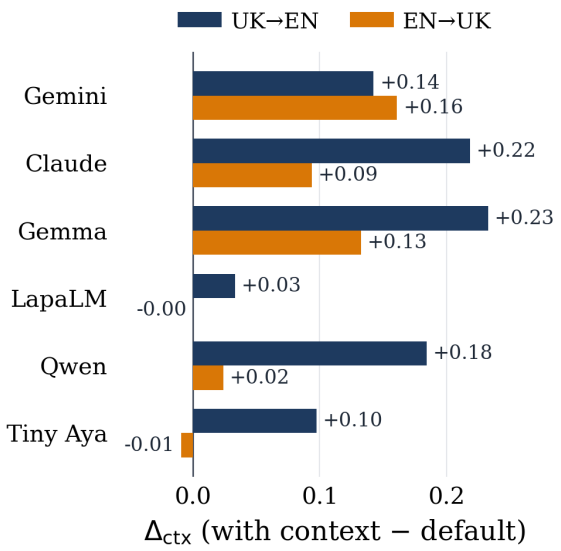


Figure 3: Context effect (Δ_{ctx}) by model and direction. Bars show the improvement from identifying the source idiom in the prompt.

Model	UK→EN (%) 1	2	3	EN→UK (%) 1	2	3
Gemini	8.2	17.8	74.0	12.2	33.8	54.1
Claude	12.8	18.2	69.1	23.1	44.9	32.0
Gemma	14.9	21.2	64.0	20.8	49.1	30.0
LapaLM	18.3	29.3	52.4	32.5	37.5	30.0
Qwen	18.4	26.2	55.4	39.2	42.1	18.8
Tiny Aya	29.4	30.2	40.4	48.6	34.4	17.0

Table 6: Score distribution by model and direction.

5.5 Semantic Transparency and Outlier Analysis

We examined whether semantic transparency — the degree to which an idiom's literal wording reflects its figurative meaning — predicts translation quality. Transparency scores were estimated automatically using sentence embedding similarity between idioms and their figurative meanings. For this analysis, we average translation scores across both prompt conditions to derive a single quality score per cluster. Of the 526 clusters with computed transparency scores, 466 have at least one translation in our evaluation set and are used in this analysis.

Figure 4 visualizes the relationship at the cluster level. We observe a weak but statistically significant positive correlation between semantic transparency and translation quality ($r = 0.136$, $p = 0.003$, $n = 466$). This indicates that more compositional expressions are marginally easier to translate, as expected.

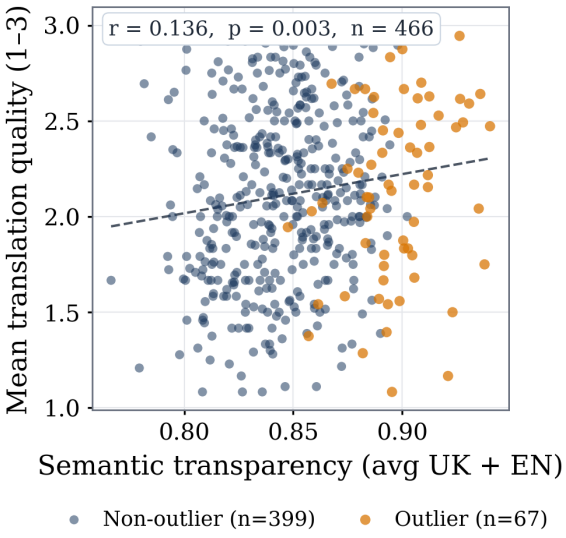


Figure 4: Semantic transparency vs. mean translation quality per cluster, averaged over both prompt conditions ($r = 0.136$, $p = 0.003$, $n = 466$). Each point is one idiom cluster.

Model	All	No outliers	Outlier only
Gemini	2.539	2.497	2.552
Claude	2.321	2.272	2.399
Gemma	2.286	2.236	2.389
LapaLM	2.153	2.066	2.295
Qwen	2.076	2.015	2.148
Tiny Aya	1.891	1.810	2.055

Table 7: Mean scores on clusters with transparency data: all clusters, non-outlier clusters, and outlier-only clusters (both prompts averaged).

Clusters flagged as outliers (77 of 526) — those with anomalously high semantic transparency — yield higher mean scores across all models (Table 7). However, removing these outlier clusters has minimal impact on overall scores (-0.04 to -0.09), confirming that the main findings are robust and not driven by quasi-compositional expressions.

6 Discussion

Our results reveal several important findings for the evaluation of idiomatic translation by LLMs.

Direction asymmetry. The pronounced performance gap between UK→EN and EN→UK demonstrates that translating *into* a morphologically rich language poses fundamentally different challenges. While models can often identify the figurative meaning of a Ukrainian idiom and render it in English, the reverse task — selecting and

correctly inflecting a Ukrainian idiom — proves substantially harder. The gap is largest for Qwen (-0.57) and smallest for Gemini (-0.24), suggesting that model-specific factors beyond language-pair difficulty are at play. We note, however, that morphological complexity and mid-resource pre-training exposure are observationally confounded for Ukrainian and we cannot fully separate their respective contributions; we discuss this further in the Limitations section.

Context utilization as a diagnostic. The differential ability to leverage context is diagnostic of model capabilities. Context helps most models in UK→EN but only Gemini and Gemma in EN→UK, suggesting that weaker models lack the morphological and syntactic competence to translate a known idiom into grammatically correct Ukrainian output. LapaLM is uniquely unable to leverage context in either direction ($\Delta_{\text{ctx}} = +0.033$ and $+0.000$), pointing to fundamental limitations in how this model incorporates auxiliary information during generation.

Ukrainian-specific fine-tuning does not help idiom translation. One of our most notable findings is that LapaLM — a model specifically adapted for Ukrainian based on the Gemma 3 architecture (Team, 2025) — performs *worse* than its base model Gemma across all conditions. In UK→EN, LapaLM scores 2.34 vs. Gemma's 2.49. In EN→UK, 1.98 vs. 2.09. The gap is even more pronounced with context: Gemma achieves context deltas of $+0.233$ (UK→EN) and $+0.133$ (EN→UK), while LapaLM's are $+0.033$ and $+0.000$. This may possibly happen due to catastrophic forgetting of pre-trained phraseological knowledge during Ukrainian-specific fine-tuning. This is consistent with observations by Paniv (2025), who found that instruction-tuned models sometimes underperform base models on Ukrainian tasks.

Model tiers and training data. The three-tier structure (Gemini > Claude/Gemma > LapaLM/Qwen/Tiny Aya) likely reflects differences in both pre-training data coverage and instruction-tuning quality for Ukrainian. API-based models with proprietary training pipelines (Gemini, Claude) generally outperform open-weight alternatives, consistent with findings on broader Ukrainian benchmarks (Paniv, 2025). The primary failure mode across all models is *literal translation*: models preserve semantic content but fail to produce target-language idioms, indicating that the challenge is figurative expression retrieval, not comprehension — a pattern also observed by Li et al. (2023) in multilingual settings.

Semantic transparency. The weak correlation between semantic transparency and translation quality ($r = 0.136$, $p = 0.003$) suggests that, while compositional idioms are marginally easier to translate, the primary difficulty lies in the model's phraseological competence rather than the transparency of individual expressions. This aligns with psycholinguistic findings that idiom processing involves both compositional and non-compositional pathways (Titone and Connine, 1999), and that even "transparent" idioms require culturally specific knowledge for appropriate translation.

7 Conclusion

We presented a comprehensive evaluation of six LLMs on the task of translating Ukrainian idiomatic expressions. Our corpus of aligned Ukrainian–English idiom pairs, enriched with figurative meanings, contextual sentences, and semantic transparency scores, provides a challenging testbed for assessing phraseological competence.

The evaluation of 65,723 translations reveals that:

1. All models perform substantially better on UK→EN than EN→UK, demonstrating that generating idiomatic Ukrainian is fundamentally harder than generating idiomatic English.

2. Identifying the source idiom in the prompt improves translation quality for most models in UK→EN but has limited or no effect in EN→UK, suggesting that morphological generation is the bottleneck.

3. Ukrainian-specific fine-tuning (LapaLM) does not improve idiom translation over the base model (Gemma), and in fact degrades both absolute performance and context utilization.

4. The primary failure mode is literal translation rather than meaning distortion: models generally preserve the intended meaning but fail to retrieve target-language idioms.

5. The model ranking is stable across conditions, with a clear three-tier structure: Gemini, Claude/Gemma, LapaLM/Qwen/Tiny Aya.

6. Semantic transparency has only a weak (though statistically significant) correlation with translation quality ($r = 0.136$, $p = 0.003$), and removing semantically transparent outlier clusters has limited impact on overall findings (-0.04 to -0.09).

These results highlight the need for evaluation frameworks that go beyond surface-level metrics when assessing figurative language translation. We release the corpus and evaluation code to support future research on idiomatic translation for Ukrainian and other mid-resource languages.

Limitations

Our study has several limitations. First, we evaluate a single language pair (Ukrainian–English); the findings may not generalize to other mid-resource languages with different morphological characteristics.

Second, we cannot fully disentangle Ukrainian's morphological complexity from its mid-resource status in pre-training data: most LLMs see substantially less Ukrainian than English, *and* Ukrainian morphology is considerably richer. Our claims about EN→UK morphological difficulty should therefore be read as joint claims about morphology and exposure; separating the two requires a controlled cross-lingual comparison, which we leave for future work.

Third, the LLM-as-judge evaluation, while scalable, may introduce systematic biases compared to human annotation. We validated judge outputs against expert annotations on a sample, but did not perform a qualitative error analysis. A related concern is that the same model family (Gemini 2.5 Flash) is used both as judge and as one of the evaluated translators, which could produce a self-preference bias; we mitigated this with golden references and explicit rubrics rather than open-ended preferences, and a fully independent judge would strengthen future evaluations.

Fourth, the corpus covers primarily literary and journalistic idioms from dictionaries and the UberText fiction subset — domain-specific idioms (e.g., legal, medical) are underrepresented, and coverage is bounded by dictionary publication dates.

Finally, transparency scores were computed with a single embedding model (Multilingual-E5-large), and our evaluation is limited to six models at a single point in time; different embeddings or future model versions may yield different results.

Ethics Statement

The idiom corpus was constructed from publicly available resources and a dictionary used with explicit permission from the Condor Publishing House. The UberText corpus is publicly available for research purposes. We used LLMs for evaluation, which may reflect biases present in their training data. The work focuses on evaluation rather than deployment, and we do not foresee direct negative societal impacts.

License and data release. The released corpus contains derived structured data (cluster identifiers, alignment graph, transparency scores, and contextual sentences retrieved from UberText) together with the English-side content from MIDAS (Kim et al., 2025). We make the following terms explicit:

- The English-side content and the cluster-graph structure are released under CC BY-SA 4.0.
- The Ukrainian idiom strings, figurative meanings, and usage examples derived from the "Ukrainian-English and English-Ukrainian Phraseological Dictionary" (Horot et al., 2024) and the "Dictionary of Phraseological Units of the Ukrainian Language" (Bilonozhenko et al., 2003) are included with the permission of the respective publishers and are provided strictly for academic research use. This portion of the corpus is *not* licensed under CC BY-SA 4.0. Users intending to use the Ukrainian portion for commercial purposes must contact the Condor Publishing House directly.
- Following the MIDAS release (Kim et al., 2025), any portion of the English-side content that originates from Wiktionary is governed by the CC BY-SA 4.0 license and inherits its requirements: attribution to the original Wiktionary entries, redistribution under the same license, and explicit indication of any modifications. This requirement applies even in academic redistribution.

We release the corpus and evaluation code to promote transparency and reproducibility.

Acknowledgements

We are grateful to the Condor Publishing House for granting permission to use the "Ukrainian-English and English-Ukrainian Phraseological Dictionary" for academic purposes.

References

Anthropic. 2025. Claude 4.5 Haiku.

V. M. Bilonozhenko, I. S. Hnatiuk, V. V. Diatchuk, N. M. Nerovnia, and T. O. Fedorenko. 2003. *Slovnyk frazeolohizmiv ukrainskoi movy [Dictionary of Phraseological Units of the Ukrainian Language]*. Naukova Dumka, Kyiv.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Sundesh Donthi, Maximilian Spencer, Om B. Patel, Joon Young Doh, Eid Rodan, Kevin Zhu, and Sean O'Brien. 2025. Improving LLM abilities in idiomatic translation. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 175–181, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Michael Flor, Xinyi Liu, and Anna Feldman. 2025. A survey of idiom datasets for psycholinguistic and computational research. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025)*.

Hui Gao, Jing Zhang, Peng Zhang, and Chang Yang. 2025. Consistency rating of semantic transparency: an evaluation method for metaphor competence in idiom understanding tasks. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10460–10471, Abu Dhabi, UAE. Association for Computational Linguistics.

Gemma Team, Google DeepMind. 2025. Gemma 3 technical report.

Google DeepMind. 2025. Gemini 2.5: Our most intelligent AI model.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Ye. I. Horot, Yu. V. Hromyk, L. K. Malimon, L. P. Pavlenko, A. B. Pavliuk, and O. O. Rohach. 2024. *Ukrainsko-anhliiskyi ta anhlo-ukrainskyi frazeolohichnyi slovnyk [Ukrainian-English and English-Ukrainian Phraseological Dictionary]*. Condor Publishing House, Kyiv, Ukraine.

Jisu Kim, Youngwoo Shin, Uiji Hwang, Jihun Choi, Richeng Xuan, and Taeuk Kim. 2025. Memorization or reasoning? exploring the idiom understanding of LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21678–21699, Suzhou, China. Association for Computational Linguistics.

Andrian Kravchenko, Yurii Paniv, and Nazarii Drushchak. 2025. UAlign: LLM alignment benchmark for the Ukrainian language. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 36–44, Vienna, Austria (online). Association for Computational Linguistics.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2023. Translate meanings, not just words: Idiomkb's role in optimizing idiomatic translation with language models.

Diego Moussallem, Mohamed Ahmed Sherif, Diego Esteves, Marcos Zampieri, and Axel-Cyrille Ngonga Ngomo. 2018. LIdioms: A multilingual linked idioms data set. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

Yurii Paniv. 2025. Isolating llm performance gains in pre-training versus instruction-tuning for mid-resource languages: The ukrainian benchmark study. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 876–883, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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

Qwen Team. 2025. Qwen3.

Nataliia Saichyshyna, Daniil Maksymenko, Oleksii Turuta, Andriy Yerokhin, Andrii Babii, and Olena Turuta. 2023. Extension Multi30K: Multimodal dataset for integrated vision and language research in Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 54–61, Dubrovnik, Croatia. Association for Computational Linguistics.

Alejandro R. Salamanca et al. 2026. Tiny aya: Bridging scale and multilingual depth.

Lapa LLM Team. 2025. Lapa LLM v0.1.2-instruct: An efficient Ukrainian open-source language model. https://huggingface.co/lapa-llm/lapa-v0.1.2-instruct. Model based on Gemma-3-12B, developed by researchers from Ukrainian Catholic University, AGH University of Krakow, Igor Sikorsky Kyiv Polytechnic Institute, and Lviv Polytechnic.

Debra A. Titone and Cynthia M. Connine. 1999. On the compositional and noncompositional nature of idiomatic expressions. *Journal of Pragmatics*, 31(12):1655–1674.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 text embeddings: A technical report.

Cai Yang, Yao Dou, David Heineman, Xiaofeng Wu, and Wei Xu. 2025. Evaluating llms on chinese idiom translation.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

A Allowed Linkages

Table 8 lists the source–field subset pairings used during corpus construction.

Left subset		Right subset
MIDAS (field: idiom)	↔	Condor EN–UK (field: idiom)
MIDAS (field: idiom)	↔	Condor UK–EN (field: translation)
Condor UK–EN (field: idiom)	↔	Bilonozhenko (field: idiom)
Condor EN–UK (field: idiom)	↔	Condor UK–EN (field: translation)

Table 8: Allowed linkages between sources and fields.

B Model Details

Table 9 provides details of the six evaluated models. All open-weight models were served using vLLM with temperature 0.3. API-based models used their respective default settings.

Model	Size	Access	Notes
Gemini 2.5 Flash	N/A	API	Google DeepMind
Claude Haiku 4.5	N/A	API	Anthropic
Gemma 3	12B	Open	Google DeepMind
Qwen 3	30B	Open	Alibaba Cloud
LapaLM	12B	Open	UA-adapted
Tiny Aya Global	3.35B	Open	Cohere Labs

Table 9: Summary of evaluated models.

C Translation Prompt Templates

For the **default** condition, models received only the source sentence and a translation instruction:

> *Translate the following text from [Ukrainian/English] to [English/Ukrainian]. Output only the translation, nothing else.*
> *Input:* `{text}`

For the **context-augmented** condition, the prompt additionally identified the source idiom by name. Note that no figurative meaning, candidate translation, or example from the corpus was provided — only the idiom string itself, sufficient to signal that a non-literal interpretation is required while keeping the prompt minimal:

> *Translate the following text from [Ukrainian/English] to [English/Ukrainian]. This text is an example usage of the idiom "`{idiom}`". Use this context to produce an accurate and natural translation. Output only the translation, nothing else.*
> *Input:* `{text}`

D Evaluation Prompt Template (UK→EN)

LLM Judge Prompt (Ukrainian → English)

Role. You are an expert in idiom translation, specializing in Ukrainian idioms, their figurative meanings, and natural English idiomatic equivalents.

Task. Evaluate whether the English translation preserves the Ukrainian idiom's figurative meaning and whether it uses a natural English idiom or fixed expression.

Test data: Ukrainian sentence: `{src}` | Idiom: `{idiom}` | Meaning: `{meaning}` | Translation: `{tgt}` | Golden equivalents: `{gld}`

Judging procedure.

1. Identify the segment in the translation corresponding to the Ukrainian idiom.
2. Determine the intended figurative meaning: (a) use idiom meaning if present; (b) else infer from golden equivalents; (c) else infer from context.
3. Decide whether the translation preserves that meaning.
4. If preserved, decide whether the rendering is idiomatic: (a) idiomatic means a conventional English idiom or widely recognized fixed expression used by native speakers; (b) a plain descriptive paraphrase is not idiomatic; (c) if golden equivalents are provided, com-

pare the phrase from the translation against them for idiomatic equivalence; (d) if golden equivalents are empty, judge by general native-speaker usage. Use golden equivalents as the strongest signal; do not penalize for wording differences, but penalize if clearly less idiomatic than the golden equivalents.

Scoring rubric. **1 pt:** Figurative meaning not preserved. **2 pts:** Meaning preserved but expressed non-idiomatically. **3 pts:** Meaning preserved and expressed idiomatically.

Output: JSON with keys: `score` (1–3), `mapped_phrase_in_tgt` (string), `meaning_preserved` (bool), `uses_english_idiom` (bool), `reasoning` (2–5 sentences).

E Evaluation Prompt Template (EN→UK)

LLM Judge Prompt (English → Ukrainian)

Role. You are an expert in idiom translation, specializing in English idioms, their figurative meanings, and natural Ukrainian idiomatic equivalents.

Task. Evaluate whether the Ukrainian translation preserves the English idiom's figurative meaning and whether it uses a natural Ukrainian idiom with the same figurative meaning.

Test data: English sentence: `{src}` | Idiom: `{idiom}` | Meaning: `{meaning}` | Translation: `{tgt}` | Golden example: `{gld}`

Definition. An idiom is a group of words established by usage as having a meaning not deducible from those of the individual words.

Judging procedure.

1. Determine the intended figurative meaning using the idiom meaning and the sentence context.
2. Identify the segment in the translation corresponding to the English idiom.
3. Decide whether the translation preserves that figurative meaning.
4. If preserved, decide whether the rendering is idiomatic: (a) a word-for-word structural copy of the English idiom (calque) is NOT idiomatic; (b) a plain descriptive paraphrase is NOT idiomatic; (c) the phrase must be a fixed, conventional expression that native Ukrainian speakers would naturally produce and recognize; (d) if a golden example is provided and the mapped phrase differs from it, treat this as a strong signal the rendering is NOT idiomatic.

Scoring rubric. **1 pt:** Figurative meaning not preserved; includes mistranslation, literal translation, calque, or unnatural phrasing. **2 pts:** Meaning preserved and translation reads naturally, but no established Ukrainian idiom is used. **3 pts:** Meaning preserved, translation reads naturally, and uses an established Ukrainian idiom that native speakers would recognize and use.

Output: JSON with keys: `score` (1–3), `mapped_phrase_in_tgt` (string), `meaning_preserved` (bool), `uses_ukrainian_idiom` (bool), `reasoning` (2–5 sentences).

F Meaning Preservation Rates

Table 10 shows meaning preservation rates. Gemini achieves the highest rate in EN→UK (93.7%), even exceeding its UK→EN rate, suggesting its EN→UK failures are predominantly literal translations rather than meaning distortions.

Model	UK→EN	EN→UK
Gemini	91.8%	93.7%
Claude	87.2%	86.0%
Gemma	85.1%	85.3%
LapaLM	81.6%	77.3%
Qwen	81.6%	71.0%
Tiny Aya	70.6%	61.2%

Table 10: Meaning preservation rates by model and direction.

G LapaLM vs. Gemma: Detailed Comparison

Table 11 compares LapaLM directly with its base model Gemma 3. Despite Ukrainian-specific adaptation, LapaLM underperforms on every metric.

	Gemma		LapaLM	
Metric	**UK→EN**	**EN→UK**	**UK→EN**	**EN→UK**
Mean score	2.49	2.09	2.34	1.98
% Perfect (3)	64.0	30.0	52.4	30.0
% Failure (1)	14.9	20.8	18.3	32.5
Meaning pres.	85.1	85.3	81.6	77.3
Δ_{ctx}	+0.233	+0.133	+0.033	+0.000

Table 11: LapaLM vs. Gemma 3 (its base model).

H Outlier Analysis by Direction

Table 12 shows the outlier effect by direction. Outlier clusters (higher semantic transparency) are easier across both directions, with slightly larger effects in EN→UK.

	UK→EN		EN→UK	
Model	**No out.**	**Out. only**	**No out.**	**Out. only**
Gemini	2.617	2.666	2.393	2.486
Claude	2.534	2.662	2.036	2.245
Gemma	2.441	2.604	2.056	2.263
LapaLM	2.249	2.478	1.905	2.188
Qwen	2.293	2.582	1.770	1.894
Tiny Aya	2.003	2.319	1.640	1.901

Table 12: Scores on non-outlier vs. outlier-only clusters, by direction (both prompts averaged).

UkrSL: Towards a Ukrainian Continuous Sign Language Dataset

Oleksandr Sobetskyi **Maryna Kosse** **Roman Kyslyi** **Angelina Savchenko**
Ukrainian Catholic University, Kyiv School of Economics, Kyiv Polytechnic Institute
lex.sobieski@gmail.com, mkosse@kse.org.ua,
rkyslyi@kse.org.ua, anhelina.savchenko@edu.kpi.ua

Abstract

We present UkrSL-Annot, an annotated dataset for Ukrainian Sign Language (USL) — one of the most underresourced sign languages in Europe. The dataset comprises 1,456 annotated clips (1,463 with cropped video segments) totalling approximately two hours of signing, sourced from six broadcast videos from Suspilne, Ukraine's public broadcaster. Each clip is annotated with a spoken Ukrainian transcription aligned to the corresponding signing segment. We describe the data collection pipeline, the annotation methodology, and provide a detailed analysis of the dataset's statistics and limitations. The dataset is being actively expanded, and we release this snapshot to support the research community and invite collaboration.

1 Introduction

Sign language recognition (SLR), sign language translation (SLT), and sign language production (SLP) have seen rapid progress in recent years, driven largely by the availability of annotated corpora for established sign languages such as German Sign Language (Koller et al., 2015), British Sign Language (Schembri et al., 2013), and American Sign Language (Duarte et al., 2021). Ukrainian Sign Language (USL), however, remains severely underresourced: no publicly available annotated continuous USL corpus exists to date, leaving Ukrainian Deaf and hard-of-hearing communities without the computational tools increasingly available to speakers of higher-resourced sign languages. Among these tasks, SLP — the automatic generation of sign language video from spoken or written language — is a natural fit for sentence-level aligned corpora such as ours, since it requires parallel data mapping text to signing and can be partially trained without gloss-level annotation, although high-quality SLP typically also benefits from fine-grained motion supervision.

The urgency of this gap has intensified in recent years. Ukrainian public discourse has expanded dramatically online, with sign language interpretation of news broadcasts, political addresses, and public information campaigns becoming widely disseminated on video platforms. These materials represent a valuable, naturally occurring source for dataset construction.

This paper introduces UkrSL-Annot, the first publicly released annotated corpus of continuous Ukrainian Sign Language. We describe the data collection and annotation pipeline, present an exploratory analysis, and discuss the challenges and ongoing efforts to expand coverage. Our contributions are:

1. A release of 1,456 annotated continuous USL clips ($\sim$2 hours of video) sourced from six publicly available broadcast videos;
2. A sentence-level annotation schema pairing signed segments with spoken Ukrainian transcriptions;
3. An open-source annotation tool and pipeline leveraging ASR for temporal alignment;
4. An analysis of dataset statistics, limitations, and a roadmap for community-driven expansion.

2 Related Work

Continuous sign language datasets have been central to the field since the release of the RWTH-PHOENIX-Weather 2014 corpus (Koller et al., 2015), which remains a primary benchmark for German SLR and SLT. Subsequent efforts include the BSL Corpus (Schembri et al., 2013), CSL-Daily for Chinese Sign Language (Zhou et al., 2021), and How2Sign for ASL (Duarte et al., 2021). These corpora range from a few hours to hundreds of hours of annotated video, and their availability has

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 53–57

directly enabled state-of-the-art transformer-based models.

For low-resource sign languages, the landscape is sparse. Yin et al. (2021) call on the NLP community to broaden its research agenda to signed languages, noting the strong asymmetry in available resources across sign languages. Existing work on Ukrainian NLP, such as the UberText corpus (Chaplynskyi, 2023) and recent shared tasks for Ukrainian LLMs (Syvokon et al., 2024), has focused on spoken and written Ukrainian, with no dedicated continuous sign language resources. The closest prior efforts for Ukrainian Sign Language are the linguistically oriented USL corpus collected by Bauer (2022) at the University of Cologne (∼3 hours of dialogues with Deaf Ukrainian refugees, intended for linguistic analysis rather than machine learning), and the rule-based machine translation work of Lozynska et al. (2019), which uses small internal sentence-level lexicons that are not released as a continuous video corpus.

Our work follows the philosophy of community-driven, incrementally growing resources used in earlier sign language corpus projects such as the NCSLGR corpus (Neidle et al., 2001) and the BSL/Auslan corpus tradition (Johnston, 2010): we treat the dataset as a living resource that grows with continuing annotation effort, rather than a one-shot release frozen at submission time.

3 Data Collection

Source videos were obtained from Suspilne (*UA:PBC*), Ukraine's public broadcasting company, which granted explicit permission for their use in this research. The videos consist of news broadcasts and public information programs featuring professional sign language interpreters, providing naturally occurring continuous USL in a controlled visual setting. Videos were selected to ensure sufficient temporal coverage and visual clarity (minimal occlusion, adequate framing of the signer's upper body and hands).

The current release draws from six source videos. The number of clips per video ranges from 18 to 508 (Table 1), with an imbalance that reflects the varying length of available source material rather than a balanced design. Addressing source diversity is a primary goal of the ongoing collection phase.

Video clips were cropped to isolate individual signing segments. All clips are encoded as H.264 MP4 files at 510×510 pixels and 30 fps, a resolution and frame rate sufficient for capturing hand and finger motion.

4 Annotation Methodology

Each source video contains both a signed and a spoken Ukrainian track. We use the Whisper large-v3 automatic speech recognition model (Radford et al., 2023) to obtain word-level transcriptions and their corresponding timestamps. The raw ASR output is then passed through a lightweight preprocessing step that merges consecutive words into sentence-level segments based on punctuation and pause heuristics. These sentence-level boundaries serve as candidate segmentation points for initial clip extraction.

Annotation tool. To support efficient and consistent annotation we developed a custom web-based annotation tool (Figure 1a), deployed on Hugging Face Spaces with Firebase Realtime Database as the backend. The interface embeds the source YouTube video alongside an editable caption table, providing playback controls (speed adjustment from $0.25\times$ to $2\times$, frame-level seeking with ± 100 ms and ± 1 s buttons) and per-caption editing fields for start time, end time, and subtitle text. Annotators can mark individual captions as *aligned* and flag entire videos as complete. The tool stores annotations in a hierarchical structure keyed by YouTube video ID, where each caption record contains the spoken text, start and end timestamps (in seconds), and an alignment flag.

Annotation workflow. For each video, an annotator: (1) reviews the Whisper-generated caption table alongside the video; (2) corrects any ASR transcription errors in the spoken Ukrainian text; (3) adjusts start and end timestamps so that each segment boundary aligns precisely with the onset and offset of the corresponding signing; and (4) marks each caption as aligned once satisfied. Annotation is carried out by fluent USL speakers from a broader pool of volunteers; two annotators contributed to the current release, compensated at 400 UAH per hour. Because our annotators are fluent signers with practical expertise in USL, we treat their annotations as reference-quality and do not measure inter-annotator agreement — a choice we revisit in the Limitations section.

Each resulting clip is paired with the spoken Ukrainian sentence as transcribed by Whisper and

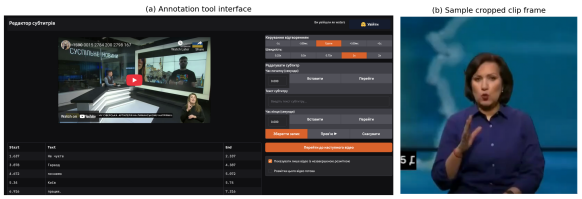


Figure 1: (a) The web-based annotation tool showing the embedded video player, playback controls, and editable caption table. (b) A sample cropped clip frame (510 × 510 px) showing a USL interpreter during a Suspilne broadcast.

Property	Value
Total annotated clips	1,456
Clips with video	1,463
Source videos	6
Annotators	2
Total video duration	∼2 h (119.8 min)
Mean clip duration	4.91 s ($\sigma = 3.65$)
Median clip duration	4.00 s
Min / Max duration	0.23 s / 36.30 s
Resolution	510 × 510 px
Frame rate	30 fps
Video codec	H.264
Total text tokens	14,514
Unique text tokens	5,876
Mean words per clip	10.0
Total file size	347.6 MB

Table 1: UkrSL-Annot dataset statistics (current release).

corrected by the annotator, representing the spoken-language transcript that accompanies the signing. The original audio track is preserved in the source videos but is not redistributed in the released clips, which contain only the cropped signer view; the spoken transcription is therefore stored as a separate textual layer aligned to the video by start and end timestamps. We note that broadcast signing may diverge from a verbatim rendering of the spoken source: in this release we adopt the spoken transcript as the reference text without explicitly flagging interpreter paraphrasing or summarisation, and we revisit this design choice in the Limitations section. Final annotations are exported to a pipe-delimited CSV with fields for clip identifier, spoken text, and annotator identity.

5 Dataset Analysis

Table 1 summarizes the key statistics of the current release. The dataset contains 1,456 annotated clips with a total video duration of approximately two hours. Clip durations are right-skewed (mean

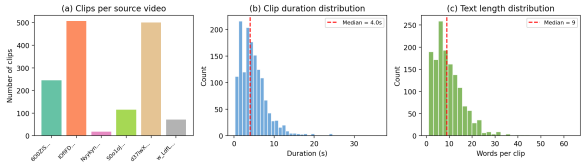


Figure 2: Dataset statistics. (a) Clip distribution across six source videos. (b) Clip duration histogram; red dashed line indicates the median (4.0 s). (c) Text length distribution (words per clip).

4.91 s, median 4.00 s), with a long tail of clips exceeding 15 seconds (Figure 2b).

The text vocabulary comprises 5,876 unique word types over 14,514 tokens. The average clip contains 10 words of spoken Ukrainian text, reflecting the sentence-level granularity of the segmentation. Figure 2c shows the distribution of text lengths across clips.

Source video coverage is uneven: the two largest source videos contribute over 1,000 clips combined, while the smallest contributes only 18 (Figure 2a). For downstream model training, we recommend treating source video identity as a stratification variable in any train/validation/test split to prevent models from exploiting video-specific visual cues rather than learning the sign language itself.

A cross-check between annotations and video files reveals that 1,463 cropped video segments are available, with 1,456 of these having corresponding annotations (99.5% coverage).

6 Ongoing Work

Active work to expand UkrSL-Annot includes: (1) expansion to additional source videos and a target of 5,000+ annotated clips; (2) addition of gloss-level annotation to enable sign language recognition tasks; (3) addition of signer identity metadata per clip, enabling signer-independent evaluation splits; (4) integration with pose estimation (MediaPipe Holistic) to extract skeletal keypoint features alongside raw video; (5) a planned train/dev/test split stratified by source video and signer.

We invite researchers and members of the Ukrainian Deaf community to contribute annotations, report errors, and suggest additional source material. The dataset and annotation guidelines are available at the repository.

7 Conclusion

We have presented UkrSL-Annot, the first publicly released annotated corpus of continuous Ukrainian

Sign Language, comprising approximately two hours of annotated video across 1,456 clips. The dataset provides sentence-level alignment between signed video segments and spoken Ukrainian transcriptions, supported by a reproducible collection and annotation pipeline. We hope this release will lower the barrier for the NLP community to engage with Ukrainian Sign Language and will serve as the foundation for a substantially larger resource in future releases.

Limitations

Although UkrSL-Annot has grown substantially, two hours of annotated video remains small compared to established benchmarks (e.g., PHOENIX-2014T contains $\sim$11 hours). The current annotation provides only sentence-level spoken text; gloss-level and sub-lexical annotations (handshape, movement, non-manual features) are not yet available, which limits applicability for phonological research and standard SLR evaluation. Source diversity is limited to six videos from a single broadcaster, which may introduce topical and stylistic biases. While we rely on fluent USL speakers for annotation quality, inter-annotator agreement has not been formally measured; future releases will include a doubly-annotated subset to quantify consistency. Because the reference text is the corrected spoken transcription rather than a gloss of what was actually signed, cases where the interpreter paraphrases, summarises or reorders the spoken source are not explicitly marked; downstream users training translation models should be aware that the alignment is text-to-signing and not gloss-to-signing. Finally, we have not yet established baseline model performance on this dataset.

Ethics Statement

All source videos were obtained with explicit permission from Suspilne (UA:PBC), Ukraine's public broadcasting company. Annotators are fluent USL speakers who are compensated for their work at a fair hourly rate. We are mindful of the ethical considerations surrounding sign language data, including the representation and agency of Deaf communities, and we commit to consulting with members of the Ukrainian Deaf community as the project expands. No personally identifiable information beyond publicly visible signer appearances in broadcast footage is included.

References

Anastasia Bauer. 2022. The Ukrainian Sign Language corpus. University of Cologne. https://ifl.phil-fak.uni-koeln.de/. Project page accessed April 2026.

Dmytro Chaplynskyi. 2023. Introducing UberText 2.0: A corpus of modern Ukrainian at scale. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 1–10.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2Sign: A large-scale multimodal dataset for continuous American Sign Language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2735–2744.

Trevor Johnston. 2010. From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics*, 15(1):106–131.

Oscar Koller, Jens Forster, and Hermann Ney. 2015. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125.

Olga Lozynska, Maksym Davydov, and Volodymyr Pasichnyk. 2019. Rule-based machine translation into Ukrainian Sign Language using concept dictionary. In *Proceedings of the International Workshop on Modern Machine Learning Technologies and Data Science (MoMLeT+DS)*.

Carol Neidle, Stan Sclaroff, and Vassilis Athitsos. 2001. SignStream: A tool for linguistic and computer vision research on visual-gestural language data. *Behavior Research Methods, Instruments, & Computers*, 33(3):311–320.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 28492–28518.

Adam Schembri, Jordan Fenlon, Ramas Rentelis, Sally Reynolds, and Kearsy Cormier. 2013. Building the British Sign Language corpus. *Language Documentation & Conservation*, 7:136–154.

Oleksiy Syvokon, Mariana Romanyshyn, and Roman Kyslyi. 2024. The UNLP 2024 shared task on fine-tuning large language models for Ukrainian. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING*.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. In

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP), pages 7347–7360.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1325.

Digitizing Old Ukrainian Texts: A Prompt-Based OCR Pipeline and Evaluation Dataset

Dmytro Chaplynskyi
I. Krypiakevych Institute
of Ukrainian Studies, lang-uk
chaplinsky.dmitry@gmail.com

Hanna Dydyk-Meush
I. Krypiakevych Institute
of Ukrainian Studies of NAS of Ukraine
hanna.dydykmeush@gmail.com

Abstract

We present a methodology and an open dataset for OCR of handwritten index cards containing a scholarly transcription of an early 17th-century Ukrainian polemical text, *Perestoroha* by Iov Boretskyi (Lviv, 1605–1606). The 430 cards, produced by 20th-century researchers, preserve the text in Old Ukrainian orthography with archaic diacritics, titlos, superscript letters, and ligatures that make automated recognition non-trivial. We develop a prompt-based OCR pipeline driven by a custom instruction set designed iteratively from the source material's orthographic conventions. The pipeline is evaluated against human-proofread ground truth in proprietary and open-source configurations using identical instructions and evaluation data. The proprietary configuration with extended thinking at maximum budget (Claude Opus 4.7, xhigh) achieves a Character Error Rate of **2.5%**; an Opus 4.6 baseline at the default 2,048-token thinking budget—used for the first batch of the released dataset—reaches **4.2%**; and two open-source Qwen3.6 variants running locally on consumer hardware reach **14.6%** (dense 27B) and **14.8%** (35B-A3B MoE). We release the fully digitized text aligned at line level to 300 DPI scanned images, as both a scholarly digital resource and training data for future OCR systems targeting Old Slavic manuscripts.[1]

1 Introduction

A significant body of Old Ukrainian linguistic material remains accessible only in handwritten form—either as original manuscripts or as scholarly transcriptions produced by researchers over the past century. Digitizing these materials is a prerequisite for computational analysis, full-text search, lexicography, and long-term preservation, yet the archaic orthographic conventions they employ render standard OCR tools ineffective.

We address this problem for a specific artifact: a set of 430 handwritten index cards constituting a scholarly transcription of *Perestoroha zelo potrebnaia na potomnye chasy pravoslavnym khrystiianom*, a polemical text by Iov Boretskyi written in Lviv in 1605–1606. The cards were produced by 20th-century researchers who carefully reproduced the original text, preserving its Old Ukrainian orthography, titlo abbreviation marks, superscript letters, and other diacritical features.

Rather than fine-tuning a dedicated handwritten text recognition model—which would require substantial annotated training data—we adopt a **prompt engineering** approach: we develop a detailed instruction set that guides a multimodal large language model to transcribe each card image into structured text while preserving all orthographic features. The instruction set was developed iteratively against Claude Opus 4.6 (Anthropic, 2025) through analysis of the source material's character inventory, abbreviation conventions, and diacritic usage patterns; we then re-evaluate the same instruction set, without modification, on Claude Opus 4.6 (production transcription baseline), Claude Opus 4.7 with extended thinking, and on two Qwen (Yang et al., 2025) variants as open-source replications.

Our contributions are: (1) a practical, reproducible prompt-based OCR pipeline for handwritten scholarly transcriptions of Old Ukrainian texts, evaluated against proprietary and fully open-source LLMs; and (2) an open dataset of 430 card transcriptions aligned to high-resolution scans.

2 Related Work

Historical HTR systems. Handwritten text recognition for historical documents has advanced substantially, driven by deep learning and large-

[1] Code: https://github.com/lang-uk/slavon_ocr/ Dataset: https://huggingface.co/datasets/lang-uk/perestoroha-ocr

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 58–66
May 29-30, 2026

scale digitization. Transkribus (Muehlberger et al., 2019) is the most widely adopted platform, achieving CERs below 5% across hundreds of public models. Kraken and eScriptorium (Kiessling, 2019; Kiessling et al., 2019) offer a fully open-source alternative designed for non-Latin and historical scripts. TrOCR (Li et al., 2023) introduced a fully Transformer-based architecture, achieving state-of-the-art results on handwriting benchmarks. CHURRO (Semnani et al., 2025), a 3B-parameter VLM fine-tuned on 155 historical corpora spanning 46 language clusters and 14 scripts, represents the current state of the art in VLM-based historical text recognition.

Old Slavic HTR. Work on Old Slavic manuscripts remains sparse compared to Western European languages. Neural HTR for Church Slavonic was pioneered via Transkribus, achieving CERs of 3–5% and identifying key error sources—superscript letters, titlo abbreviations, and word separation—that directly correspond to error categories in our pipeline (Rabus, 2019). The first generic HTR model for Ukrainian handwriting (CER 4.2%) was trained on 19th–20th century manuscripts (Tikhonov and Rabus, 2024); no existing model covers the 17th-century orthography targeted here. Recent work on HTR postprocessing of pre-modern Slavic texts identifies the same challenges with superscript letters and titlos that we encounter (Lendvai et al., 2024). Critically, all existing approaches rely on Transkribus or Kraken models requiring annotated training data.

Multimodal LLMs for OCR. Recent work shows that multimodal LLMs can match or exceed specialized HTR systems. On 18th–19th century English documents, GPT-4o, Claude, and Gemini achieve CERs of 5.7–7% (Humphries et al., 2025), and systematic comparisons find that LLMs significantly outperform all conventional methods on historical handwritten records (Kim et al., 2025). A multilingual HTR benchmark shows Claude 3.5 Sonnet outperforms other models in zero-shot settings but with weaker non-English performance (Crosilla et al., 2025). For Slavic text specifically, an evaluation of 12 multimodal LLMs on 18th-century Cyrillic finds that models exhibit "over-historicization"—inserting archaic characters from incorrect periods (Levchenko, 2025).

Prompt engineering for documents. Layout-aware prompting improves document understanding by 263% in zero-shot settings (Wang et al., 2023). Comparisons of prompting strategies for historical handwriting find that detailed document descriptions with few-shot examples yield optimal results (Kim et al., 2025). Structured prompt engineering with JSON output schemas enables effective extraction from specialized text without fine-tuning (Dagdelen et al., 2024). In low-resource settings, where annotated training data for HTR is unavailable, recent benchmarks of LLM-based OCR on under-served scripts (Sohail et al., 2024) report that careful prompt design closes a substantial part of the gap to specialized systems—a finding that motivates the instruction-set engineering approach we adopt for 17th-century Ukrainian.

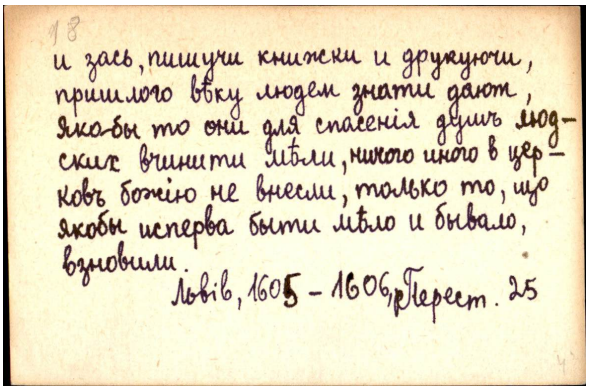


Figure 1: Example index card from the *Perestoroha* collection. The handwriting preserves 17th-century orthographic conventions. The bottom line shows the source reference (Lviv, 1605–1606, Perest. 25).

3 The Perestoroha Image Dataset

The source material consists of 430 index cards, scanned at 300 DPI, all belonging to a single scholarly artifact. Each card reproduces a fragment of *Perestoroha* with source attribution noting library provenance, approximate date, and folio references. The cards were handwritten by 20th-century researchers, making the handwriting itself relatively legible; the difficulty lies in the orthographic system being reproduced.

The text employs Old Ukrainian orthography of the 17th century. Key features include: ѣ (yat), ѡ (omega as ot-ligature), ї (yi with two dots), ѫ (big yus); **titlo** marks over abbreviated sacra nomina; superscript letters indicating abbreviation expansions; acute stress marks on vowels; and the coexistence of є and е, ꙋ and у—visually similar pairs requiring per-instance discrimination rather than any default substitution rule.

4 OCR Pipeline

4.1 Overview

The pipeline consists of five stages: (1) splitting scanned PDF files into individual card images; (2) an EXIF-orientation preprocessing pass that bakes any rotation flag into the pixel buffer (§4.5); (3) instruction-driven multimodal LLM transcription of each image into structured JSON; (4) import into a web-based proofreading editor for expert review; and (5) export of corrected transcriptions into a browsable dataset. Each card is transcribed independently—no context carries over between cards—ensuring reproducibility and enabling parallel processing. The code is available at `https://github.com/lang-uk/slavon_ocr/`.

4.2 The Instruction Set

The core of the pipeline is a custom instruction set for Claude Opus 4.6 (Anthropic, 2025), developed iteratively over multiple rounds of analysis and error correction. Its design drew on several sources:

- **Three ground-truth cards** with matched manual transcriptions, which revealed the scholar's encoding conventions: parentheses for expanded abbreviations, specific Unicode codepoints for diacritical marks, and character-level encoding decisions.
- **A 12,000-word reference corpus** of the same text (already digitized[2]), which yielded a full character inventory, uncovering characters initially missed—Latin *s* for zelo, the positional distribution of ꙋ vs у, and a complete inventory of sacra nomina with titlos.
- **Lecture materials** on Old Ukrainian phonology and orthography, which contributed precise descriptions of rarely encountered marks: paieryk, kamora, and prydykh.

The instruction set specifies a detailed character encoding table with Unicode codepoints, rules for diacritic handling, the convention for representing abbreviation expansions, and a structured JSON output schema. Key components are provided in Appendix A.

[2]The digitized portion of Кройніка used here as a reference corpus results from two phases of philological work. In 1981, the linguists Valentyna Cherniak and Oleksandra Zakharkiv selectively transcribed the monument for the Card Index of the *Dictionary of the Ukrainian Language of the 15th – first half of the 17th centuries*. In 2013–2014, Hanna Dydyk-Meush and Oksana Shpyt worked with the original of Кройніки and transcribed an extended portion in full; that latter transcription is what we use here.

4.3 Anti-Hallucination Measures

Early experiments revealed a critical failure mode: the model substitutes familiar words for unfamiliar archaic ones when letter shapes are ambiguous. For example, Спудєй was misread as Судєй, and испервa as неперва—plausible modern forms replacing rare historical ones.

We adopted a three-pronged mitigation strategy: (1) explicit "read letter by letter, not word by word" instructions; (2) a verification step listing concrete examples of documented error types (є/е confusion, ꙋ/у substitution, spurious trailing ъ); and (3) a "no defaults" policy—both members of each visually similar pair (є/е, ꙋ/у) are flagged as requiring per-instance verification against the card image.

4.4 Thinking Spiral Mitigation

With extended thinking (chain-of-thought reasoning) enabled at large reasoning budgets on Claude Opus 4.6, we observed a failure mode in which the model entered an unbounded deliberation loop on a card with many ambiguous characters, consuming its entire output token budget without producing any transcription. This occurred at both 4K and 32K thinking budgets. We did *not* observe the issue on *Perestoroha*—all 152 cards in the proofread corpus completed via the thinking-on path on both the Opus 4.6 production run (default 2,048-token thinking budget) and the Opus 4.7 evaluation (xhigh budget)—but it surfaced on a different handwritten source we work with, and colleagues working with other handwritten Slavic sources have reported the same pattern privately, suggesting it is not corpus-specific.

The root cause is that extended thinking expands to fill whatever budget is available, and the instruction set's emphasis on precision amplifies this tendency. Our pipeline therefore applies a budget-aware retry strategy: each card is first transcribed with extended thinking enabled at the configured budget; if the call exceeds the budget without producing a transcription, the card is retried with extended thinking disabled. On *Perestoroha* this fallback is never invoked, but it degrades gracefully on the small number of cards prone to deliberation loops in other corpora. A prompt-level mitigation—a "deliberation discipline" section instructing the model to make one pass, verify once, and commit—proved helpful but insufficient on its own.

4.5 EXIF Orientation Preprocessing

A non-trivial fraction of the scans—roughly 25% (over 100 files)—carried EXIF rotation flags whose target orientation differed from the underlying pixel buffer. Multimodal models honor these flags at decode time, so the model was shown the image rotated 90° or 180° from the writer's intended orientation. On the worst-affected cards this was catastrophic: a single card hit a CER of approximately 137%, with the model producing essentially gibberish in an attempt to read inverted handwriting. We added a one-time prepass that reads the EXIF tag, applies the implied rotation to the pixel data itself, and strips the metadata; both Claude and Qwen runs consume the resulting auto-oriented image directory. After the fix, the previously affected cards dropped to in-distribution CERs in the 3–5% range. In an ablation against the proofread evaluation corpus on the Opus 4.6 production configuration, removing the prepass raised aggregate CER from 4.16% to approximately 5.3% (+1.1 pp) and WER by several points, because the long tail of mis-rotated cards contributes disproportionately to both metrics: each misread compounds with re-segmentation noise across subsequent lines.

4.6 Proofreading Editor

We built a custom web-based proofreading editor (Flask + SQLite) using Claude Code. The editor displays the scanned card image alongside the model's transcription in a side-by-side view (Figure 2), with a toolbar for inserting Old Ukrainian diacritics and archaic characters, enabling a domain expert to correct errors character by character. Corrected transcriptions are exported as parallel JSON files, preserving the original OCR output for comparison. Although each card's initial draft was produced by Claude, every character in the resulting ground truth was verified against the source image by the annotator; the proofread reference is therefore independent of the model's lexical choices, even where the draft seeded the typing.

Proofreading was performed by a co-author with domain expertise in historical Ukrainian linguistics. The annotator's familiarity with the source material's orthographic system was sufficient to resolve ambiguities without formal adjudication guidelines.

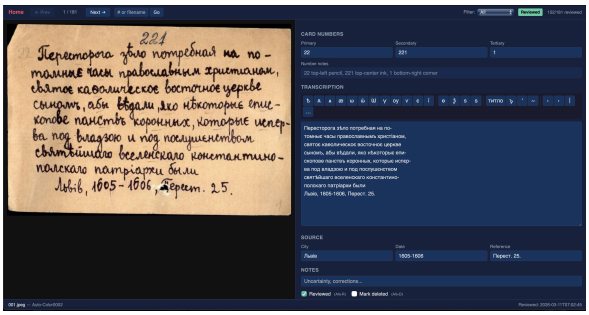


Figure 2: The proofreading editor showing a card image (left) alongside the model's transcription (right), with fields for card numbers, source metadata, and a diacritics toolbar.

5 Evaluation

5.1 Experimental Setup

We evaluate the pipeline in three configurations using identical instructions and evaluation data:

- **Claude Opus 4.7 with extended thinking** (Anthropic, 2025)—proprietary, API access, maximum reasoning budget (`xhigh`)—our headline configuration, applied to the proofread evaluation set.
- **Claude Opus 4.6** (Anthropic, 2025)—proprietary, API access, extended thinking enabled at the default 2,048-token budget—the production configuration used for the first batch of cards in the released dataset, reported here as a same-prompt baseline against which to measure the impact of the model upgrade.
- **Two open-source Qwen3.6 variants** (Yang et al., 2025), run locally on a single NVIDIA RTX 4090 GPU via `llama.cpp`: a dense **Qwen3.6-27B** model and a **Qwen3.6-35B-A3B** mixture-of-experts model with 3B active parameters per token. Both use 4-bit Q4_K_M GGUF quantizations from the Unsloth distribution.[3] The same instruction set was applied to both without modification.

5.2 Metrics

We report Character Error Rate (CER) and Word Error Rate (WER). CER is the primary metric, as word boundaries in this orthographic system are not always unambiguous, and many errors involve single-character substitutions within diacritical marks. Tokenization for WER uses the spaCy

[3] https://huggingface.co/unsloth/Qwen3.6-27B-GGUF and https://huggingface.co/unsloth/Qwen3.6-35B-A3B-GGUF.

Configuration	CER	WER
Claude Opus 4.7 (xhigh thinking)	2.51%	12.16%
Claude Opus 4.6 (default 2K thinking)	4.16%	18.10%
Qwen3.6-27B (dense, local)	14.56%	39.44%
Qwen3.6-35B-A3B (MoE, local)	14.84%	40.34%

Table 1: OCR accuracy on the proofread 152-card evaluation set, identical instruction set across configurations. WER tokenisation uses spaCy `uk_core_news_sm`; whitespace and punctuation tokens are dropped.

`uk_core_news_sm` pipeline; whitespace and punctuation tokens are dropped before the edit-distance computation, and hyphenated line-break splits are not rejoined.

5.3 Results

Results are shown in Table 1. Claude Opus 4.7 with extended thinking achieves 2.51% CER (542 character errors over 21,635 reference characters; 432 word errors over 3,552 word tokens), comparable to Transkribus-based results on Old Slavic texts (Rabus, 2019; Tikhonov and Rabus, 2024). Holding the instruction set fixed, switching from Opus 4.6 at the default 2K thinking budget (our production configuration) to Opus 4.7 with xhigh thinking reduces total character-level edits from 901 to 542—a 40% relative drop—and lifts the share of perfect cards (CER = 0) from 11.2% to **12.5%**. The two Qwen variants land within a hair of each other at 14.56% / 14.84% CER—roughly 6× higher than Opus 4.7—with the smaller dense 27B narrowly outperforming the larger MoE variant on every aggregate metric. A small tail of failure-mode cards drives most of the gap to Claude (§5.4).

An additional reference point: GPT-5.4 via Codex. As a sanity check on whether the proprietary advantage was Claude-specific, we ran the same instruction set through OpenAI's Codex CLI driving GPT-5.4 at the highest reasoning-effort setting on the 30-card test split. It reached a CER of 5.13% and WER of 24.08%, against 2.88% / 13.96% for Opus 4.7 (xhigh thinking) and 3.39% / 15.34% for Opus 4.6 on the same split, while taking roughly an order of magnitude longer per card in wallclock time than our Opus 4.6 production harness (we did not directly time-match it against the Opus 4.7 evaluation harness, which uses different parallelism). Although competitive with Opus 4.6 in absolute accuracy terms, GPT-5.4 offered no accuracy advantage on this task at substantially higher inference latency, so we did not pursue a full-corpus evaluation of this configuration.

5.4 Error Analysis

Character-level edit operations for Claude Opus 4.7 (xhigh thinking) on the 152-card proofread corpus comprise 542 total edits over 21,635 reference characters (376 substitutions, 63 insertions, 103 deletions). For comparison, the same instruction set on Opus 4.6 at the default 2K thinking budget produces 901 edits (550 / 193 / 158)—a 40% relative reduction at corpus scale, distributed unevenly across error classes. The most frequent residual error categories on Opus 4.7 are:

1. є/е **confusion** (U+0454 ↔ U+0435)—159 substitutions in both directions (є→е: 122; е→є: 37), accounting for 42% of all substitutions on Opus 4.7. The absolute count barely moves between Opus 4.6 (164) and 4.7 (159), while every other error class shrinks substantially—a strong indicator that this confusion reflects genuine visual ambiguity in the scholar's handwriting rather than a model failure mode that further prompt engineering will reduce.
2. **Spurious** ъ **insertion**—20 inserted hard signs, still the single largest non-whitespace insertion category, but down from 125 on Opus 4.6 (∼6× reduction). The model continues to over-apply the convention of word-final ъ from Church Slavonic templates, but only on a small residual set of cases.
3. ү/у **normalization** (U+04AF → U+0443)—27 cases where the rare straight-у variant is flattened to the dominant у, consistent with bias toward the more frequent letter form.
4. і/и **substitution** (U+0456 → U+0438)—13 cases where dotted і is read as plain и; plain и is roughly 4× more common in *Perestoroha* than і.
5. ѣ **(yat) misreadings**—10 substitutions spread across seven distinct target characters, reflecting the visual similarity of yat to several other letters when the writer's stroke is rushed.

Qwen-specific patterns. Both Qwen variants exhibit the same є/е confusion as Claude at substantially higher rates—173 substitutions on the 27B (є↔е sum), versus 159 on Opus 4.7—and a much stronger pull toward word-final ъ: 165 spurious insertions on the 27B and 161 on the 35B-A3B,

against just 20 on Opus 4.7 (∼8× more). They also show a modern-Ukrainian-bias pattern almost absent from the Claude error profile, including о→а substitutions (32 / 35 cases for 27B / 35B-A3B) and ы→и (21 / 39 cases for 27B / 35B-A3B). Severe whole-word hallucinations—the model generating plausible-looking Old Ukrainian text rather than transcribing what is on the card—are concentrated on a small tail of cards: per-card CER exceeds 50% on only 4/152 cards for the 27B and 3/152 for the 35B-A3B (roughly 2–3% of the evaluation). When those outlier cards are excluded, both Qwen variants produce CERs of approximately **8.8%**, suggesting that the aggregate gap to Claude is driven less by systematic recognition error than by a small number of failure-mode cards. The 35B-A3B's MoE architecture (3B active per token) produces marginally more deletions and fewer fabrications than the dense 27B, but the two are otherwise functionally equivalent on aggregate metrics. This comparison may not be entirely fair: the instruction set was developed for Claude, and both Qwen configurations used aggressively quantized GGUF builds. Prompt tuning specifically for the open-source models, or using less quantized variants, could plausibly close the residual gap further.

6 Dataset

We release the fully digitized text of *Perestoroha* as transcribed from 430 index cards.[4] The first batch of approximately 150 cards was transcribed with Opus 4.6 using an earlier version of the instruction set; the remainder was transcribed with Opus 4.7 (xhigh thinking) using the final instruction set described in §A. The dataset consists of JSON files paired with 300 DPI card images, aligned at line level. Each JSON record contains card numbers (primary, secondary, tertiary where present), an array of transcribed lines preserving exact line breaks for image-text alignment, source metadata (provenance, date, folio), and free-text notes flagging uncertainty.

The dataset serves two purposes: (1) as a **historical linguistic resource**—a searchable, citable digital text of a significant early 17th-century Ukrainian source that is otherwise difficult to access; and (2) as **OCR training and evaluation data** for future systems targeting Old Slavic Cyrillic handwriting.

[4]Dataset: https://huggingface.co/datasets/lang-uk/perestoroha-ocr

7 Discussion

Prompt engineering as a lightweight alternative to fine-tuning. The instruction set—approximately 3,000 words of encoding rules and error mitigation—required no training data beyond a few reference cards and a character inventory corpus. This makes the approach accessible to digital humanities practitioners who lack the resources for model fine-tuning.

Open-source viability. The two Qwen variants achieve CERs of 14.56% (dense 27B) and 14.84% (35B-A3B MoE) using the same instruction set without modification, with the smaller dense 27B narrowly outperforming the larger MoE variant on every aggregate metric—a useful finding for institutions that cannot use proprietary APIs and prefer a smaller, simpler deployment. While the aggregate numbers are still too high for unsupervised digitization, when the small tail of severely-hallucinated cards is excluded (∼2–3% of the evaluation) both variants produce CERs of approximately 8.8%, suggesting that the gap to Claude is driven primarily by failure-mode cards rather than by systematic recognition errors. This makes the open-source configuration a viable first pass for reducing manual transcription effort, and the released dataset is now large enough to support targeted fine-tuning aimed at suppressing the hallucination tail.

Broader implications. Ukrainian archives hold substantial collections of handwritten linguistic material from the 16th–18th centuries. The prompt-based approach demonstrated here can be adapted to other artifacts by developing source-specific instruction sets, potentially enabling large-scale digitization without the overhead of training dedicated models for each orthographic convention.

8 Conclusion

We presented a prompt-based OCR pipeline for digitizing handwritten scholarly index cards containing a 17th-century Ukrainian text. The pipeline, driven by a carefully engineered instruction set, achieves a CER of 2.5% with Claude Opus 4.7 (extended thinking), 4.2% with the Opus 4.6 production configuration, and 14.6–14.8% with two open-source Qwen3.6 variants run locally—improving to ∼8.8% once a small tail of failure-mode cards is excluded. We release the fully digitized text of *Perestoroha*—430 cards aligned to scanned

images—as an open resource for Ukrainian historical linguistics and OCR research.

Future work includes extending the pipeline to additional artifacts with different orthographic conventions (experiments are underway and show promising results), fine-tuning an open-source model on the released dataset, and building a searchable lexicographic resource from the digitized text.

Limitations

The pipeline was developed and evaluated on a single artifact with a specific orthographic system. Generalization to other Old Ukrainian texts—particularly those from different centuries or with different transcription conventions—requires partial re-engineering of the instruction set. The ground truth was produced by a single domain expert, and inter-annotator agreement has not been measured. The instruction set itself was iteratively refined—using Claude Code as a development assistant—against a 32-card sample drawn from the dev split during development on Opus 4.6, in order to reduce CER and WER on recurring failure modes; the 30-card test split was held out throughout. The full-corpus Opus 4.6 numbers therefore mix held-out and in-distribution performance, and readers wanting a pure held-out estimate should consult the test-split numbers in §5.3. The Opus 4.7 result is computed by running the same unchanged prompt on a newer model that the prompt was *not* tuned against, so the 4.6 → 4.7 delta partly reflects this looser coupling. The dataset is split deterministically into a 122-card development split and a 30-card test split (seed 20250101); the per-split numbers used for the GPT-5.4 comparison are reported separately to support cleaner held-out comparisons in future work. The proprietary configuration (Claude Opus) incurs API costs that may be prohibitive for large-scale digitization; the open-source alternative (Qwen) offers a cost-free option but at substantially lower accuracy, with a small tail of whole-word hallucinations and stronger biases toward Church Slavonic and modern-Ukrainian normalization that limit its utility without expert proofreading.

Ethical Considerations

This work involves digitization of historical scholarly materials in the public domain. No personal or sensitive data is processed. We used AI-based writing assistance tools (Claude) in preparing this manuscript, as well as Claude Code for developing the OCR pipeline and the proofreading interface. The open-source release of the dataset and instruction set is intended to enable reproducibility and further research.

Acknowledgments

We thank the staff of the I. Krypiakevych Institute of Ukrainian Studies of the National Academy of Sciences of Ukraine for access to the *Perestoroha* card collection and for support of this work. We thank Oksana Shpyt for her work on the philological transcription of Кройніка that produced our digitized reference corpus, Yurii Paniv for assistance with the Qwen experiments, and Mariana Romanyshyn for reviewing early drafts of this manuscript.

References

Anthropic. 2025. Claude Opus 4 system card. https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-opus-4-system-card.pdf.

Giorgia Crosilla, Lukas Klic, and Giovanni Colavizza. 2025. Benchmarking large language models for handwritten text recognition. *Journal of Documentation*, 81(7):334–354.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15:1418.

Mark Humphries, Lianne C. Leddy, Quinn Downton, Meredith Legace, John McConnell, Isabella Murray, and Elizabeth Spence. 2025. Unlocking the archives: Using large language models to transcribe handwritten historical documents. *Historical Methods*, 58(3):175–193.

Benjamin Kiessling. 2019. Kraken — a universal text recognizer for the humanities. In *Digital Humanities 2019 Book of Abstracts*, Utrecht.

Benjamin Kiessling, Robin Tissot, Peter A. Stokes, and Daniel Stökl Ben Ezra. 2019. eScriptorium: An open source platform for historical document analysis. In *2019 ICDAR Workshops*, pages 19–24.

Seorin Kim, Julien Baudru, Wouter Ryckbosch, Hugues Bersini, and Vincent Ginis. 2025. Early evidence of how LLMs outperform traditional systems on OCR/HTR tasks for historical records. *arXiv preprint arXiv:2501.11623*.

Piroska Lendvai, Maarten van Gompel, Anna Jouravel, Elena Renje, Uwe Reichel, Achim Rabus, and Eckhart Arnold. 2024. A workflow for HTR-postprocessing, labeling and classifying diachronic and regional variation in pre-modern Slavic texts. In *Proceedings of LREC-COLING 2024*, pages 2039–2048. ELRA and ICCL.

Maria Levchenko. 2025. Evaluating LLMs for historical document OCR: A methodological framework for digital humanities. In *Proceedings of LM4DH 2025*, pages 75–85.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. TrOCR: Transformer-based optical character recognition with pre-trained models. In *Proceedings of AAAI 2023*, volume 37, pages 13094–13102.

Guenter Muehlberger, Louise Seaward, Melissa Terras, Sofia Ares Oliveira, Vicente Bosch, Maximilian Bryan, Sebastian Colutto, Hervé Déjean, Markus Diem, Stefan Fiel, Basilis Gatos, Albert Greinoecker, Tobias Grüning, Guenter Hackl, Vili Haukkovaara, Gerhard Heyer, Lauri Hirvonen, Tobias Hodel, Matti Jokinen, and 35 others. 2019. Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*, 75(5):954–976.

Achim Rabus. 2019. Recognizing handwritten text in Slavic manuscripts: A neural-network approach using Transkribus. *Scripta & e-Scripta*, 19:9–32.

Sina J. Semnani, Han Zhang, Xinyan He, Merve Tekgürler, and Monica S. Lam. 2025. CHURRO: Making history readable with an open-weight large vision-language model for high-accuracy, low-cost historical text recognition. In *Proceedings of EMNLP 2025*.

Muhammad Atif Sohail, Sarfaraz Masood, and Hammad Iqbal. 2024. Deciphering the underserved: Benchmarking LLM OCR for low-resource scripts. *arXiv preprint arXiv:2412.16119*.

Alexey Tikhonov and Achim Rabus. 2024. Handwritten text recognition of Ukrainian manuscripts in the 21st century: Possibilities, challenges, and the future of the first generic AI-based model. *Kyiv-Mohyla Humanities Journal*, 11:226–247.

Wenjin Wang, Yunhao Li, Yixin Ou, and Yin Zhang. 2023. Layout and task aware instruction prompt for zero-shot document image question answering. *arXiv preprint arXiv:2306.00526*.

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

A Instruction Set Components

We provide the three key components of the OCR instruction set: the character encoding table (§A.1), the verification step (§A.2), and the JSON output schema (§A.3). The full instruction set is approximately 3,000 words; we include the components most critical for reproducibility.

A.1 Character Encoding Table (excerpt)

The instruction set specifies exact Unicode codepoints and encoding conventions for each archaic character and diacritical mark:

```
ABBREVIATIONS AND EXPANSIONS
- Letters expanded from titla or
  superscript go in PARENTHESES:
  e.g., v"shy(t)ko, ye(d)no
- Titlo mark U+0483 goes INSIDE
  parentheses for abbreviated
  letters: (l"), (ch"), (ts")
- Titlo on sacra nomina stays
  OUTSIDE parentheses on the base
  letter: g"" for gospod'

ARCHAIC CHARACTERS (never modernize)
- omega (U+03C9): ot-ligature as
  omega(t); plain: omega zmiiyu
- yat (U+0463): virimo, zviri
- big yus (U+046B): tu, zovu(t")
- little yus (U+0467): sya, movyachi
- Latin s for zelo: distinct from
  z and dze
- dze (U+0455): distinct from z
  and Latin s

VOWEL LETTERS
- ye (U+0454) and e (U+0435) are
  BOTH valid, often coexist on one
  card. Do NOT default to either.
  Check EVERY instance against card.
- u (U+04AF) and u (U+0443) are
  BOTH valid. Transcribe as written.
```

A.2 Verification Step

After initial transcription, the instruction set requires a line-by-line comparison against the source image, checking for six specific error types:

```
VERIFY BEFORE SAVING:
Re-read the image and compare against
your transcription line by line.
Check for:
- INSERTED LETTERS: characters added
  that are not on the card
- SUBSTITUTED WORDS: a familiar word
  replacing an unfamiliar one
- ye/e CONFUSION: writing ye where
  the card shows e, or vice versa.
  Check EVERY instance.
- ADDED WORD ENDINGS: adding hard
  sign or other characters where
  the card does not have them
- MODERNIZED SPELLING: archaic
  letters silently replaced with
  modern equivalents
```

```
- MISSING CHARACTERS: letters on
  the card that were skipped
```

A.3 JSON Output Schema

Each card is transcribed into a structured JSON record:

```
{
  "filename": "Scan0042.jpg",
  "card_numbers": {
    "primary": "18",
    "secondary": null,
    "tertiary": null,
    "notes": "18 top-left ink"
  },
  "lines": [
    "line 1 of transcription",
    "line 2 of transcription",
    "..."
  ],
  "source": {
    "city": "Lviv",
    "date": "early 17th c.",
    "reference": "Kron. 3 rev."
  },
  "notes": ""
}
```

The `lines` array preserves exact line breaks from the card, enabling alignment between image lines and transcribed text. The `source` object captures provenance metadata written at the bottom of each card. Card numbers track multiple numbering systems (primary/secondary/tertiary) reflecting successive cataloguing efforts.

Semantic Fidelity Versus Literary Quality: A Construct Validity Study of Neural Machine Translation Metrics

Dmytro Chaplynskyi[1,2,3] **Maria Shvedova**[4,5] **Ivan Kulynych**[6] **Lesia Ivashkevych**[7]

[1]Ukrainian Catholic University [2]I. Krypiakevych Institute of Ukrainian Studies [3]lang-uk
[4]NTU "Kharkiv Polytechnic Institute" [5]University of Jena
[6]Grammarly [7]NTUU "Igor Sikorsky Kyiv Polytechnic Institute"

chaplynskyi.dmytro@ucu.edu.ua, corpus.textiv@gmail.com,
ivankulynych4@gmail.com, lesia.ivashkevych@gmail.com

Abstract

Automatic machine translation metrics are the de facto standard for evaluating translation quality. Yet, it remains unclear what they actually measure. We investigate this question using a unique multilingual corpus: seven human Ukrainian translations of George Orwell's *Animal Farm*, alongside three architecturally distinct AI systems (GPT-5.2, DeepL, and Lapa, a Ukrainian-tuned LLM). Across seven neural metrics, four reference-free and three reference-based, all three AI translations rank at the top. However, stylometric analysis exposes that these same AI translations are not as lexically rich as human ones (-18% MTLD), underuse Ukrainian particles (up to $2\times$ fewer) and diminutive morphology ($2.6\times$ fewer), and converge on near-identical outputs (LaBSE pairwise similarity 0.941 vs. 0.711 for human pairs). A controlled LLM-as-a-judge experiment demonstrates a clear preference reversal: when the English source is visible, AI ranks first; when it is hidden and the judge evaluates literary quality alone, humans rise to the top and AI falls to the lower ranks. Human evaluation (1,034 pairwise judgments) is balanced across both patterns. We argue that current MT metrics reward semantic fidelity and surface fluency — properties optimized by AI systems — while failing to capture the lexical richness, cultural adaptation, and stylistic voice that characterize skilled literary translation.

1 Introduction

Neural MT metrics — COMET, COMETKiwi, XCOMET, MetricX — are trained on human judgments from news and general-domain settings. They reward closeness to the source and surface fluency. Whether this transfers to literary translation, where voice, style, and cultural adaptation matter, is an open question since literary translation differs structurally from metric training regimes. The quality of literary translation depends largely on how the text reads in the target language — on lexical richness, pragmatic nuance, rhythm, and culturally specific expression. A translation can be semantically faithful yet stylistically thin.

To test construct validity, we derive four falsifiable predictions from the hypothesis that neural metrics primarily reward semantic fidelity:

1. **Metric dominance.** Systems optimized for fidelity — including modern AI systems — should rank highest across neural metrics.

2. **Convergence.** If metrics reward proximity to the source, metric-preferred systems should be both closer to the English original and closer to one another in semantic space.

3. **Stylistic compression.** Features associated with Ukrainian literary expressiveness — lexical diversity, particles, diminutive morphology, and stylistic dispersion — should not correlate with metric rankings and may be reduced in metric-preferred systems.

4. **Preference reversal.** If fidelity drives metric scores, then removing access to the source during evaluation should alter system rankings. When evaluators judge only the Ukrainian text as literary prose, systems optimized for fidelity should lose their advantage.

We test these predictions using seven neural MT metrics (four reference-free and three reference-based), cross-lingual embedding similarity (LaBSE), multi-dimensional stylometric analysis, and two controlled LLM-as-a-judge experiments that differ only in source visibility. We further compare these results to human pairwise evaluation aggregated with TrueSkill.

2 Related Work

Automatic MT evaluation has shifted decisively from lexical-overlap metrics such as BLEU toward

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 67–79

neural approaches — including COMET (Rei et al., 2020), COMETKiwi (Rei et al., 2022), XCOMET (Guerreiro et al., 2024), and MetricX-24 (Juraska et al., 2024) — which correlate much more strongly with human adequacy judgments (Freitag et al., 2022). This progress, however, raises a largely overlooked question: what exactly do these metrics measure, and is correlation with adequacy the right criterion for evaluating literary translation? Zouhar et al. (2024) show that COMET is biased toward adequacy and penalizes valid paraphrases, while Läubli et al. (2018) demonstrate that human-parity claims dissolve at document-level evaluation.

For example, in general-domain English–Ukrainian parallel data, Chaplynskyi and Zakharov (2025) found that an ensemble of six Quality Estimation (QE) models explains only ∼60% of the variance in human quality judgments, with a nonlinear relationship between metric scores and human perception, suggesting that the gap widens further for literary text.

Prior research shows that neural systems produce fluent output but struggle with stylistic consistency, figurative language, and cultural nuance (Toral and Way, 2018; Wang et al., 2023; Karpinska and Iyyer, 2023) — precisely the dimensions that define literary quality. Vanmassenhove et al. (2019) show that NMT reduces lexical and morphological richness compared to human translation, and Zhang et al. (2025) find that automatic metrics consistently prefer machine-generated literary translations over professional translators. Existing work on literary MT either remains qualitative or applies general-domain metrics without questioning their validity for literary texts.

Stylometry offers a long-standing tradition of analyzing these dimensions. Features such as function-word frequencies, lexical diversity, and morphological patterns have been used to identify translator voice and stylistic distinctiveness (Burrows, 2002; Rybicki, 2012; Baker, 2000). Yet this tradition has not been meaningfully integrated into MT evaluation research. Zheng et al. (2023) show that GPT-4 matches both controlled and crowdsourced human preferences, and LLMs have been established as state-of-the-art MT evaluators (Kocmi and Federmann, 2023). However, Huang et al. (2024) find that source information can be counterproductive for LLM-based evaluation — a finding our source-visibility experiment directly extends. We bring these strands into dialogue.

This paper sits at the intersection of MT evaluation, literary translation, and stylometric analysis. Unlike prior work that evaluates literary MT using standard metrics, we ask whether those metrics validly measure literary quality. Unlike qualitative critiques of MT in literary contexts, we provide quantitative evidence across multiple independent dimensions.

ID	Year	Translator / System	Type
T1	1947	Ivan Cherniatynskyi	Human
T2	1984	Iryna Dybko	Human*
T3	1991	Oleksii Drozdovskyi	Human
T4	1991	Yurii Shevchuk	Human
T5	1992	Natalia Okolitenko	Human†
T6	2020	Bohdana Nosenok	Human
T7	2021	Viacheslav Stelmakh	Human
T8	—	Lapa (v0.1.2-instruct)	AI (tuned LLM)
T9	—	GPT-5.2	AI (general LLM)
T10	—	DeepL	AI (commercial NMT)

Table 1: List of Ukrainian translations of George Orwell's *Animal Farm* included in the corpus. *Free cultural adaptation. †Translated from Russian, not English.

3 Corpus

To test whether neural MT metrics capture literary translation quality, we need multiple translations of the same source, enabling comparison of translational strategies independent of source variation. Our corpus comprises ten Ukrainian translations of George Orwell's *Animal Farm* (seven human, three AI) aligned into 1,367 sentence-level segments.[1] The seven human translations are drawn from the ParaFarm corpus (Maslij (Kalashnyk) and Shvedova, 2025; Kalashnyk, 2025). All translate the identical English source, so observed differences reflect strategy rather than content.

The human translations span 1947–2021, representing diaspora (Cherniatynskyi, Dybko), early-independence (Drozdovskyi, Shevchuk, Okolitenko), and contemporary professional (Nosenok, Stelmakh) contexts, which provides natural variation in norms and conventions.

We include Dybko (1984) as a deliberate test. Because it departs substantially from the English source, we expect fidelity-oriented metrics to penalize it. We therefore retain it in full metric analyses but exclude it from certain group-level comparisons between human and AI systems.

We generate three AI translations representing distinct architectures and training regimes:

[1] Sentence boundaries do not always coincide across translations; alignment follows the source segmentation.

- **GPT-5.2**[2] — a general-purpose large language model prompted for translation (see Appendix D).
- **DeepL**[3] — a commercial neural machine translation system (used via API without additional prompting).
- **Lapa (v0.1.2-instruct)** (Paniv et al., 2025) — a 12B-parameter LLM (Gemma-3) adapted for Ukrainian, state-of-the-art on English–Ukrainian translation benchmarks, and fine-tuned on multimodal data including the filtered English–Ukrainian parallel corpus from Chaplynskyi and Zakharov (2025)

The three AI systems differ in architecture and training, so convergent stylistic patterns would point to properties of AI translation in general rather than to any single model.

The corpus has three methodological advantages: one controlled source, seven human baselines spanning 74 years, and natural variation in translation norms. The retranslation literature (Deane-Cox, 2014; Stasiuk, 2019; Lepokhin, 2023) suggests that successive translators and translation editors may deliberately diverge from predecessors, which would further increase human–human variation independently of AI convergence. The result is a rare setting for disentangling semantic fidelity, stylistic range, and metric bias.

4 Methodology

Our experimental design tests the four predictions outlined in Section 1, with each analytical component targeting a distinct aspect of the construct-validity hypothesis.

To evaluate metric dominance (Prediction 1), we assess ten Ukrainian translations using seven state-of-the-art neural MT metrics spanning reference-free and reference-based paradigms. The reference-free metrics (COMETKiwi-22, COMETKiwi-XL, XCOMET-XXL, and MetricX-24 QE) estimate translation quality using only the source sentence and candidate translation. The reference-based metrics (COMET-22, XCOMET, and MetricX-24) are implemented in a round-robin design in which each translation is scored against the other nine as pseudo-references, yielding a 10×10 pairwise matrix. For each system, we report the mean score across all nine references. Higher scores indicate better quality for COMET-family metrics, whereas lower scores indicate better quality for MetricX.

If neural metrics operationalize semantic fidelity and fluency, AI systems optimized for fidelity should rank highest.

To test convergence and source proximity (Prediction 2), we compute cross-lingual semantic similarity using LaBSE (Feng et al., 2022). For each pair of systems, aligned segments are embedded, and cosine similarity is computed per segment, with mean similarity reported. We calculate system–system similarity (AI–AI, human–human, AI–human) as well as system–source similarity between each translation and the English original.

If neural metrics reward source proximity, AI systems should exhibit higher similarity to the source, tighter clustering in semantic space, and greater separation from human translations.

To assess stylistic compression (Prediction 3), we conduct a multi-dimensional stylometric analysis capturing features of Ukrainian literary expressiveness not reducible to semantic fidelity. Lexical diversity is measured using Measure of Textual Lexical Diversity (MTLD), Moving Average Type Token Ratio (MATTR), and hapax ratio (via lexical richness and pymorphy3[4]). We select two morphosyntactic features identified by Ukrainian linguists as salient markers of literary expressiveness. Discourse particles (e.g., ж ‘indeed’, таки ‘after all’, ось ‘here/just’, бо ‘because’, аж ‘even’, ну ‘well’, мов/наче ‘as if’) have no direct English equivalents and must be actively introduced by the translator; Shevelov (1963) describes them as “the microorganisms of language, which lend it colour and flavour,” making their density a marker of literary voice. Diminutive morphology — suffixes such as -еньк-, -очк-, -ик, -оньк-, -ечк- signaling affection, intimacy, or attenuation — encodes evaluative, emotional, and pragmatic functions beyond literal smallness (Ruda, 2021), requiring the translator to perceive the emotional register of the scene. We quantify particle frequency and diversity through lemma matching and detect diminutive morphology using regex over morphologically analyzed tokens. Surface overlap between translations is measured with pairwise chrF and BLEU scores computed with sacrebleu,[5] and stylistic distance is estimated using Cosine Delta, a variant of Burrows’

[2] https://openai.com/index/gpt-5/
[3] https://www.deepl.com/
[4] https://github.com/no-plagiarism/pymorphy3
[5] https://github.com/mjpost/sacrebleu

Delta (Evert et al., 2017) based on z-score normalized function-word frequencies. We also compute the standard deviation of per-segment Ukrainian-to-English word-count ratios to assess consistency in expansion and compression. If neural metrics fail to capture stylistic richness, top-ranked systems may exhibit reduced lexical diversity, lower particle and diminutive usage, and lower stylistic dispersion.

To test preference reversal under controlled source visibility (Prediction 4), we conduct two LLM-as-a-judge experiments using GPT-5.2 with identical sampling and aggregation procedures. Note that GPT-5.2 serves a dual role: it is both one of the evaluated translation systems (T9) and the judge. We discuss the resulting self-preference risk in the Limitations section.

- *Translation-focused:* the model evaluates pairs with access to the English source, mirroring the human evaluation setup.

- *Literary-focused:* the model evaluates the same pairs without access to the source and is instructed to judge solely on literary naturalness, expressiveness, and stylistic richness in Ukrainian.

Human evaluation is conducted as a pairwise preference tournament following Romanyshyn et al. (2024). Four professional English–Ukrainian translators serve as annotators. They are shown the English source and two anonymized Ukrainian translations and select the better translation or indicate a tie. Judgments emphasize meaning preservation alongside fluency and literary quality. Because evaluators have access to the source text, this procedure reflects adequacy-oriented judgment under source visibility, comparable to the conditions under which neural metrics are trained.

All three evaluations — both LLM-as-a-judge experiments and the human evaluation — draw from the same pool: 1,128 valid segments $\times$ 45 system pairs $=$ 50,760 items, globally shuffled with a fixed seed. This ensures that even early subsets cover all 45 system pairs approximately uniformly. Left/right presentation order is randomized per pair to control for position bias. TrueSkill ratings (Herbrich et al., 2006) are computed with default parameters ($\mu_0 = 25$, $\sigma_0 = 25/3$).

If neural metrics primarily reward fidelity to the source, removing source visibility should systematically alter system rankings, providing causal evidence of construct misalignment.

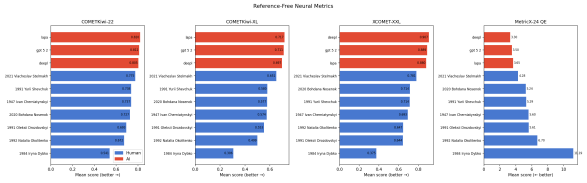


Figure 1: Comparison of reference-free neural metrics across the systems.

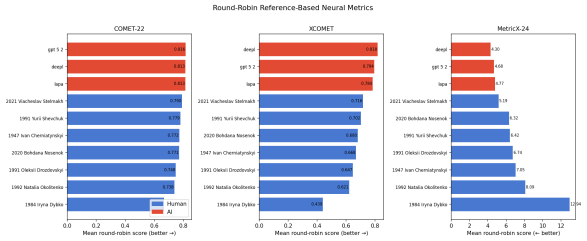


Figure 2: Round-robin comparison across the systems.

5 Results

We present results corresponding to the four predictions made in Section 1.

5.1 Prediction 1: Metric Dominance

Across every MT metric, a consistent pattern emerges: all three AI systems rank in the top three. No metric ranks any human translator above any AI system. The highest-ranked human translation (Stelmakh, 2021) consistently places fourth.

For example:

- On COMETKiwi-22, AI systems average 0.812 versus 0.724 for human translators (excluding Dybko), a gap of +0.088.

- On MetricX-24 QE (lower is better), AI systems average 3.48 compared to 5.45 for humans.

The pattern holds across COMETKiwi-XL, XCOMET-XXL, COMET-22, and round-robin MetricX-24. Figure 1 (reference-free metrics) and Figure 2 (round-robin metrics) show consistent separation between AI and humans.

As expected, Dybko's free cultural adaptation ranks last on every metric, confirming that the metrics are highly sensitive to source deviation. Notably, Lapa — a 12B-parameter model adapted from Gemma-3 — tops the COMETKiwi rankings, matching or exceeding GPT-5.2, a much larger general-purpose model. On metric scores, a

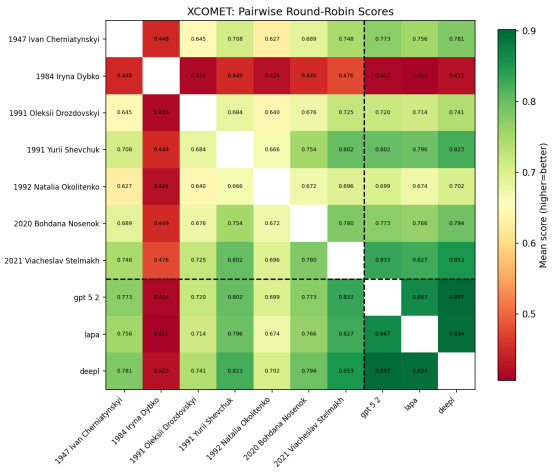


Figure 3: Heatmap of pairwise XCOMET round-robin scores (note the bottom-right AI block).

Measure	H–H	AI–AI	Gap
LaBSE cosine sim.	0.711	**0.941**	+0.230
XCOMET round-robin	0.627	**0.886**	+0.259
COMET-22 round-robin	0.750	**0.881**	+0.131
MetricX-24 round-robin	7.61	**3.38**	−4.23
chrF (surface overlap)	33.6	**43.1**	+9.5

Table 2: AI–AI vs. Human–Human average pairwise scores. For MetricX-24, lower values indicate greater similarity. Bold signifies the more similar group.

domain-tuned small model can reach parity with a system orders of magnitude larger.

These results confirm Prediction 1: neural MT metrics systematically favor AI translations in this literary corpus. Not only that, the round-robin heatmaps reveal a uniformly high-scoring 3×3 AI–AI block, while the 7×7 human–human block shows much more variation.

5.2 Prediction 2: Convergence

We next examine pairwise similarity among translations. The AI–AI LaBSE similarity averages 0.941, compared to 0.711 for human–human pairs — a +0.230 gap.

This agreement is consistent across independent measures:

The individual AI–AI LaBSE pairs — Lapa–DeepL (0.952), GPT-5.2–DeepL (0.940), Lapa–GPT-5.2 (0.932) — represent near-identical translations. The convergence holds on every measure, from neural embeddings to raw character n-grams. Three architecturally distinct systems have arrived at the same output.

AI systems are also measurably closer to the

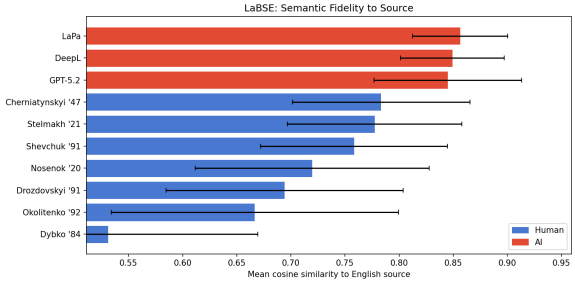


Figure 4: LaBSE semantic similarity to English source.

English source than any human translator in cross-lingual embedding space. On LaBSE, all three AI systems exceed 0.845 similarity to the source; the closest human (Cherniatynskyi, 1947) reaches 0.783. The human average (excluding Dybko) is 0.733. This gap may reflect AI systems selecting the most frequent translation equivalents, while human translators employ richer or less default lexical choices that increase semantic distance from the source.

Surface overlap (chrF/BLEU) confirms this trend. AI translations are 15–20% more similar to each other at the surface level (chrF) than any pair of human translations. The three AI systems form a tight cluster; humans spread across a much wider range — a gap that may be amplified by a deliberate "repulsion effect," as later translators often read and consciously diverge from their predecessors (Deane-Cox, 2014).

Despite architectural differences, a general-purpose LLM (GPT-5.2), a commercial NMT system (DeepL), and a domain-tuned LLM (Lapa) produce near-identical translations. The convergence persists from neural metrics to surface-level chrF scores.

This confirms Prediction 2: systems that neural metrics favor are both closest to the source and highly clustered in semantic space.

5.3 Prediction 3: Stylistic Compression

At the same time, AI translations systematically lack Ukrainian literary expressiveness. AI systems average an MTLD of 311 vs. 377 for humans (excluding Dybko) — an 18% gap. The hapax ratio (words used exactly once) tells the same story: AI 0.182 vs. human 0.215, a 15% deficit. AI translations cycle through a narrower vocabulary, concentrating 39.4% of their text within the 100 most common words (vs. 37.7% for humans).

Ukrainian particles (ж 'indeed', таки 'after all',

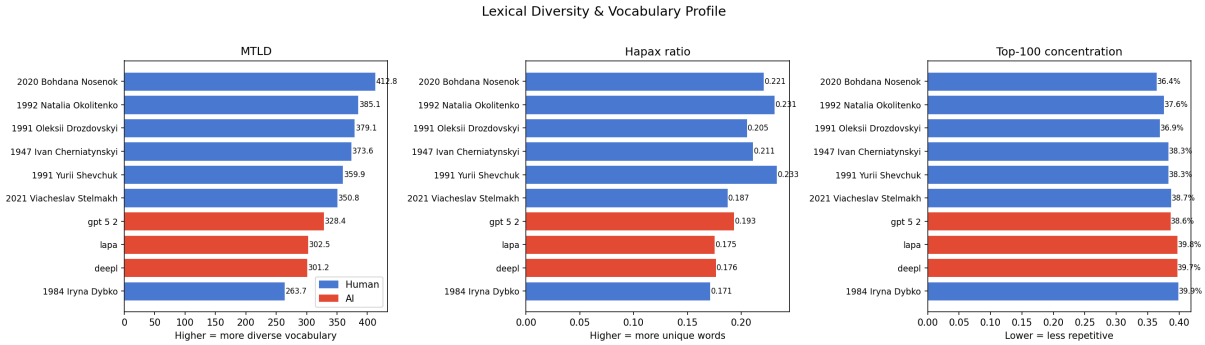


Figure 5: Lexical diversity (MTLD, Hapax, Top-100).

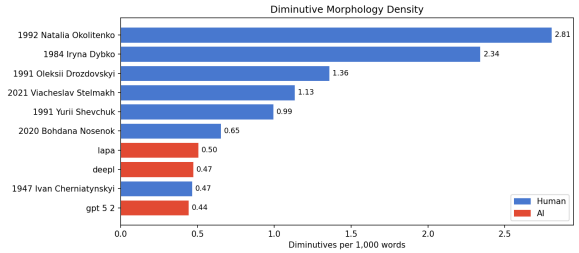


Figure 6: Diminutive morphology across translators and AI systems.

ось ‘here’, бо ‘because’, аж ‘even’, ну ‘well’, мов/наче ‘as if’) encode pragmatic nuances — emphasis, surprise, hedging — with no direct English equivalents. They must be *added* by the translator. GPT-5.2 and DeepL produce approximately 2× fewer particles than the human average.

Diminutive suffixes (-еньк-, -очк-, -ик, -оньк-; e.g., мама ‘mom’ → мамочка ‘mommy’) are a core expressive device in Ukrainian, signaling affection, irony, or intimacy. Linguists studying Ukrainian diminutives note that the category often encodes much more than literal smallness, including evaluative, emotional, or pragmatic functions in discourse (Ruda, 2021).

AI systems average 0.47 diminutives per thousand tokens; humans average 1.23 — a 2.6× gap. All three AI systems rank at the bottom, below every human translator. Lapa, despite literary fine-tuning, is indistinguishable from GPT-5.2 and DeepL on this measure.

Using function-word lemma frequencies — content-independent markers of translatorial voice — Cosine Delta measures how stylistically distinct each system is from the others. AI systems have the smallest mean pairwise distance (DeepL: 1.058, GPT-5.2: 1.059, Lapa: 1.068), placing them closest to each other and to the corpus centroid. Human translators range from 1.095 (Nosenok) to 1.160 (Dybko). By this measure, AI systems occupy the same narrow corner of the stylistic space.

Lapa ($\sigma = 0.127$) and DeepL ($\sigma = 0.132$) are the two most uniform systems, maintaining near-constant word count ratios across segments. Human translators vary more — they expand descriptive passages and compress dialogue, adapting to content. AI produces uniformly adequate output without the peaks and valleys that characterize human stylistic choices.

These results confirm Prediction 3: the systems that neural metrics rank highest exhibit reduced lexical diversity, fewer discourse particles and diminutives, lower stylistic dispersion, and more uniform expansion ratios — a systematic pattern of stylistic compression.

5.4 Prediction 4: Preference Reversal

In order to validate these results, we ran two LLM-as-a-judge experiments (∼750 pairwise comparisons each, with rankings stable between the 500- and 750-pair checkpoints) and a human evaluation (1,034 judgments) with an identical setup except for one variable: the presence of the English source.

- **LLM-as-a-judge Translation (Source Visible):** Judge sees the English source + two Ukrainian translations.
- **LLM-as-a-judge Literary (Source Hidden):** Judge sees only two Ukrainian sentences.
- **Human eval:** same setup as experiment 1.

We compute TrueSkill ratings for each experiment and compare them with metric rankings and human evaluations (1,034 matches). The results are in Table 3; the ranking shifts for AI systems are in Table 4.

System	Metrics	LLM: Translation	LLM: Literary	Human Eval
Lapa	**#1**	#4 (27.4)	#9 (22.4)	#7 (24.9)
GPT-5.2	**#2**	**#1** (30.3)	#6 (24.4)	#5 (25.3)
DeepL	**#3**	**#3** (29.2)	#8 (23.8)	**#2** (26.4)
Stelmakh 2021	#4	**#2** (29.8)	**#2** (28.6)	**#1** (27.1)
Shevchuk 1991	#5	#5 (26.7)	**#3** (26.6)	**#3** (26.3)
Cherniatynskyi 1947	#6	#6 (25.0)	#7 (24.3)	#8 (24.1)
Nosenok 2020	#7	#8 (23.5)	#5 (24.7)	#4 (25.3)
Drozdovskyi 1991	#8	#7 (24.0)	**#1** (29.5)	#6 (25.2)
Okolitenko 1992	#9	#9 (21.4)	#4 (25.2)	#9 (23.3)
Dybko 1984	#10	#10 (13.9)	#10 (21.4)	#10 (18.1)

Table 3: System rankings across four evaluation paradigms, ordered by metric rank. TrueSkill μ in parentheses. Bold = top 3 in each column.

AI system	Metric	Transl.	Literary	Shift
GPT-5.2	#2	#1	#6	↓5
DeepL	#3	#3	#8	↓5
Lapa	#1	#4	#9	↓5

Table 4: AI rank shifts when the English source is removed from evaluation.

This reversal is uniform across all AI systems and persists even when the judging model is held constant. The main experimental manipulation here is the presence or absence of the source sentence. **The top five positions in the literary ranking are all human translators.**

The most dramatic individual reversal is that of Drozdovskyi (1991). This translation ranks #8 on COMETKiwi-22 (0.693) and #7 on the translation judge — firmly in the bottom half. But on the literary judge, it leaps to #1 ($\mu = 29.5$), surpassing every system, including Stelmakh. The translation is rich in particles (10.1/1k — the highest in the corpus) and diminutives (1.36/1k). These are exactly the features metrics penalize and literary judgment rewards.

Okolitenko (1992), who translated from Russian rather than English, shows a similar pattern: #9 on metrics and human evaluation, but #4 on the source-free literary judge. The low fidelity scores are expected given the different source language, yet the Ukrainian prose is judged favorably when evaluated on its own terms.

Stelmakh (2021) is the only translation to rank in the top two across all non-metric rankings: #1 in human eval (27.1), #2 in literary judge (28.6), #2 in translation judge (29.8). On metrics, it ranks #4 — the best human. Stelmakh represents the rare translator whose work satisfies both fidelity-oriented and literary-oriented evaluation: semantically faithful enough for metrics, stylistically rich enough for readers.

The 1,034-match human evaluation places Stelmakh clearly first ($\mu = 27.1$), but positions #2–#7 form a compressed cluster ($\mu = 24.9$–26.4) with overlapping confidence intervals. DeepL ranks #2 and Shevchuk #3 — both above GPT-5.2 (#5) and Lapa (#7). Humans value fidelity enough to keep DeepL near the top, but not enough to replicate the metric-based ranking in which all three AI systems dominate.

This suggests that human judges — even when given access to the source — weigh stylistic and expressive qualities more heavily than neural metrics do. In other words, human assessment appears to balance semantic fidelity with literary naturalness, whereas neural metrics disproportionately reward the former. The human results thus reinforce the central claim of construct misalignment: the metrics that identify what is "best" do not fully correspond to what readers perceive as high-quality literary translation.

These results confirm Prediction 4: evaluation criteria shift when the source is removed, demonstrating that metric-aligned fidelity is not equivalent to literary quality.

6 Discussion

Our results suggest a clear answer: neural MT metrics primarily measure semantic fidelity to the source and surface fluency in the target language. These are the properties AI systems maximize (via RLHF, parallel training data, and instruction tuning), and these are the properties on which AI systems achieve the highest scores.

This is not a flaw in the metrics per se — semantic fidelity and fluency are genuine dimensions of translation quality. The problem is one of construct validity: the metrics claim to measure "translation

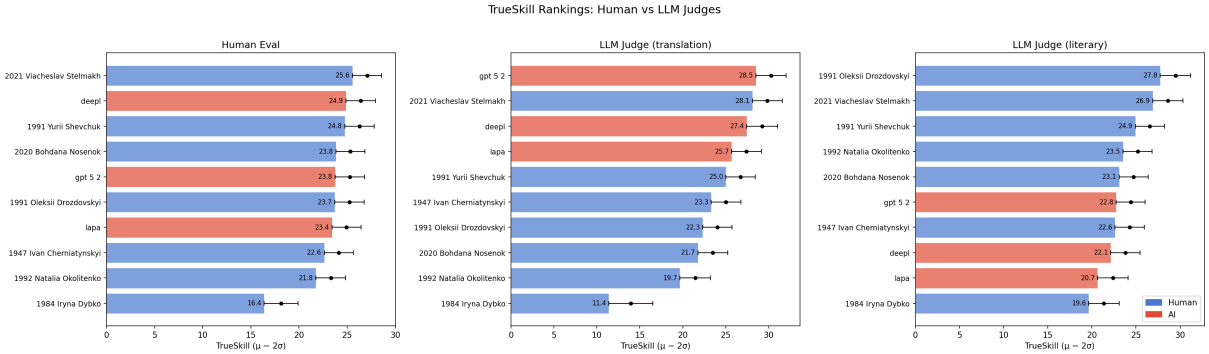


Figure 7: TrueSkill rankings across three evaluation paradigms.

quality" but actually measure a subset of quality that is systematically correlated with AI output. In the literary domain, this subset is insufficient.

The preference reversal is the strongest evidence for this claim. When the source is visible, the judge sees the same information the metrics do and produces the same ranking. When the source is hidden, the judge must evaluate the Ukrainian text on its own merits — and the ranking inverts. **The quality that survives without source access is a different quality from what metrics capture.**

Perhaps the most striking finding is the convergence of three architecturally distinct AI systems. GPT-5.2, DeepL, and Lapa produce translations with pairwise LaBSE similarity of 0.941. By every measure (neural metrics, surface overlap, cross-lingual embeddings, cosine delta), they are closer to each other than any pair of human translators is to each other.

This has implications beyond evaluation. If AI translation converges on a single point in the space of possible translations, then:

1. Switching between AI systems may not always increase diversity. While individual differences exist, the overall stylometric profiles of the three AI systems we tested are far more similar to each other than to any human translator. Whether this convergence generalizes to a broader population of MT systems is an open question requiring larger-scale replication.

2. Domain tuning closes the metric gap but not the stylistic one. Lapa (12B parameters) matches the much larger GPT-5.2 on metric scores — a strong result for a compact, domain-adapted model. Yet on stylometric measures, Lapa is indistinguishable from GPT-5.2 and DeepL: fine-tuning achieves metric parity without producing a distinct literary voice.

The stylometric deficits we document — in particles, diminutives, and lexical diversity — are not arbitrary choices but features that Ukrainian linguists identify as markers of skilled prose (see Section 4). Their absence in AI translations is not a matter of personal preference but of cultural competence: AI systems translate *what is said* but not *how it is said*.

These findings have practical consequences. For MT evaluation research, current metrics should not be applied to literary translation without explicit caveats. A literary-specific evaluation metric is needed — one that rewards vocabulary richness, cultural perceptiveness, and stylistic identity alongside semantic fidelity. For AI-assisted translation, AI output may serve as a useful starting point — human evaluators ranked DeepL second overall — but the stylistic deficits documented here suggest that post-editing for lexical and cultural richness remains necessary; across the three systems we tested, switching between AI providers did not yield greater stylistic diversity, though wider replication is needed before generalizing.

6.1 Future Directions

One avenue for improvement is prompt design. The LLM-based translations (GPT-5.2 and Lapa) were generated with the minimal translation prompt used by the Lapa team in their evaluation pipeline (Paniv et al., 2025) — adopted here for direct comparability with their reported benchmark scores — without specific instructions targeting stylistic features such as particles or diminutives. Because large language models are sensitive to instruction framing, prompts explicitly emphasizing lexical variation, pragmatic particles, and affective nuance may reduce stylistic compression without requiring fine-

tuning. Controlled comparisons between general and feature-targeted prompts would clarify this possibility.

A second direction concerns evaluation. Literary-aware metrics could incorporate stylometric dispersion, lexical richness, and discourse-pragmatic features alongside semantic fidelity. Hybrid systems combining adequacy metrics with source-free literary judgment may better reflect the multidimensional nature of translation quality.

The particles and diminutives analyzed here are salient examples that Ukrainian linguists have already highlighted, but the space of expressive markers is likely much larger. Preliminary analysis of augmentative and pejorative suffixes (-ищ(е), -иськ(о), -юг(а), -юр(а)) showed zero occurrences in all AI translations versus sparse but non-zero usage by humans — a pattern consistent with our findings, though the signal in this corpus is too weak for statistical claims. Systematic identification of further morphosyntactic markers of literary expressiveness is a promising direction for future work.

Finally, replication across genres and language pairs, particularly those with different morphological and expressive resources, would determine whether stylistic compression is language-specific or structural to current AI systems.

7 Conclusion

Across seven neural metrics, three AI translations consistently rank at the top. Yet these same translations are not as lexically rich as human ones, lack common signs of Ukrainian literary voice, and converge on near-identical output. When an LLM-as-a-judge evaluates the translations without access to the source — judging only whether the Ukrainian text reads as skilled literary prose — the AI advantage disappears, and human translators take the top five positions.

Although the metrics accurately measure what they were trained to measure — semantic fidelity to the source and surface fluency in the target — literary translation requires more than fidelity and fluency. It requires voice, cultural adaptation, and expressive richness — qualities that current metrics cannot detect and that AI systems do not produce.

Limitations

The study is limited to a single novella and a single language pair (English→Ukrainian), restricting generalization. The number of systems (seven human, three AI) constrains statistical breadth, though TrueSkill mitigates uncertainty. The seven human translations span 74 years (1947–2021) and reflect shifts in Ukrainian orthographic and stylistic norms across diaspora, Soviet-era, and post-independence periods; this temporal range is part of what makes the corpus interesting, but it also means within-human variation conflates translator voice with diachronic norm change. We note, however, that within-human pairwise LaBSE similarity (0.711) remains substantially lower than within-AI similarity (0.941), so temporal and norm-driven variation in the human set does not approach the magnitude of the AI–human gap that drives our central findings. The 1,034-match human evaluation is directionally stable, but the #2–#7 cluster is tight with overlapping confidence intervals. GPT-5.2 serves as both a translation system and the LLM judge, creating a potential self-preference bias; however, GPT-5.2 ranks only #6 in the literary condition and #5 in human evaluation, suggesting that any self-preference does not dominate the results. Replication with an architecturally distinct judge model (e.g., Claude or Gemini) would further strengthen this conclusion and is a clear next step. Our stylometric analysis targets specific morphosyntactic features (particles, diminutives, lexical diversity) but does not measure figurative language or cultural adaptation, which are also central to literary quality. Finally, sentence-level pairwise judgments cannot capture long-range narrative qualities such as sustained voice or rhythm, and our human evaluation was conducted only under source-visible conditions; a source-hidden human pairwise evaluation, mirroring the LLM literary judge, is the natural next experiment to fully decouple the source-removal effect from the human/LLM-judge contrast.

Ethical Considerations

This study uses published literary translations and publicly available AI systems. Human annotators were professional translators who participated in the study on a voluntary basis. No personally identifiable information was collected. The AI translations were generated for research purposes. AI writing assistance (Claude, Anthropic) was used

for editing and formatting the manuscript. We acknowledge that our findings about AI translation limitations should not be used to devalue human translators' work but rather to highlight the irreplaceable qualities they bring to literary translation.

Acknowledgments

We thank Vladislav Demyanov, Kateryna Buchina, and Serhiy Snihur for serving as expert translators in the human evaluation, and Taras Yaroshko for his contributions during the early stages of this research.

References

Mona Baker. 2000. Towards a methodology for investigating the style of a literary translator. *Target*, 12(2):241–266.

John Burrows. 2002. 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287.

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

Sharon Deane-Cox. 2014. *Retranslation: Translation, Literature and Reinterpretation*. Bloomsbury Academic.

Stefan Evert, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt. 2017. Understanding and explaining delta measures for authorship attribution. *Digital Scholarship in the Humanities*, 32(suppl_2):ii4–ii16.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of ACL 2022*.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU — neural metrics are better and more robust. In *Proceedings of WMT 2022*, pages 46–68.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. TrueSkill: A Bayesian skill rating system. In *Advances in Neural Information Processing Systems 19*.

Xu Huang, Zhirui Zhang, Xiang Geng, Yichao Du, Jiajun Chen, and Shujian Huang. 2024. Lost in the source language: How large language models evaluate the quality of machine translation. In *Findings of ACL 2024*, pages 3546–3562.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation (WMT 2024)*, pages 492–504.

Viktoriia Kalashnyk. 2025. Creation of a multi-variant parallel corpus of Ukrainian translations of George Orwell's "Animal Farm" and its use for studying variability of the Ukrainian language. In *Language Space of the Modern World: Proceedings of the IX All-Ukrainian Scientific Conference*, pages 113–119. NaUKMA.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of WMT 2023*.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of EAMT 2023*, pages 193–203.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? A case for document-level evaluation. In *Proceedings of EMNLP 2018*, pages 4791–4796.

Yevhenii Lepokhin. 2023. Vasyl Stefanyk's short story "All alone" as interpreted in translations by Olha Kobylianska, Mary Skrypnyk and Danylo Struk in English and German. *Slavia: Journal for Slavic Philology*, 92(3):322–353.

Viktoriia Maslij (Kalashnyk) and Mariia Shvedova. 2025. ParaFarm: English-Ukrainian multiple-translation corpus (1.1). Data set.

Yurii Paniv, Bohdan Didenko, Mykola Haltiuk, Vladyslav Humennyy, Andrian Kravchenko, Roman Kyslyi, Viktoriia Makovska, Artem Orlovskyi, Bohdan Ruban, Maksym-Yurii Rudko, Anastasiia Senyk, Nazarii Drushchak, Dmytro Chaplynskyi, and Mariana Romanyshyn. 2025. Lapa LLM v0.1.2 — the most efficient Ukrainian open-source language model.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of EMNLP 2020*.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-Unbabel 2022 submission for the quality estimation shared task. In *Proceedings of WMT 2022*.

Mariana Romanyshyn, Oleksiy Syvokon, and Roman Kyslyi. 2024. The UNLP 2024 shared task on fine-tuning large language models for Ukrainian. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 67–74.

Nataliia Ruda. 2021. Diminutive contronyms in Ukrainian. *Studia Slavica Academiae Scientiarum Hungaricae*, 65(2):341–350.

Jan Rybicki. 2012. The great mystery of the (almost) invisible translator: Stylometry in translation. In *Quantitative Methods in Corpus-Based Translation Studies*, pages 231–248. John Benjamins.

George Y. Shevelov. 1963. *The Syntax of Modern Literary Ukrainian: The Simple Sentence*. Mouton.

Bohdan Stasiuk. 2019. Conflict of editorial versions of old translations and the problem of their republishing. *Zapysky Naukovoho tovarystva imeni Shevchenka*, 272:496–514.

Antonio Toral and Andy Way. 2018. What level of quality can neural machine translation attain on literary text? In *Translation Quality Assessment*, pages 263–287. Springer.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of EMNLP 2023*.

Ran Zhang, Wei Zhao, and Steffen Eger. 2025. How good are LLMs for literary translation, really? literary translation evaluation with humans and LLMs. In *Proceedings of NAACL 2025*, pages 10961–10988.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36*.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024. Pitfalls and outlooks in using COMET. In *Proceedings of WMT 2024*, pages 1272–1288.

A Appendix: Additional Metric Heatmaps

This appendix reports the full pairwise round-robin score matrices for the reference-based metrics, complementing the XCOMET heatmap in Figure 3.

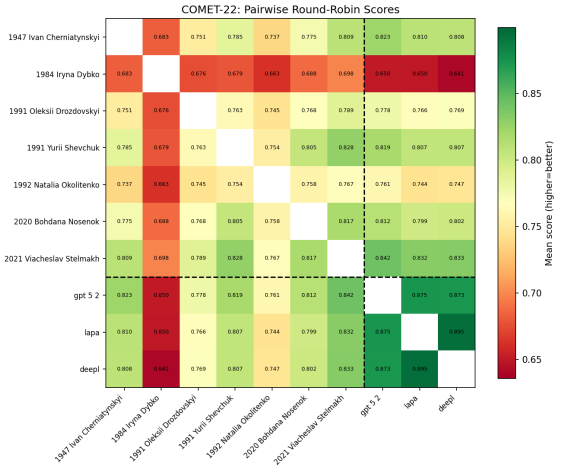


Figure 8: Heatmap of pairwise COMET-22 round-robin scores.

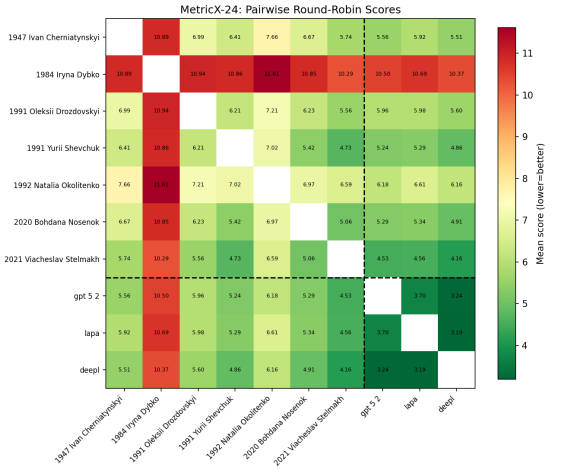


Figure 9: Heatmap of pairwise MetricX-24 round-robin scores.

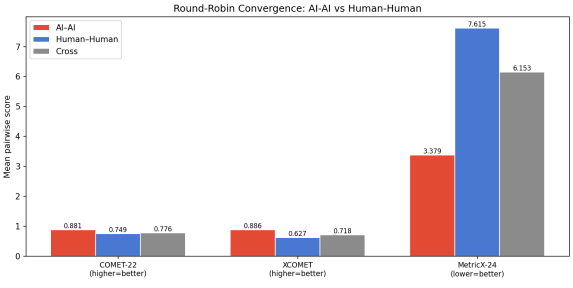


Figure 10: Round-robin convergence across MT metrics.

B Appendix: Additional LaBSE Analyses

This appendix presents additional cross-lingual similarity analyses based on LaBSE sentence em-

beddings.

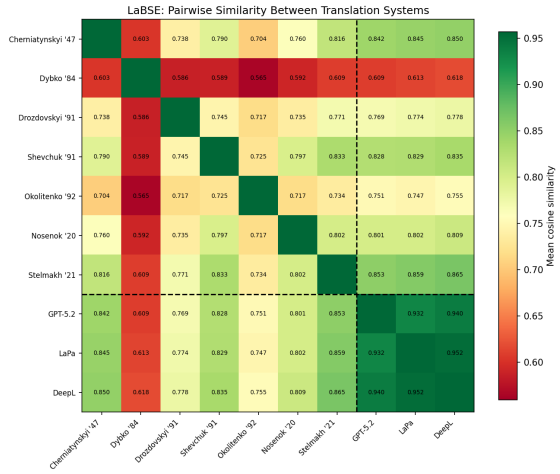


Figure 11: LaBSE pairwise similarity heatmap.

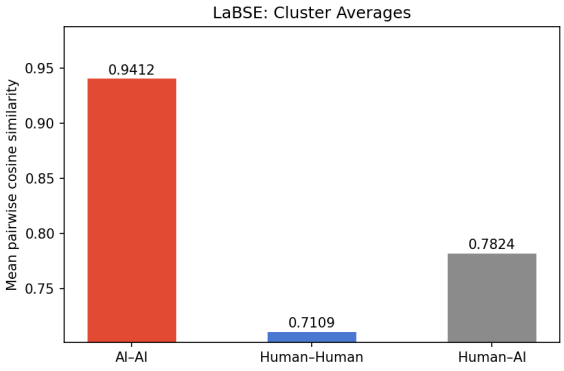


Figure 12: LaBSE cluster averages (AI–AI, Human–AI, Human–Human).

C Appendix: Additional Stylometric Analyses

This appendix collects the remaining stylometric comparisons across translators and AI systems.

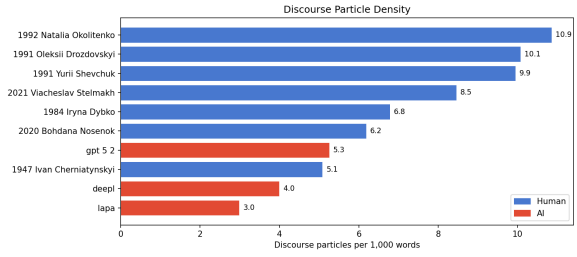


Figure 13: Discourse particle frequency across translators and AI systems.

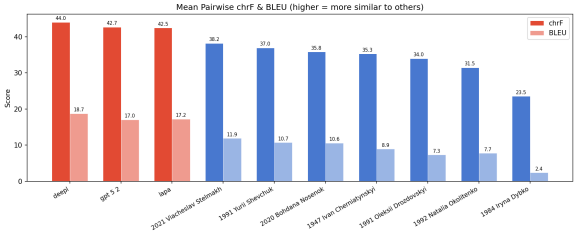


Figure 14: Mean pairwise chrF and BLEU scores.

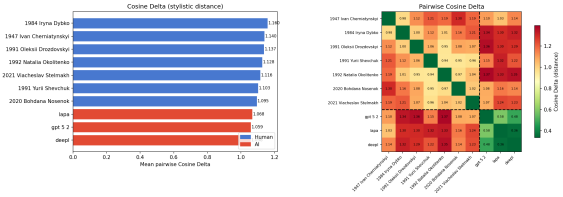


Figure 15: Cosine Delta heatmap showing stylistic distances between translators.

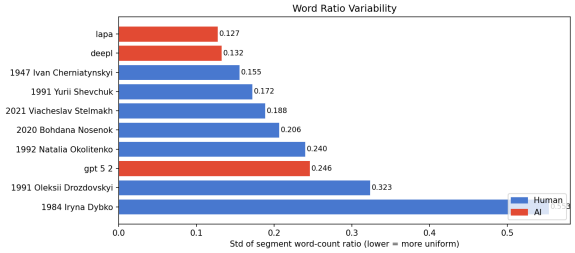


Figure 16: Word ratio uniformity across translators and AI systems.

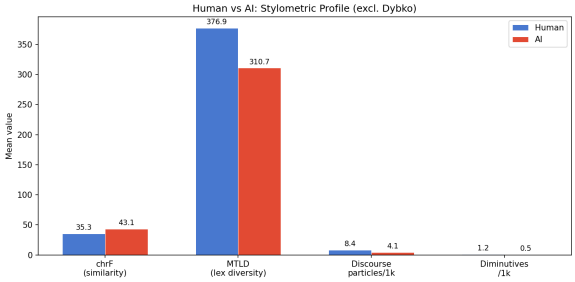


Figure 17: Stylometric convergence summary across all measures.

D Appendix: Prompts

AI translation prompt (Lapa, GPT-5.2).

> Translate the following English text to Ukrainian. Output only the translated text without any additional words or formatting, start with the translated text:

LLM judge: translation (source visible). System prompt:

> You are an expert literary Ukrainian translator evaluating translation quality. You will be given one English sentence and two Ukrainian translations. Choose the better translation.
>
> How to decide: 1. Meaning preservation — Does the translation convey the core meaning and intent of the English sentence? 2. Fluency and literary quality — Which translation reads more natural, expressive, and appropriate for literary Ukrainian?
>
> Rules: Prefer the translation that best balances intent and natural literary expression. Use "tie" only if you genuinely cannot decide. Judge each sentence independently.
>
> Respond with EXACTLY one of: `system1`, `system2`, `tie`.

User template: `English: {english} | system1: {system1} | system2: {system2}`

LLM judge: literary (source hidden). System prompt:

> You are an expert in Ukrainian literature. You will be given two Ukrainian sentences. Choose the one that sounds more literary — as if written by a skilled Ukrainian author for a published book.
>
> Judge only literary quality: naturalness, expressiveness, and stylistic richness of the Ukrainian language.
>
> Rules: Use "tie" only if you genuinely cannot decide. Judge each sentence independently.
>
> Respond with EXACTLY one of: `system1`, `system2`, `tie`.

User template: `system1: {system1} | system2: {system2}`

Both experiments use GPT-5.2 with temperature 0.0 and max 16 output tokens.

E Appendix: Reproducibility

All computational analyses are reproducible from the accompanying repositories: the translation metrics and stylometric pipeline at https://github.com/vanneqq/translation-metrics, and the human evaluation tournament implemented in Vulyk Arena at https://github.com/lang-uk/vulyk-arena. The three AI translations will be submitted as an update to the ParaFarm corpus to enable replication and further research.

Graph-Based Detection of Disinformation Narrative Diffusion between Russian and Ukrainian Telegram Channels

Yuliia Vistak
Ukrainian Catholic University
vistak.pn@ucu.edu.ua

Viktoriia Makovska
Ukrainian Catholic University
makovska.pn@ucu.edu.ua

Vera Schmitt
Technische Universität Berlin
vera.schmitt@tu-berlin.de

Veronika Solopova
Technische Universität Berlin
veronika.solopova@tu-berlin.de

Abstract

Detecting disinformation narratives on social media is challenging due to the scale of amplification, rapid evolution, and linguistic variability of online content. We propose a graph-based framework for identifying and analyzing disinformation narratives in Telegram ecosystems by combining weak supervision with propagation graph analysis. The approach aggregates semantically related claims into narrative-level clusters and models their diffusion across interconnected channels. This enables the detection of coordinated narrative amplification that is difficult to capture through post-level analysis alone. Our results demonstrate that integrating textual signals with network structure provides a scalable method for detecting disinformation narratives and offers insights into how they propagate within large-scale messaging environments.

1 Introduction

Disinformation at scale remains a persistent challenge for modern information ecosystems, with content volumes far exceeding the capacity of manual verification and fact-checking (Wardle and Derakhshan, 2017). This challenge is especially acute in conflict settings, where disinformation evolves rapidly and is repackaged across platforms and communities. That's why narrative-level representation offers a practical abstraction for analyzing high-volume environments: rather than tracking individual posts and factuality of their claims, they aggregate semantically related claims into higher-level frames that remain stable despite linguistic variation and cross-platform adaptation (Nikolaidis et al., 2025). Prior work in computational propaganda and fake news analysis shows that rhetorical framing and narrative structure capture systematic signals that generalize beyond surface text, enabling scalable trend detection across large corpora (Rashkin et al., 2017).

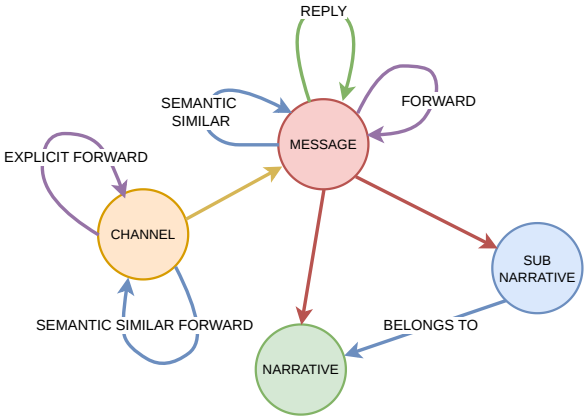


Figure 1: Schema of the graph used in our analysis, with channel, message, narrative, and sub-narrative nodes, and edges for posting, reply, explicit forwarding, semantic similarity, and narrative assignment.

Telegram is a particularly important platform for studying these dynamics, with a particularly high share of media space in Eastern Europe (Makhortykh et al., 2025). With moderation enforcement increased on mainstream social media in the early 2020s, harmful and conspiratorial communities migrated toward lower-moderation spaces such as Telegram (Kalkbrenner et al., 2025). Its channel-based broadcast architecture and native forwarding mechanism make information diffusion *structurally observable*: forwarded/forwarded-from metadata enables reconstruction of cross-channel propagation networks (La Morgia et al., 2025). Telegram is also a central venue for Kremlin-related and anti-Kremlin communications during the Russia-Ukraine conflict, operating at substantial scale (Bawa et al., 2025). Despite this, misinformation detection research remains heavily skewed toward English-language and platforms such as X (Kalkbrenner et al., 2025). Content-only classifiers are often brittle across languages and domains, while disinformation is frequently better characterized by amplifiers (e.g., botnets, troll

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 80–96

farms, and coordinated sockpuppet accounts) and their patterns, rather than by the lexicon of the messages. Propagation-aware methods exploit the empirical observation that deceptive and credible information spreads differently, thereby providing more language-agnostic and manipulation-resistant signals (Monti et al., 2019). Therefore, in this study, we pursue network-driven disinformation narrative analysis for the Ukrainian Telegram ecosystem, building on MisinfoTeleGraph (Kalkbrenner et al., 2025) while adapting it to a conflict-specific, narrative-centric setting. We collect posts from a manually curated set of Telegram channels and construct a directed *share graph* based on cross-channel forwarding (Figure 1). Our supervision target is *multi-class narrative assignment*: each message may activate multiple disinformation narratives. We operationalize the label space extending narrative taxonomy from the VoxCheck Propaganda Diary (database of narratives, created by fact-checking organization VoxCheck) and match messages against a fine-grained inventory of **380** pro-Russian disinformation subnarratives. [1] [2]

Methodologically, we compare three families of weak labeling approaches: semantic similarity to narrative descriptions, multilingual NLI-based zero-shot classification, and instruction-tuned LLM zero-, few-shot prompting. These methods are evaluated on a seed set of human-labeled messages, with model selection or ensembling guided by precision–coverage trade-offs. The resulting weak labels are attached to message nodes and, where needed, aggregated to the channel level. This labeled content is then analyzed in two complementary graph layers: an explicit share graph induced by Telegram forwarding metadata, and a semantic propagation graph induced by cross-channel message similarity.The full codebase, including labeling pipelines, graph construction, and evaluation scripts, is publicly available at GitHub repository. [3]

2 Background and Related Work

This section reviews the main strands of work that inform our approach: disinformation narrative analysis, weak supervision for large-scale labeling, and graph-based modeling of information spread. Together, these lines of research motivate our focus on narrative-level detection in Telegram and our combination of weak labeling with graph-based propagation analysis.

2.1 From Claims to Narratives

Narratives are recurring ideological structures that organize multiple claims into coherent worldviews (Hellman, 2024). Their effectiveness lies not in factual accuracy but in emotional resonance and identity alignment, which makes them resilient to corrective information (Dahlstrom, 2014). As a result, disinformation detection increasingly focuses on identifying narrative patterns rather than isolated falsehoods. Prior work explicitly models narratives as hierarchical or multi-level structures. PolyNarrative introduces a multilingual dataset with coarse- and fine-grained narrative labels (Nikolaidis et al., 2025), while datasets such as EUvsDisinfo provide benchmarks for detecting pro-Kremlin EU-targeting disinformation narratives in news articles (Leite et al., 2024). Domain-specific studies have also examined narrative ecosystems in contexts such as climate change denial (Upravitelev et al., 2026a,b). We build on this by adopting a hierarchical narrative representation tailored to the Ukrainian information environment.

2.2 Weak Supervision for Narrative Detection

Expert annotation remains a bottleneck for narrative-level disinformation detection. Programmatic Weak Supervision addresses this challenge by combining multiple noisy labeling sources (known as labeling functions) to generate scalable, probabilistic supervision (Ratner et al., 2017). Such sources may include heuristic rules, semantic similarity, or outputs of zero-shot classifiers.

In multilingual settings, weak labeling functions commonly rely on semantic similarity between texts and narrative descriptions, natural language inference (NLI)–based entailment scoring, or zero-shot classification with instruction-tuned language models (Yin et al., 2019; Schick and Schütze, 2021; Ratner et al., 2017). Weak supervision is well-suited for Telegram, where narrative inventories can be externalized and calibrated using small expert-labeled seed sets (Kalkbrenner et al., 2025).

2.3 Disinfo Spreading Networks Analysis via Graphs

Graph-based approaches model misinformation not only through message content, but also through the structure of its spread. Early work showed that

[1] https://russiandisinfo.voxukraine.org/
[2] https://voxukraine.org/en/voxcheck
[3] https://github.com/yuliavistak/TeleNarratives.git

credibility can be inferred in part from diffusion and interaction patterns in social media streams (Castillo et al., 2011). Large-scale evidence from Twitter demonstrated that false news spreads faster, farther, and more broadly than true news, which motivated propagation-aware approaches as an alternative to content-only classification (Vosoughi et al., 2018). Building on this, later work explicitly represented misinformation spread as graphs (Monti et al., 2019; Bian et al., 2020). Our work shifts the emphasis from message- or story-level misinformation detection to *narrative-level* propagation analysis in Telegram.

3 Methodology

Our methodology consists of four main components: (i) data collection, (ii) definition of a multi-label narrative assignment task using VoxCheck subnarratives, (iii) weak labeling via multiple few-shot and similarity-based labeling functions, and (iv) share graph construction from Telegram forwarding metadata.

3.1 Data Collection

We manually compiled a set of public Telegram channels relevant to war-related political discourse. The channel selection heavily relied on data from the fact-checking organization such as Spravdi. [4] We began with channels they identified as "unreliable" or "suspicious" and supplemented these with popular Ukrainian news channels exceeding 300,000 followers. This resulted in a total of **98** channels. Although we recognize that any curated list involves some bias, our aim was to make the final dataset as balanced as possible.

We also adopt the **VoxCheck Propaganda Diary**, which organizes pro-Russian disinformation into a structured taxonomy of **26** high-level **narratives** and **360** fine-grained **sub-narratives**, identified and curated during 2022–2023. Some of these narratives directly relate to the Russian–Ukrainian conflict (e.g. "The war in Ukraine demonstrates the supremacy of Russian weapons"), while others address broader societal allegations (e.g. "The Russian language is being suppressed in Ukraine"). This taxonomy reflects real-world monitoring practices and enables multi-class narrative assignment, capturing the fact that individual messages may activate multiple narrative elements simultaneously.

[4] Spravdi channel list.

The messages dataset includes (i) message text, (ii) timestamps, (iii) channel and message identifiers, and (iv) native forwarding provenance fields (`is_forwarded`, `fwd_from_channel_id`, `fwd_from_message_id`). This design tracks how content spreads across channels without using any private user data.

The resulting messages corpus is primarily bilingual, containing messages in both **Ukrainian** and **Russian**. Our primary objective was to analyze the discourse throughout 2025 and to include the early months of 2026, up to the date of our final data extraction. Consequently, the dataset consists of **1,352,668 messages** published between **16 December 2024** and **27 February 2026**. Figure 2 shows the temporal distribution of messages during the observation period.

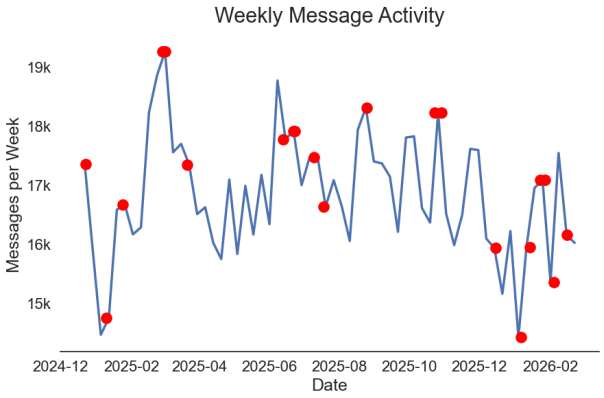


Figure 2: Weekly message activity within the curated Telegram corpus. Red markers denote significant political or conflict-related events (see Table 4 in the appendix for a comprehensive list). The peaks in message volume frequently coincide with these events.

To investigate the propagation of specific narratives and the underlying network topology between channels, we focused our analysis on the subset of messages involved in forwarding chains. This includes both **forwarded messages** and the **original posts** from which content was forwarded. From the initial corpus, this refined subset consists of **70,595** messages.

Some channels do not use the official "forward" button, copying and pasting text with only tiny changes. We call this *implicit forwarding*. To catch these cases, we looked for messages with very similar meanings. For that reason, the decision was to transform messages into embedding vectors via `baai/bge-m3`, compute the similarity score between them, and take pairs with high score

in further analysis. [5]

This step enabled the inclusion of **10,872** semantically similar pairs (representing **10,774** additional messages). By tracking both official forwards and these "copy-paste" reposts, we get a much clearer picture of how news and narratives actually spread.

The final dataset analyzed in this study consists of **81,369** messages, which were subsequently annotated for further investigation.

3.2 Multi-Class Narrative Assignment

We formulate disinformation detection as a **multi-class narrative assignment** problem. The narratives dataset is organized into a two-level hierarchy:

Narratives (26 total): Broad, general themes of disinformation (e.g., "The West controls Ukraine and uses it for its own purposes").

Sub-narratives (360 total): Specific claims and real-world examples that fit within those broader themes (e.g., "The US is using Ukraine to weaken Europe").

For manual annotation, we sampled 412 messages (*'golden set'*). To capture the corpus's temporal structure and account for activity fluctuations throughout the year, we employed **stratified random sampling** by week of the year.

Three independent annotators conducted the labeling in two phases. First, to establish inter-annotator agreement (IAA), they labeled *a shared subset* of 103 messages. Second, each annotator independently labeled an additional 103 distinct messages to expand the dataset. Annotators assigned each message to the best-matching narrative from the 26 predefined ones, or marked it as *not containing a narrative*.

During analysis, we observed that many messages contained pro-Russian narratives absent from the original taxonomy. This likely occurred because the messages in the dataset were relatively recent, while the existing set of narratives had been defined earlier and therefore did not fully capture newly emerging disinformation narratives. To address this gap, annotators identified **20 new sub-narratives**, which were subsequently harmonized and grouped into **6 new narrative categories**.

Furthermore, because single messages frequently contained multiple contextually related narratives, we grouped these interrelated narratives under broader, overarching meta-narratives.

The final dataset, **TeleNarratives**, is organized as a three-level taxonomy with **380 sub-narratives**, **32 narratives**, and **9 meta-narratives**. The complete resource, including a **Neo4j**[6] **graph database dump**, is publicly available at the project repository.

3.3 Weak Labeling

Given the scale of the corpus, full manual annotation was impractical due to limited human resources. Therefore, we adopted a **weak labeling** approach.

To implement this approach, we explored three different methods. The core idea behind these strategies was to compare each message with the sub-narratives rather than broader narratives. It allows models to perform more precise semantic comparisons between the message content and the narrative descriptions. Our goal was to evaluate the performance of these strategies and select the most effective one based on its agreement with the **golden set**.

Semantic similarity-based labeling. This approach utilizes **sentence-transformer models** to represent messages and sub-narratives as embedding vectors. Using cosine similarity, each message is assigned to the most similar sub-narrative, provided the score exceeds a predefined threshold. Messages that do not meet this criterion are classified as **not containing a narrative**.

To generate embeddings, we experimented with three multilingual models: `paraphrase-multilingual-MiniLM-L12-v2`, `BAAI/bge-m3`, `text-embedding-3-large (OpenAI)`[7] [8] [9]. These models were selected to compare different multilingual embedding approaches, including lightweight open-source models and larger, high-capacity proprietary systems.

Natural Language Inference (NLI)-based labeling. Natural Language Inference (NLI) models identify the logical relationship between two text segments as *entailment*, *contradiction*, or *neutral*. In this study, NLI models assess whether a message supports a particular sub-narrative.

Under this framework, the message acts as the *premise*, and the sub-narrative description

[5] https://huggingface.co/BAAI/bge-m3

[6] https://neo4j.com/

[7] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

[8] https://huggingface.co/BAAI/bge-m3

[9] https://developers.openai.com/api/docs/models/text-embedding-3-large

as the *hypothesis*. The model calculates the probability of entailment for each pair; the message then receives the label of the sub-narrative with the highest score, provided it exceeds a predefined threshold. Messages that do not meet this criterion are classified as **not containing a narrative**. We experimented with `mDeBERTa-v3-base-xnli-multilingual-nli`[10] as a multilingual NLI model. We selected this model for its rare Ukrainian language support and strong community validation on Hugging Face.

LLM-based labeling with zero-shot and few-shot prompting. In this approach, the model identifies whether a message contains a specific sub-narrative and provides a confidence score. In this sense, the LLM acts as an 'additional annotator', producing labels that can be compared with human annotations.

We explored two prompting approaches: **zero-shot** and **few-shot**. In the zero-shot setting, the model receives only task instructions. In the few-shot setting, the prompt also includes several labeled examples with brief explanations. These examples show the model how to identify narratives in practice.

To evaluate the performance of this method and identify the optimal price-quality ratio for the task, the following LLMs were selected: `GPT-4.1`[11], `GPT-4o-mini`[12], `Gemini-2.5-flash`[13], `Claude-sonnet-4-20250514`[14].

3.4 Share Graph Construction

We model message diffusion using two complementary layers: (i) an explicit channel-level share graph derived from Telegram forwarding metadata, and (ii) a semantic message-level graph designed to recover repost-like diffusion that is not marked as a native forward.

Forward-based channel share graph. Let $G_{\text{share}} = (V_C, E_F, w_F)$, where V_C is the set of observed Telegram channels. We add a directed edge $(u \to v) \in E_F$ when channel v contains a message forwarded from channel u. Edge weights count observed forwarding events:

$$w_F(u, v) = \sum_{m \in M_v} \mathbf{1}[\texttt{fwd\_from\_channel}(m) = u] \,,$$

where M_v denotes the set of messages posted in channel v. In implementation, we retain a forward edge only when the referenced source message is present in the collected corpus. This restriction avoids links to unobserved sources and ensures that all edges in G_{share} are supported by directly observed data.

Semantic message graph for repost-like diffusion. Explicit forwarding provides a high-precision, platform-native signal of diffusion, but it does not capture copy-paste reposts or lightly edited message reuse. To approximate such hidden propagation, we construct a message-level semantic graph with edges of type `SIMILAR_TO` marked in the dataset.

Before embedding, we normalize message text and exclude duplicates attributable to explicit forwarding chains. We then encode the remaining Ukrainian- and Russian-language messages using `baai/bge-m3` and retrieve approximate nearest neighbors with an HNSW index (Malkov and Yashunin, 2018).

For each message, we retrieve the top-k neighbors with $k = 80$, $M = 48$, $ef_{\text{construction}} = 200$, and $ef_{\text{search}} = 200$; these settings preserved recall for subtle multilingual paraphrases while keeping index size and query latency tractable, with only marginal gains beyond them. We retain only high-confidence cross-channel pairs above a calibrated cosine threshold $\tau = 0.985$, excluding same-channel pairs and pairs already linked by explicit forwarding metadata. We selected τ using 3,215 mapped hidden-forward pairs: recall was 81.0% at 0.98, 76.39% at 0.985, and 69.58% at 0.99, while the number of retained pairs after filtering decreased from 35,092 to 24,692 and 19,095, respectively, providing a practical balance between recall and edge inflation.

We selected `baai/bge-m3`, `intfloat/multilingual-e5-large`, and `gemini-embedding-001` as pilot candidates because all three are strong multilingual embedding models suitable for cross-lingual semantic retrieval. In pilot calibration on manually inspected Ukrainian/Russian message pairs, `baai/bge-m3` showed the clearest separation among these candidates.

[10] https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7

[11] https://developers.openai.com/api/docs/models/gpt-4.1

[12] https://developers.openai.com/api/docs/models/gpt-4o-mini

[13] https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash

[14] https://platform.claude.com/docs/en/about-claude/models/overview

In a 16-pair pilot containing exact-like positives and manually identified hard negatives, the hard-negative pairs had a mean cosine similarity of 0.328 under `baai/bge-m3`, compared with 0.744 for `intfloat/multilingual-e5-large` and 0.762 for `gemini-embedding-001`. We therefore used `baai/bge-m3` for the full embedding run.

4 Experiments, Evaluation, Results

We evaluate the proposed framework in three stages. First, we measure IAA to verify the reliability of the manually labeled data. Second, we compare the weak labeling strategies from Section 3.3 and identify the most effective setup for large-scale narrative assignment. Third, using the selected labels, we analyze the diffusion graphs to study source behavior, cross-group sharing, multi-hop propagation, and narrative-level dissemination patterns. This organization allows us to move from label quality to structural findings and to assess whether combining weak supervision with graph analysis yields an informative view of the spread of disinformation narratives on Ukrainian Telegram.

4.1 Annotator Agreement

To evaluate the quality of the manual annotation, we computed **Cohen's Kappa**, **Fleiss' Kappa**, and **Krippendorff's Alpha**. Cohen's Kappa measures agreement between pairs of annotators while accounting for chance agreement. Fleiss' Kappa generalizes this measure to multiple annotators and categorical data. Krippendorff's Alpha provides a flexible measure of inter-annotator reliability that can handle multiple annotators, missing data, and different data types.

The results are presented in Table 1. Binary classification achieved near-perfect agreement ($\kappa/\alpha \approx 0.887$), while meta-narrative labeling reached substantial agreement (≈ 0.719). Fine-grained narrative annotation showed moderate agreement (≈ 0.626), suggesting that this task is at the narrative level relatively subjective and harder to annotate reliably. Notably, all three metrics (Cohen's Kappa, Fleiss' Kappa, and Krippendorff's Alpha) produce almost identical within each category, which strongly confirms the reliability and consistency of these measurements.

Metric	Narrative	Meta-narrative	Binary
Cohen's Kappa (mean)	0.626	0.719	0.887
Fleiss' Kappa	0.626	0.718	0.887
Krippendorff's Alpha	0.627	0.719	0.887

Table 1: Inter-annotator agreement results.

4.2 Weak Labeling Results

The models were executed to evaluate each labeling strategy across the taxonomy. Since each sub-narrative is linked to a specific narrative and meta-narrative, identifying a sub-narrative automatically determines its higher-level categories. This mapping allows predictions to be evaluated at all levels of the hierarchical structure.

Due to label imbalance (i.e., a majority of non-narrative messages and an uneven distribution of specific narratives), we focused on metrics robust to skewed data: **F1** for binary classification (whether a message contains a narrative), **Weighted F1** for multi-class classification, **Matthews Correlation Coefficient (MCC)** for both.

First, we evaluated the models used in the **semantic similarity**, **multilingual NLI**, and **LLM zero-shot prompting** strategies. However, their performance on *the shared subset* of 103 messages was relatively limited (see Table 4.2 for detailed results).

We also explored **ensemble approaches** that combined predictions from multiple LLMs. In particular, we tested ensembles consisting of `Claude`, `GPT-4.1`, and `Gemini`, as well as `GPT-4o-mini`, `GPT-4.1`, and `Gemini`, assigning a higher weight to Gemini due to its stronger individual performance. However, these ensemble configurations did not lead to a meaningful improvement in overall results (as Gemini consistently outperformed them across all evaluated metrics).

We found that the main difficulty is that the relationship between a message and a narrative is often **implicit**. As a result, the message may not be semantically similar to the corresponding narrative description. Instead, the narrative needs to be inferred from the common (sometimes political) context.

Because of this, the **semantic similarity** and **multilingual NLI** approaches were not suitable for our task. We therefore modified the LLM-based strategy by using **few-shot prompting**, adding several annotated examples to the prompt to illustrate the implicit relationship between messages and narratives. This approach produced significantly better results. To ensure that the method performs consistently on a larger sample, we repeated the evaluation on the full set of 412 manually annotated messages. The results remained strong ($F_1 = 0.82$,

$\text{MCC}_{\text{bin}} = 0.71$, W-$F_1 = 0.76$, $\text{MCC}_{\text{meta}} = 0.57$). The final version of the prompt used in the experiments is provided in Appendix E.

Classifier	Binary		Meta	
	F1	MCC	W-F1	MCC
Semantic Similarity				
MiniLM-L12-v2*	0.51	0.14	0.41	0.11
BGE-M3*	0.37	0.08	0.54	0.13
OpenAI Text-3*	0.41	-0.07	0.36	0.03
Multilingual NLI				
mDeBERTa-v3-base*	0.54	0.15	0.32	0.13
LLM (Zero-Shot)				
GPT-4o-mini*	0.51	0.23	0.53	0.17
GPT-4.1*	0.65	0.60	0.69	0.50
Gemini-2.5-Flash*	0.76	0.63	0.74	0.52
Claude-Sonnet-3.5*	0.60	0.55	0.67	0.40
LLM (Ensemble)				
GPT-4.1 + Claude + Gemini*	0.69	0.61	0.72	0.50
GPT-4.1 + GPT-4o-mini + Gemini*	0.71	0.61	0.70	0.44
LLM (Few-Shot)				
Gemini-2.5-Flash*	**0.85**	**0.78**	**0.80**	**0.62**
Gemini-2.5-Flash**	**0.82**	**0.71**	**0.76**	**0.57**

Table 2: LM-based approaches significantly outperform Semantic Similarity and NLI baselines, with **Gemini-2.5-Flash (Few-Shot)** achieving the highest overall performance. Results marked with (*) denote the *shared subset*, while (**) represents the full *golden set*. Meta denoted Meta-narrative.

4.3 Share Graph Results

We next analyze the share graph to understand what its structure reveals about how narratives spread across channels. Our analysis focuses on three questions: which channels are likely to occupy upstream positions, how often sharing occurs across channel groups, and how these patterns differ between the explicit-forwarding and semantic-similarity layers. Formal definitions of the metrics and additional robustness details are provided in Appendix A.

Narrative source detection. We first examine whether explicit forwarding structure can identify likely upstream sources of narrative dissemination. For this analysis, a channel is classified as *narrative-active* if it satisfies two conditions: (i) it has at least 50 labeled messages, and (ii) at least 60% of those labeled messages receive a narrative label. Under this rule, 41 of the 98 observed channels are classified as narrative-active. The remaining 57 channels are treated as non-narrative for this grouping analysis, including 25 channels that meet the minimum label-count requirement but fall below the 60% narrative threshold, and 32 channels with limited evidence.

To characterize source behavior in the explicit forwarding graph, we use three complementary channel-level metrics: total forwarding volume (*spread_events*), the number of distinct recipient channels reached (*downstream_channels*), and a normalized source tendency score (*source_share*) that distinguishes net sources from channels that mostly relay content from others.

Across the 41 narrative-active channels, we observe 3,016 explicit spread events. By forwarding volume, `rian_ru` is the dominant source with 612 forwarding events, accounting for 20.3% of all spread events in this subset. It is followed by `depzdravzo` (337), `Pukhov_M` (303), `hersonruss` (228), and `readovkanews` (209). `rian_ru` also ranks first in dissemination breadth, reaching 17 distinct downstream channels, and has a source-share score of 1.00, indicating a purely source-like position in the observed forwarding network. These results show that the explicit share graph provides a useful signal for identifying likely upstream narrative spreaders, while still reflecting source positions only within the collected network rather than absolute first creators outside it.

Explicit graph hop calibration and sensitivity. We selected the hop budget by examining the shortest-path distribution over reachable ordered pairs. The smallest value covering at least 85% of reachable ordered pairs is $h = 4$, which covers 92.28% of them. At $h = 4$, overall cross-group reachability is 0.0211, with reachability of 0.0153 from narrative-active to non-narrative channels and 0.0268 in the reverse direction. Increasing the hop budget to $h = 8$ raises reachability from non-narrative to narrative-active channels to 0.03, while reachability from narrative-active to non-narrative channels remains 0.0153.

Cross-group sharing analysis. We next examine if explicit forwarding crosses channel groups defined by channel-level narrative presence. To quantify this, we compare within-group and cross-group forwarding shares and then examine whether

cross-group contact becomes more common when short multi-hop paths are allowed.

Direct cross-group forwarding is rare: among 34,967 forwarding events, only 121 cross group boundaries, corresponding to a share of 0.0035. These events are limited in both directions, with 57 events from narrative-active to non-narrative channels and 64 in the reverse direction. Allowing short multi-hop paths does not substantially change this conclusion. Using a hop budget calibrated on the shortest-path distribution, overall cross-group reachability in the explicit forwarding graph is only 0.02, with somewhat higher reachability from non-narrative channels into the narrative-active group than in the reverse direction. Overall, most forwarding remains in-group, and bridge-like cross-group routes are uncommon.

Semantic graph flow and reachability details. In the semantic similarity graph, message-level semantic links are oriented by time and aggregated into channel-level semantic flow counts, where $\mathcal{W}^{\text{sem}}_{u \to v}$ denotes the number of semantic propagation events from channel u to channel v. Direct cross-group semantic flow is summarized using the same within-group and cross-group share logic as in the explicit forwarding graph. For indirect connectivity, we calibrate the hop budget on the shortest-path distribution over reachable ordered channel pairs. The smallest cutoff covering at least 85% of reachable ordered pairs is $h = 3$, corresponding to 85.65% coverage. At $h = 3$, overall cross-group reachability is 0.5385, with reachability of 0.5049 from narrative-active to non-narrative channels and 0.5721 in the reverse direction. A larger hop budget, $h = 8$, increases these values to 0.669 and 0.6387, respectively, with 0.6538 overall.

Cross-group sharing in the semantic similarity graph. To complement the explicit forwarding analysis, we examine cross-group diffusion in the semantic similarity graph, where message-level semantic links are time-oriented and aggregated to channel-level semantic flows. We then compare within-group and cross-group semantic flows and assess whether the two groups are connected via short multi-hop paths.

Direct cross-group semantic flow remains limited but is more common than in the explicit forwarding graph. Among 52,414 semantic flow events, 1,078 cross group boundaries, corresponding to a share of 0.02. The directional split is also asymmetric: 256 events run from narrative-active to non-narrative channels, while 822 run in the reverse direction. The main contrast appears in the multi-hop setting. At hop budget $h = 3$, overall cross-group reachability rises to 0.54, far above the corresponding value in the explicit forwarding graph. This suggests a repost-like narrative transfer frequently connects the two groups through intermediate channels even when direct cross-group links remain relatively uncommon.

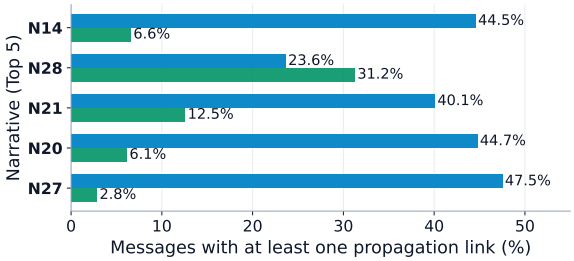


Figure 3: Coverage-based propagation comparison for the five most frequent narratives: N14 (Discrediting or ridiculing representatives of Ukrainian authorities), N28 (Russia improves life in occupied territories), N21 (Ukraine's victory is impossible), N20 (The West controls Ukraine and uses it for its own goals), and N27 (Discrediting the EU and the West). Blue bars show the share of messages with at least one explicit forward (`FORWARD_FROM`); green bars show the share of messages with at least one semantic forward (`SIMILAR_TO`).

4.4 Narrative Distribution

We next examine narrative distribution in the labeled graph scope. Of the 81,369 labeled messages, 29,649 (36%) receive a narrative assignment. Figure 3 compares propagation coverage for these five narratives across the explicit and semantic graph layers. At the macro-family level, the distribution is dominated by *Discrediting Ukraine and its institutions* (33%), followed by *Narratives about Russian welfare and "liberating" role* (17.2%) and *Discrediting the EU and the West* (14.2%) (See Appendix D). This suggests that narrative prominence in the corpus and propagation strength do not align uniformly across diffusion layers (explicit forward and semantic similar forward).

5 Discussion and Conclusion

This work introduces a graph-based framework for detecting and analyzing disinformation narratives in Telegram ecosystems by combining weak supervision with propagation graph analysis. Modeling disinformation at the narrative level enables the

system to capture semantically related claims that appear in different linguistic forms, addressing the challenge of repeated paraphrasing and reposting common in Telegram channels. The propagation graph further provides a structural perspective on information spread, revealing how narratives circulate across interconnected channels rather than appearing as isolated posts. Graph representation highlights clusters of channels that repeatedly amplify similar narratives, offering insights into how disinformation spreads through platform-specific communication patterns. Beyond improving large-scale detection, this approach provides a framework for studying narrative diffusion and monitoring emerging disinformation campaigns in rapidly evolving online environments such as Telegram, where traditional post-level analysis often fails to capture broader information dynamics.

Limitations

Our approach has several limitations. First, the framework relies on weak supervision for narrative labeling, which can introduce label noise. While this enables scalable annotation of large Telegram datasets, automatically generated labels may include ambiguous or incorrectly assigned instances, potentially affecting downstream analysis. In the study, we estimated the amount of errors which one can expect from our best-performing solution. However, on different channels and over time the performance might degrade or perform better.

Second, the narrative taxonomy used for labeling may be incomplete and subject to drift over time. As new narratives emerge or existing ones evolve, the predefined taxonomy may fail to capture all relevant claims. In our experiments, we observe indications of narrative drift when additional narratives are introduced, suggesting that narrative boundaries are dynamic and context-dependent.

Third, the dataset is limited to a selected set of Telegram channels, which may introduce sampling bias, especially as we only subsample from the messages that have forwarding connections. The analyzed network may therefore not fully represent the broader Telegram information ecosystem nor the channels where they come from.

Finally, we do not investigate variability of Ukrainian language and code-switch varieties, and how performance of narrative detection decreases on such instances, while this is expected behaviour based on prior research Shynkarov et al. (2025).

Ethical Considerations

Our analysis relies on publicly accessible Telegram channels. Although these data are publicly available, users may not anticipate their messages being analyzed in large-scale computational studies. To mitigate potential privacy risks, we focus on aggregate patterns of narrative propagation and big public channels rather than individual user behavior.

Automated systems for detecting disinformation may produce false positives or misclassify legitimate content. Such errors could contribute to unfair labeling of channels or narratives if used without appropriate human oversight. Our framework is intended as a research tool for analyzing information dynamics rather than as a standalone moderation system.

Methods for detecting narrative propagation may also inform adversarial actors about how such systems operate. Increased awareness of detection strategies could encourage actors to adapt their communication patterns to evade analysis.

Using an external narrative inventory (VoxCheck) embeds expert judgments about what constitutes a disinformation narrative. While grounded in professional fact-checking practice, such inventories reflect particular epistemic and institutional perspectives and may not capture all interpretations of contested claims.

Acknowledgements

This research was partially supported by ELEKS through a grant dedicated to the memory of Oleksiy Skrypnyk. The work on this paper is partially performed in the scope of the project “VeraXtract” (16IS24066) funded by the German Federal Ministry for Research, Technology and Aeronautics (BMFTR).

References

Apaar Bawa, Ugur Kursuncu, Dilshod Achilov, Valerie L. Shalin, Nitin Agarwal, and Esra Akbas. 2025. Telegram as a Battlefield: Kremlin-Related Communications During the Russia-Ukraine Conflict. *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1):2361–2370.

Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. *Preprint*, arXiv:2001.06362.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, page 675–684, New York, NY, USA. Association for Computing Machinery.

Michael F. Dahlstrom. 2014. Using narratives and storytelling to communicate science with nonexpert audiences. *Proceedings of the National Academy of Sciences*, 111(Supplement 4):13614–13620.

Maria Hellman. 2024. *Narrative Analysis and Framing Analysis of Disinformation*, pages 101–121. Springer Nature Switzerland, Cham.

Lu Kalkbrenner, Veronika Solopova, Steffen Zeiler, Robert Nickel, and Dorothea Kolossa. 2025. MisinfoTeleGraph: Network-driven misinformation detection for German telegram messages. In *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*, pages 179–191, Vienna, Austria. Association for Computational Linguistics.

Massimo La Morgia, Alessandro Mei, and Alberto Maria Mongardini. 2025. Tgdataset: Collecting and exploring the largest telegram channels dataset. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, KDD '25, page 2325–2334, New York, NY, USA. Association for Computing Machinery.

João A. Leite, Olesya Razuvayevskaya, Kalina Bontcheva, and Carolina Scarton. 2024. Euvsdisinfo: A dataset for multilingual detection of pro-kremlin disinformation in news articles. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 5380–5384, New York, NY, USA. Association for Computing Machinery.

Mykola Makhortykh, Aytalina Kulichkina, and Kateryna Maikovska. 2025. Evolution of wartime discourse on telegram: A comparative study of ukrainian and russian policymakers' communication before and after russia's full-scale invasion of ukraine. *Preprint*, arXiv:2510.11746.

Yu. A. Malkov and D. A. Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *Preprint*, arXiv:1603.09320.

Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M. Bronstein. 2019. Fake News Detection on Social Media using Geometric Deep Learning. *Preprint*, arXiv:1902.06673.

Nikolaos Nikolaidis, Nicolas Stefanovitch, Purificação Silvano, Dimitar Iliyanov Dimitrov, Roman Yangarber, Nuno Guimarães, Elisa Sartori, Ion Androutsopoulos, Preslav Nakov, Giovanni Da San Martino, and Jakub Piskorski. 2025. PolyNarrative: A multilingual, multilabel, multi-domain dataset for narrative extraction from news articles. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31323–31345, Vienna, Austria. Association for Computational Linguistics.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: rapid training data creation with weak supervision. *Proc. VLDB Endow.*, 11(3):269–282.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Yurii Shynkarov, Veronika Solopova, and Vera Schmitt. 2025. Improving sentiment analysis for Ukrainian social media code-switching data. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 179–193, Vienna, Austria (online). Association for Computational Linguistics.

Max Upravitelev, Veronika Solopova, Charlott Jakob, Premtim Sahitaj, Sebastian Möller, and Vera Schmitt. 2026a. Retrieving climate change disinformation by narrative. *Preprint*, arXiv:2603.22015.

Max Upravitelev, Veronika Solopova, Charlott Jakob, Premtim Sahitaj, Sebastian Möller, and Vera Schmitt. 2026b. Retrieving climate change disinformation by narrative. *Preprint*, arXiv:2603.22015.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Claire Wardle and Hossein Derakhshan. 2017. Information disorder: Toward an interdisciplinary framework for research and policy making. Technical report, Council of Europe.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

A Details for Share-Graph Results

Source metrics in the explicit forwarding graph. Let $W_{u\to v}$ denote the number of explicit forwarding events in which channel v forwards content originally posted by channel u. We summarize each channel's source behavior using three forwarding-based metrics:

$$\begin{aligned} \textit{spread\_events}(u) &= \sum_{v\neq u} W_{u\to v}, \\ \textit{downstream\_channels}(u) &= \left|\{v \neq u : W_{u\to v} > 0\}\right|, \\ \textit{source\_share}(u) &= \frac{\sum_{v\neq u} W_{u\to v}}{\sum_{v\neq u}(W_{u\to v} + W_{v\to u})}. \end{aligned}$$

The first metric captures total propagation volume, the second captures dissemination breadth across distinct recipient channels, and the third measures net-source tendency on a bounded $[0, 1]$ scale.

Cross-group forwarding metrics. To quantify within-group versus cross-group forwarding in the explicit share graph, we compute

$$\begin{aligned} \textit{same\_label\_share} &= \frac{\sum_{u\neq v} W_{u\to v}\,\mathbf{1}[y(u) = y(v)]}{\sum_{u\neq v} W_{u\to v}}, \\ \textit{cross\_label\_share} &= 1 - \textit{same\_label\_share}. \end{aligned}$$

where $y(u)$ denotes the group label of channel u.

Cross-group reachability definition. For ordered channel pairs (u, v), let $d(u, v)$ be the directed hop distance from u to v. We define cross-group reachability under hop budget h as

$$\begin{aligned} R_{\text{cross}}(h) &= \frac{|\{(u, v) \in C : d(u, v) \leq h\}|}{|C|}, \\ C &= \{(u, v) : y(u) \neq y(v)\}. \end{aligned}$$

B Cross-Group Shares

Table 3 summarizes the direction of cross-group events, separately for explicit forwarding and semantic similarity.

Additional examples of such messages, along with their detected narratives and sub-narratives, are presented in Table 5 for cross-group explicit-forwarding from narrative-active to non-narrative channels, Table 6 for cross-group explicit-forwarding from non-narrative to narrative-active channels, Table 7 for cross-group semantic-similarity from narrative-active to non-narrative channels, and Table 8 for cross-group semantic-similarity from non-narrative to narrative-active channels.

C Chronological List of Events

A chronological list of events corresponding to markers in Figure 2 is shown in Table 4.

From	To	Events	Share
Explicit forwarding			
Narrative-active	Non-narrative	57	0.4711
Non-narrative	Narrative-active	64	0.5289
Semantic similarity			
Narrative-active	Non-narrative	256	0.2375
Non-narrative	Narrative-active	822	0.7625

Table 3: Summary of cross-group events for explicit forwarding and semantic similarity.

D Narrative Distribution

Supplementary visualizations for the narrative distribution analysis discussed in Section 4.4. Table 4 shows narrative distribution at the meta-narrative level.

E Final LLM Prompt

This appendix provides both the original Ukrainian version (Table 9) and the English translation (Table 10) of the final prompt used in the LLM-based weak labeling approach. The prompt was executed using `Gemini-2.5-flash`.

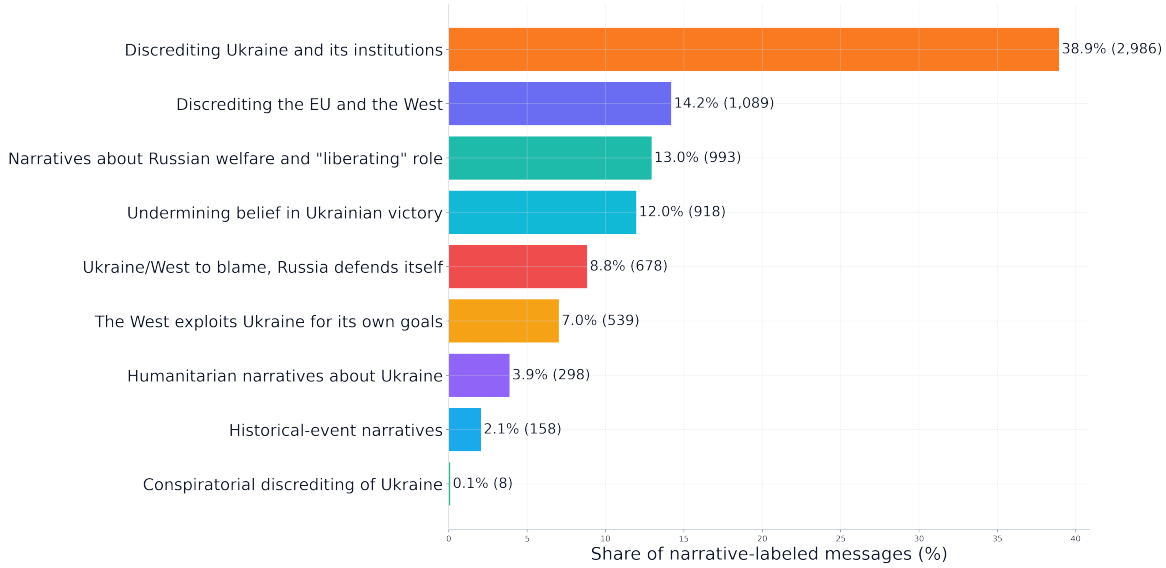


Figure 4: Narrative distribution at the meta-narrative level.

Date	Event
2024-12-19	EU agreed on a €90B loan programme for Ukraine (2026–2027).
2024-12-22	Slovak PM Fico visits Moscow; protests begin.
2025-01-10	Large-scale protests in Slovakia.
2025-01-24	New protests following "attempted coup" statements.
2025-02-28	Ukraine-US meeting at White House on strategy.
2025-03-02	London Summit on Ukraine.
2025-03-21	German Bundestag approves €3B military aid.
2025-06-13	Iran's massive missile strike on Israel.
2025-06-22	US air strikes on Iranian nuclear facilities.
2025-06-23	Iran attacks US base Al-Udeid in Qatar.
2025-07-10	EU announces €2.3B in reconstruction funding.
2025-07-18	EU adopts 18th package of sanctions.
2025-08-25	EU announces additional €4B in aid.
2025-10-23	EU adopts 19th package of sanctions.
2025-10-29	Updated EU-Ukraine trade agreement in force.
2025-12-15	European leaders propose peace plan.
2026-01-06	Declaration for multinational security force.
2026-01-14	EC proposes €90B in aid for 2026–2027.
2026-01-23	1st round of negotiations (Abu Dhabi).
2026-02-04	2nd round of negotiations (Abu Dhabi).
2026-02-15	Munich Security Conference 2026.
2026-02-23	Hungary blocks new EU sanctions/aid.

Table 4: Chronological list of events corresponding to markers in Figure 2.

Message (English translation)	Share count	Narrative	Sub-narrative
In Kremenchuk, TCC officers chased a young man and deliberately crashed into parked cars. The video shows how the TCC men deliberately ram the car, because they know that the law does not exist for them. A day before that, these same draft officers in uniform opened fire. The head of the Main Directorate of the National Police in Poltava oblast remains silent.	1	Discrediting the Ukrainian army	Mobilization in Ukraine violates all standards
The Wall Street Journal, citing sources in the American administration: the United States has removed the key restriction on Ukraine's use of Western long-range missiles, which will allow Kyiv to strike deep into Russia. This shift coincided with President Trump's attempt in early October to pressure the Kremlin to begin negotiations on ending the war.	1	The West controls Ukraine and uses it for its own goals	Ukraine is an instrument of the United States
Trump has become more detached on the issue of the Ukrainian conflict and no longer lashes out at Russia and Ukraine over the lack of progress in the negotiations, ABC reports, citing American officials.	1	Ukraine and the West refuse to start peace negotiations	Ukraine must immediately begin peace negotiations and make compromises with Russia

Table 5: Representative cross-group explicit-forwarding examples from narrative-active to non-narrative channels.

Message (English translation)	Share count	Narrative	Sub-narrative
It feels as though the wrong people were called occupiers. This is how the TCC tried to abduct a serf in front of his pregnant wife and children. People fought them off, but what moral freaks they are. Ze mobilization is the genocide of the Ukrainian people.	3	Discrediting or ridiculing representatives of Ukrainian authorities	Zelensky wants to sacrifice Ukrainians for his own interests
To call Putin an old fantasist while having 78-year-old Trump next to him is a brilliant move. I love it when this idiot buries himself.	1	Discrediting or ridiculing representatives of Ukrainian authorities	Ridiculing representatives of the Ukrainian authorities
"Ukraine is an artificially created country. When Putin takes Odesa, everything will be fine for us," just a survey on the streets of Odesa. More than 3 years of full-scale war. What is in their heads?	1	Narrative about historical events	Ukraine developed as an artificial state

Table 6: Representative cross-group explicit-forwarding examples from non-narrative to narrative-active channels.

Message (English translation)	Share count	Narrative	Sub-narrative
War has been declared today on the Russian world — Vladimir Putin.	2	Actions of Ukraine and the West forced Russia to start the war	Russia was forced to enter the war to ensure its survival
Putin supported Trump's idea of a mutual refusal by Russia and Ukraine for 30 days to strike energy infrastructure and gave such an order to the military — Kremlin.	2	Ukraine and the West refuse to start peace negotiations	Unlike Ukraine, Russia demonstrates readiness for negotiations
Poland does not want simply to be in solidarity with Ukraine; it wants to profit from it. "We will no longer help in a naive way. It will not be that Poland is solidaristic and others profit from the reconstruction of Ukraine," said Polish Prime Minister Tusk.	2	The West controls Ukraine and uses it for its own goals	Western elites are the main beneficiaries of the war

Table 7: Representative cross-group semantic-similarity examples from narrative-active to non-narrative channels.

Message (English translation)	Share count	Narrative	Sub-narrative
In Pokrovsk, almost 15 million were allocated to support the media, 73 million to the water utility, 34 million to beautification, and almost 30 million to the functioning of heat energy, — Telegraf, citing the community budget for 2025. Fifty million is provided for officials' salaries, 98 million for the elimination of emergency situations, and 147 million for education. According to DeepState, the Russian Armed Forces are already 2–3 km from the city. As of the end of the year, 70% of residential buildings, 80% of social facilities, and 95% of industrial facilities had been damaged or destroyed. There is no electricity, gas, heating, or water supply in the city.	3	Discrediting or ridiculing representatives of Ukrainian authorities	Corruption in Ukraine's political and military leadership
Slovakia threatens Ukraine with stopping electricity supplies because of the halt in gas supplies. "If necessary, we will stop electricity supplies, which Ukraine critically needs during power outages," Fico said.	2	Ukraine is left alone and will lose	Partners abandoned Ukraine
The Biden administration informed Trump's team that Ukraine will have to "resolve the issue of lowering the mobilization age," — Sullivan. In his opinion, the shortage of people remains an acute problem while the United States is providing "a huge amount of ammunition and military equipment."	2	The West controls Ukraine and uses it for its own goals	Ukraine is an instrument of the United States

Table 8: Representative cross-group semantic-similarity examples from non-narrative to narrative-active channels.

	Ukrainian version
System instructions	Ти уважний український класифікатор. Поверни ЛИШЕ валідний JSON. Без зайвого тексту. НЕ вигадуй. НЕ використовуй markdown. НЕ перекладай і не перефразовуй текст наративів. У відповіді повертай ТІЛЬКИ narrative_id та sub_narrative_id (без тексту наративів).
Task instructions	Завдання: Визнач, чи повідомлення ПРЯМО або НЕПРЯМО ('натякаючи') просуває будь-який наратив зі списку. Правила: • Спочатку вирішi, чи є хоча б один відповідний наратив. • Відповідай «так» лише якщо твердження у повідомленні прямо підтримує наратив. • Якщо це загальні новини або нейтральний опис подій — НЕ став жодних наративів. • Може бути таке, що повідомлення може містити кілька наративів. Обери той, що найбільше підходить, або найбільш чіткий. • Читай уважно. Дивись на повідомлення з української перспективи. • Використовуй ТІЛЬКИ narrative_id та sub_narrative_id зі списку. • Якщо жоден наратив не підходить — поверни null для narrative_id і sub_narrative_id та confidence 0. • Додай confidence від 0 до 1. • Вивід має бути валідним JSON. • Наприклад: – "кличко заявив що україна стикається з браком солдатів і пропонує знизити вік мобілізації деталі за посиланням" – це повідомлення нав'язує думку, що Кличко хоче знизити мобілізаційний вік. Це страшить громадян і у них виникає погана думка про політика Кличка, тому тут є наратив "Дискредитація чи висміювання представників української влади" – "мужики красавчики понимая что ударные дроны вылетают з контейнера на прицепе фуры начали забрасывать его камнями рискуя своей жизнью конечно они понимали об опасности было бы хорошо если бы их нашли и наградили просто красавчики neoficialniybezsonov" – повідомлення стимулює 'захоплюватись' сміливістю росіян і хоче переконати, що перемога росіян однозначна, тому тут присутній наратив "Перемога України неможлива"

Table 9: Original version of final prompt.

	English version
System instructions	You are a diligent Ukrainian classifier. Return ONLY valid JSON. No extra text. Do NOT make anything up. Do NOT use Markdown. Do NOT translate or paraphrase the narrative text. In your response, return ONLY the narrative_id and sub_narrative_id (without the narrative text).
Task instructions	Task: Determine whether the message DIRECTLY or INDIRECTLY ('by implication') promotes any of the narratives listed below. Rules: • First, decide whether there is at least one relevant narrative. • Answer 'yes' only if the statement in the post directly supports the narrative. • If it is general news or a neutral description of events — DO NOT assign any narratives. • It may be the case that a post contains several narratives. Choose the one that is most appropriate or the clearest. • Read carefully. View the post from a Ukrainian perspective. • Use ONLY narrative_id and sub_narrative_id from the list. • If no narrative is suitable — return null for narrative_id and sub_narrative_id and confidence 0. • Set confidence to a value between 0 and 1. • The output must be valid JSON. • For example: – "Klitschko has stated that Ukraine is facing a shortage of soldiers and is proposing to lower the conscription age – details via the link" – this message seeks to suggest that Klitschko wants to lower the conscription age. This frightens citizens and gives them a negative impression of the politician Klitschko; therefore, there is a narrative "Discrediting or ridiculing representatives of the Ukrainian authorities" – "A bunch of cool guys, realising that attack drones were taking off from a container on the back of a lorry, started pelting it with stones, risking their lives. Of course, they knew the danger involved. It would have been great if they'd been found and rewarded – just a bunch of cool guys. neoficialniybezsonov" – The post encourages people to 'marvel' at the Russians' bravery and aims to convince them that a Russian victory is inevitable, so the narrative here is "A Ukrainian victory is impossible"

Table 10: English version of the prompt used for weak labeling (translated from Ukrainian via DeepL). To ensure reproducibility, please refer to the original Ukrainian prompt in the Table 9.

Professional Translators Versus Quality Estimation Models: Reliability and Agreement in English-Ukrainian Translation Evaluation

Dmytro Chaplynskyi
Ukrainian Catholic University
I. Krypiakevych Institute
of Ukrainian Studies, lang-uk
chaplynskyi.dmytro@ucu.edu.ua

Kyrylo Zakharov
UNHCR
kirillzakharov13@gmail.com

Lesia Ivashkevych
NTUU "Igor Sikorsky Kyiv
Polytechnic Institute"
lesia.ivashkevych@gmail.com

Abstract

We extend a prior study comparing automatic Quality Estimation (QE) models with crowdsourced student judgments for English–Ukrainian parallel corpus evaluation. Eight professional translators each rate 1,000 sentence pairs on a continuous 0–100 scale under one of two paradigms: holistic quality scoring or a two-stage fluency-plus-adequacy protocol, with a repeated task for test–retest reliability. Professionals using the holistic scale achieve significantly higher inter-rater reliability than both linguistics students and professionals using separate fluency and adequacy scales, contradicting the expectation that multidimensional evaluation improves agreement. Adequacy correlates strongly with holistic judgments while fluency emerges as a largely independent dimension. Experts also exhibit a significant leniency drift over the session, alongside increasing evaluation speed. We additionally evaluate three LLMs as translation quality judges (Gemini 3 Flash, GPT-5.4, Gemma 3 27B) and find that the two larger models modestly outperform dedicated QE models in correlation with expert scores ($r = 0.814$–0.821 vs. $r \leq 0.747$). When prompted for separate fluency and adequacy scores, the LLMs replicate the adequacy-dominance pattern, confirming that meaning preservation drives holistic quality perception across both human and machine judges.

1 Introduction

Evaluating the quality of machine translation (MT) output is essential for building and filtering parallel corpora, particularly for low- to mid-resource languages where training data quality directly affects downstream model performance. Automatic Quality Estimation (QE) models predict translation quality from the source–target pair alone, without requiring a human reference; this distinguishes them from reference-based metrics such as BLEU or chrF and makes them practical for corpus filtering at scale. Recent neural QE systems—COMET (Rei et al., 2020), COMETKiwi (Rei et al., 2022, 2023), xCOMET (Guerreiro et al., 2024), and MetricX (Juraska et al., 2023, 2024)—have become standard tools for this task, offering scalable sentence-level quality predictions.

However, the relationship between automatic QE scores and human quality perception is not straightforward. In a prior study (Chaplynskyi and Zakharov, 2025), we applied six QE models to score 55 million English-Ukrainian sentence pairs and conducted a human evaluation of a stratified sample of 9,775 pairs using linguistics students as annotators. The best ensemble model explained approximately 60% of the variance in averaged human ratings, with a non-linear relationship between automatic scores and human perception. Critically, inter-rater agreement among students was only moderate (ICC = 0.43–0.54), raising the question of whether limited agreement reflects genuine subjectivity in quality assessment or insufficient evaluator expertise.

The present paper addresses this question directly. We make four contributions:

1. **Professional evaluation at scale.** We recruit eight professional translators and collect over 8,000 individual evaluations on 1,000 sentence pairs, enabling direct comparison with student data from our prior study.

2. **Holistic versus multidimensional evaluation.** We employ two paradigms in parallel: a single holistic quality score and a two-stage fluency-plus-adequacy protocol (Graham et al., 2013; Lommel et al., 2014). Contrary to expectations, holistic scoring yields higher inter-rater agreement.

3. **Systematic reliability analysis.** We examine inter-rater and test–retest reliability, evaluator learning effects, leniency drift, and the role

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 97–107

of text complexity and evaluation time in explaining score variance.

4. **LLM-as-a-Judge comparison.** We evaluate three LLMs as translation quality judges and find that the two larger models outperform dedicated QE models in agreement with professional translators, while a smaller model does not benefit from structured prompting.

2 Related Work

Human evaluation of MT quality. The dominant paradigms for human MT evaluation include Direct Assessment (DA) on a continuous 0–100 scale (Graham et al., 2013, 2016), the Multidimensional Quality Metrics (MQM) framework based on error annotation (Lommel et al., 2014), and pairwise ranking. DA with continuous scales has been adopted by WMT shared tasks as the standard for collecting human judgments at scale (Freitag et al., 2021, 2022). The separation of fluency and adequacy as distinct evaluation dimensions has a long history in MT evaluation (Castilho et al., 2017; Görög, 2014), though recent work has debated whether holistic scoring produces comparable results with lower annotator burden.

Evaluator expertise. The effect of annotator expertise on MT evaluation reliability remains underexplored. Freitag et al. (2021) showed that expert evaluators produced higher agreement and different system rankings compared to crowdsourced judgments. Our prior study (Chaplynskyi and Zakharov, 2025) used linguistics students, achieving moderate inter-rater reliability (ICC = 0.43–0.54). The present study extends this by recruiting professional translators to test whether domain expertise improves reliability.

Quality estimation models. Neural QE models have advanced rapidly. The COMET family (Rei et al., 2020, 2022, 2023) combines COMET's architecture with OpenKiwi's predictor-estimator setup. xCOMET (Guerreiro et al., 2024) adds error span detection. The MetricX family (Juraska et al., 2023, 2024) uses a two-stage fine-tuning strategy on human-labeled data. These models achieve high agreement with human judgments on high-resource language pairs (Freitag et al., 2022), but their performance on low-resource pairs such as English-Ukrainian is less established.

LLM-as-a-Judge. Recent work has demonstrated that large language models can serve as effective MT evaluators. Kocmi and Federmann (2023) showed that GPT-based evaluation achieves state-of-the-art correlation with human judgments, and Zheng et al. (2023) established the LLM-as-a-Judge paradigm more broadly. We evaluate three LLMs as translation quality judges: Gemma 3 27B (Gemma Team, 2025), Gemini 3 Flash, and GPT-5.4, alongside traditional QE models.

3 Methodology

3.1 Data Sample

The evaluated texts were drawn from the OPUS Open Parallel Corpora (Tiedemann, 2016) and consist of 1,000 English-Ukrainian sentence pairs. Of these, 720 pairs overlap with the sample evaluated by linguistics students in Chaplynskyi and Zakharov (2025), enabling direct cross-study comparison. The remaining 280 pairs were randomly selected from the same corpus under the constraint that they were not corrupted and did not contain inappropriate or sensitive content.

3.2 Expert Recruitment

Nine professional translators were recruited, all with a minimum of three years of professional experience translating from English into Ukrainian. Candidates were identified through professional networks, with preference for translators whose reliability could be vouched for by colleagues. The translators' professional backgrounds span technical, legal, literary, media, marketing, and military translation domains, providing diversity in evaluation perspectives.

Translators were randomly assigned to two groups: four evaluated using a holistic quality scale (Group 1), and five evaluated using separate fluency and adequacy scales (Group 2). One translator in Group 2 completed only 39 evaluations and was excluded from the analysis, leaving four active evaluators per group and eight professional translators in total.

3.3 Evaluation Protocol

Evaluations were collected using Vulyk,[1] an open-source crowdsourcing platform, with two task plugins developed for this study: one for holistic transla-

[1] https://github.com/mrgambal/vulyk/

tion evaluation[2] and one for the two-stage fluency-and-adequacy protocol.[3] Twenty evaluation tasks were constructed, each containing 50 translation pairs sampled without replacement from the pool of 1,000 pairs. The order of pairs within each task was randomized. To assess test–retest reliability, the first task was repeated at the end of the evaluation sequence. Before beginning, all evaluators were shown five worked examples. Breaks were permitted between tasks but not during a 50-pair task. Timestamps were recorded for each evaluation from presentation to submission.

Holistic evaluation (Group 1). Evaluators were presented simultaneously with the English source text and its Ukrainian translation and instructed: "Rate the translation (0–100)." Ratings were provided using a continuous slider with a red-to-green color gradient and the following descriptors: 0–10 *Incorrect translation*, 11–29 *A few correct keywords, but the meaning is different*, 30–50 *Major mistakes in translation*, 51–69 *Understandable but contains typos or grammatical errors*, 70–90 *Preserves semantics closely*, 91–100 *Perfect translation*. This replicates the protocol used for students in the prior study (Chaplynskyi and Zakharov, 2025).

Fluency and adequacy evaluation (Group 2). Evaluators assessed translations in two successive stages. In the fluency stage, only the Ukrainian translation was displayed, and evaluators rated linguistic quality on a 0–100 scale (0–25 *Incomprehensible*, 25–50 *Disfluent*, 50–75 *Good*, 75–100 *Flawless*). After completing the fluency rating, the English source was revealed, and evaluators rated adequacy: "How much of the meaning expressed in the source text is also expressed in the target translation?" (0–25 *None*, 25–50 *Little*, 50–75 *Most*, 75–100 *All*). The sequential design prevents the source text from biasing fluency judgments.

3.4 Automatic Quality Estimation

Machine translation quality was assessed using nine automatic metrics. Six are dedicated QE models from two families: the COMET family (wmt22-cometkiwi-da, wmt23-cometkiwi-da-xl, wmt23-cometkiwi-da-xxl, and xCOMET-XXL) and the MetricX family (MetricX-23 and MetricX-24). We also include bicleaner-ai (Zaragoza-Bernabeu et al., 2022), a parallel-corpus cleaning classifier trained to flag noisy sentence pairs; cosine similarity of LaBSE sentence embeddings (Feng et al., 2022), a multilingual sentence encoder that supports cross-lingual semantic similarity; and Gemma 3 27B (Gemma Team, 2025) as an LLM-as-a-Judge baseline using a holistic scoring prompt matching the human evaluation rubric.

MetricX scores were rescaled from their native 0–25 inverted scale to 0–1 using $\text{score}_{\text{adj}} = 1 - \text{score}/25$. Gemma 3 scores were rescaled from 0–100 to 0–1.

3.5 Statistical Analysis

Expert scores were analyzed in raw, z-score-normalized, and percentile-rank-transformed forms. We report results primarily on z-score-normalized data, as this transformation yielded the most precise reliability estimates.

Inter-rater reliability. We summarised inter-rater agreement using the Intraclass Correlation Coefficient (ICC), which expresses the share of total score variance attributable to genuine differences between sentence pairs as opposed to disagreement among evaluators. We estimated ICC with a two-way random-effects, absolute-agreement, single-measures model (Shrout and Fleiss, 1979; Koo and Li, 2016) and interpreted values following Cicchetti (1994): below 0.40 = poor, 0.40–0.59 = fair, 0.60–0.74 = good, 0.75–1.00 = excellent.

Test–retest reliability. For individual evaluators, test–retest ICC was computed on the 50 pairs evaluated twice during the first and last task (Gisev et al., 2013). Systematic bias was assessed using paired t-tests.

Mixed-effects models. To test hypotheses about evaluation dynamics, we fitted linear mixed-effects models using lme4 (Bates et al., 2015), with random intercepts for sentence pairs and evaluators. Evaluations exceeding 120 seconds were excluded as likely reflecting breaks rather than sustained attention.

Text complexity. Readability indices (ARI, Coleman-Liau, FORCAST, nWS, RIX) and lexical diversity measures were computed on the English source texts using quanteda.

[2] https://github.com/lang-uk/vulyk-translations

[3] https://github.com/lang-uk/vulyk-fluency-adequacy

Expert	Type	N	Mean	Med.	SD
Exp. 1	Student	709	9.8	7.8	7.9
Exp. 2	Student	707	12.6	9.7	11.0
Exp. 3	Student	705	29.5	25.5	18.0
Exp. 4	Holistic	1017	15.4	12.1	13.4
Exp. 5	Holistic	911	43.5	37.4	26.7
Exp. 6	Holistic	933	19.9	16.0	16.0
Exp. 7	Holistic	889	41.5	33.6	27.2
Exp. 8	F&A	1009	28.0	24.8	16.6
Exp. 9	F&A	924	38.6	29.1	28.0
Exp. 10	F&A	993	34.4	30.8	17.9
Exp. 12	F&A	1011	20.3	16.8	14.3

Table 1: Evaluation duration in seconds per evaluator (evaluations $<$120 s). F&A = fluency and adequacy.

Expert	Scale	ICC	95% CI	Bias
Exp. 4	Holistic	0.21	(−0.07; 0.46)	−10.5
Exp. 5	Holistic	0.55	(0.32; 0.72)	16.7
Exp. 6	Holistic	0.58	(0.36; 0.74)	4.9
Exp. 7	Holistic	0.83	(0.72; 0.90)	12.2
Average	Holistic	0.85	(0.75; 0.91)	5.8
Exp. 8	Fluency	0.69	(0.51; 0.81)	5.8
Exp. 9	Fluency	0.78	(0.64; 0.87)	17.6
Exp. 10	Fluency	0.65	(0.46; 0.79)	18.1
Exp. 12	Fluency	0.76	(0.61; 0.86)	7.2
Average	Fluency	0.86	(0.77; 0.92)	12.2
Exp. 8	Adequacy	0.62	(0.42; 0.77)	9.6
Exp. 9	Adequacy	0.62	(0.42; 0.77)	12.6
Exp. 10	Adequacy	0.46	(0.21; 0.66)	16.6
Exp. 12	Adequacy	0.63	(0.43; 0.77)	14.7
Average	Adequacy	0.79	(0.66; 0.88)	13.4

Table 2: Test–retest reliability for individual experts (50 repeated pairs). Bias = mean score increase from first to last session. All paired t-tests significant ($p < 0.05$).

Group	ICC	95% CI	
Experts (holistic)	0.72	0.69	0.74
Adequacy	0.66	0.64	0.69
Fluency	0.59	0.56	0.62
Students	0.57	0.53	0.61
COMETKiwi+wmt22	0.86	0.84	0.87
xCOMET+MetricX-24	0.80	0.77	0.83
Machines (no bicleaner)	0.69	0.67	0.72
LLM judges	0.82	0.76	0.86
LLM fluency	0.68	0.57	0.75
LLM adequacy	0.80	0.72	0.85

Table 3: Group-level ICC (two-way, agreement, single measures) on z-score-normalized ratings.

4 Results

4.1 Descriptive Statistics

The dataset comprises 10,566 individual evaluations: 2,127 from three linguistics students (prior study), 4,200 from four professional translators using the holistic scale, and 4,239 from four professional translators using the fluency and adequacy scales. Table 1 summarizes the evaluation time by evaluator.

Median evaluation times range from 7.8 s (fastest student) to 37.4 s (slowest holistic expert). Professional translators in both conditions are generally slower than students, consistent with more deliberate evaluation.

4.2 Test–Retest Reliability

Individual test–retest ICC values, computed on the 50 pairs evaluated in both the first and last task, are reported in Table 2. For all experts, the time between the first and the last batch was no less than six days, except for Expert 12, who completed all tasks within two days. Previous methodological research suggests that test–retest intervals are typically chosen to balance recall bias and true change, most commonly ranging from a few days to approximately two weeks (Marx et al., 2003). Nevertheless, the re-test ICC for Expert 12 was not statistically different from the ICCs of the other experts.

Several patterns emerge. First, Expert 4 demonstrates poor test–retest reliability (ICC = 0.21, CI includes zero), suggesting inconsistent scoring behavior. Second, all experts show a positive bias—scores in the repeated task are higher than in the initial task—indicating a systematic leniency drift over the evaluation session. Third, averaged expert scores yield good to excellent reliability (ICC = 0.79–0.86), substantially exceeding individual reliability, confirming that aggregation across evaluators stabilizes judgments. Fourth, fluency test–retest reliability (average ICC = 0.86) is numerically higher than adequacy (0.79), though the difference is not statistically significant given overlapping confidence intervals.

4.3 Inter-Rater Reliability

We computed group-level ICC for each evaluator category using standard score ratings (z-score), which we determined to yield the most precise reliability estimates across normalization methods tested (raw, percentile, rank-based inverse normal transformation).

Figure 1 and Table 3 report the group-level ICC values for each evaluator category using z-score-normalized scores.

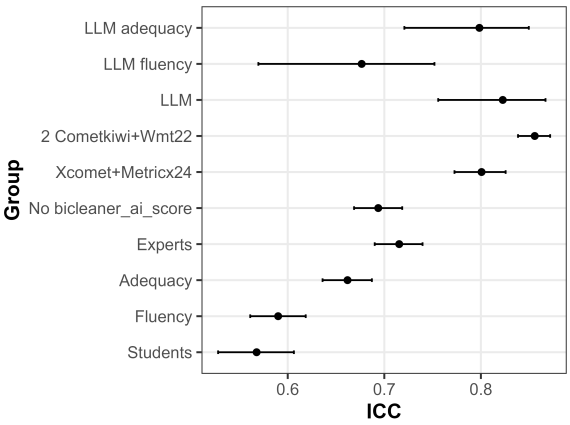


Figure 1: ICC with 95% confidence intervals for evaluator groups (z-score normalized).

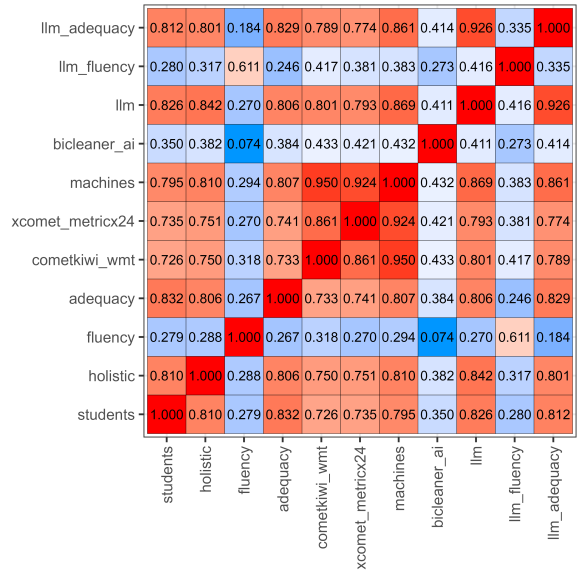


Figure 2: Correlation matrix between averaged group scores and machine model scores (z-score normalized).

Hypothesis 1: Professional translators show higher inter-rater reliability than students. Professional translators using the holistic scale achieved significantly higher inter-rater agreement (ICC = 0.72, 95% CI 0.69–0.74) than linguistics students (ICC = 0.57, 95% CI 0.53–0.61): the confidence intervals do not overlap. This confirms that evaluator expertise matters for translation quality assessment and that the moderate agreement observed in Chaplynskyi and Zakharov (2025) was at least partly attributable to evaluator inexperience rather than inherent task subjectivity.

Hypothesis 2: Multidimensional evaluation yields higher agreement. Contrary to our hypothesis, the holistic single-score evaluation (ICC = 0.72) yielded higher inter-rater agreement than either the adequacy (ICC = 0.66) or fluency (ICC = 0.59) scales. The adequacy ICC confidence interval (0.64–0.69) does not overlap with the holistic group (0.69–0.74), confirming that holistic evaluation produces significantly higher agreement. Fluency shows the lowest agreement among professional groups. This finding suggests that decomposing quality judgment into separate dimensions does not improve—and may slightly reduce—evaluator consistency, possibly because the cognitive task of isolating fluency from adequacy introduces additional judgment uncertainty.

Leave-one-out analysis. Leave-one-out ICC analysis confirmed that Expert 4 is an outlier: removing this evaluator substantially increases the holistic group's ICC. Expert 11 (already excluded due to insufficient evaluations) would similarly degrade the fluency-adequacy group if included. Among machine models, bicleaner-ai exhibits a distinct scoring pattern; excluding it increases the machine group ICC, consistent with its low correlation with other metrics observed in Chaplynskyi and Zakharov (2025).

4.4 Holistic Versus Multidimensional Scores

Hypothesis 3: Holistic scores correlate with adequacy more than with fluency. The correlation analysis confirms this hypothesis (Figure 2). Averaged holistic expert scores show a strong correlation with averaged adequacy scores ($r = 0.806$), while the correlation with fluency scores is notably weaker ($r = 0.288$). This indicates that when translators assign a single quality score, they weight meaning preservation (adequacy) more heavily than linguistic surface quality (fluency).

Fluency emerges as a largely independent dimension. Its highest correlation with any other group is only $r = 0.318$ (with COMETKiwi+wmt22), while even fluency–adequacy correlation is low ($r = 0.267$). This raises the question of whether fluency, as measured here, captures a quality dimension that is relevant to translation quality assessment in the context of parallel corpus filtering.

Student and expert holistic scores correlate strongly ($r = 0.810$), but expert scores show slightly stronger correlation with machine model scores ($r = 0.810$ vs. $r = 0.795$ for students), suggesting that professional judgment aligns more closely with what QE models capture.

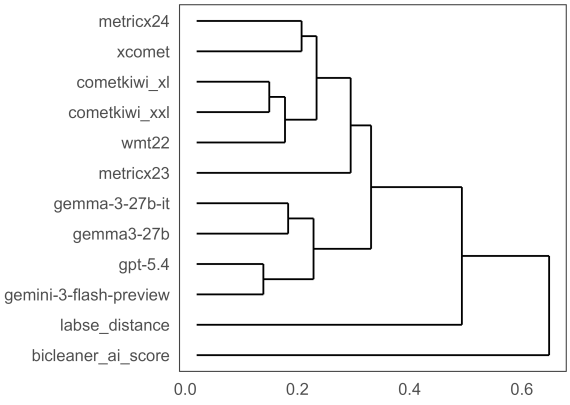


Figure 3: Hierarchical clustering of machine QE models and LLM judges based on Spearman correlation distance (z-score normalized scores).

4.5 Correlation with Automatic Metrics

The same non-linear (quadratic) relationship between machine and expert scores observed in Chaplynskyi and Zakharov (2025) persists in the professional evaluation data: QE models and human experts diverge most for translations of moderate quality, where automatic scores tend to be more optimistic than human judgments.

Among the machine models, the COMET family and xCOMET show the strongest correlations with expert holistic scores ($r = 0.750$ for COMETKiwi+wmt22, $r = 0.751$ for xCOMET+MetricX-24), while bicleaner-ai demonstrates weak correlations with all human and machine scores ($r = 0.382$ with experts). The hierarchical clustering of machine model scores (Figure 3) reveals two distinct clusters: one containing the COMETKiwi variants and wmt22-cometkiwi-da, and another containing xCOMET and MetricX-24. This clustering motivated our decision to compute group-level ICC separately for these model families (Table 3), as treating all machine models as a single group would conflate metrics that capture partially different aspects of translation quality.

An additional cluster composed of the LLM judges sits further from the traditional QE metrics, suggesting that the divergence reflects not only model differences but also the evaluation paradigm: LLM-based judgments depend on instruction-following and prompt design, introducing additional variability compared with regression-based QE metrics.

Model	N	r	**Mean**
LLM-as-a-Judge (rubric prompt, 0–100)			
Gemini 3 Flash	999	0.821	75.7
GPT-5.4	1000	0.814	68.1
Gemma 3 27B	1000	0.722	76.3
QE models			
Gemma 3 27B (QE)	710	0.747	—
COMETKiwi-XXL	710	0.740	—
xCOMET	710	0.735	—
Expert avg	1000	—	70.9

Table 4: Spearman correlation (r) of automatic scores with averaged expert holistic scores. Gemma 3 27B appears twice: as a QE baseline from the prior study (simple scoring, rescaled 0–1) and as an LLM-as-a-Judge with an explicit evaluation rubric (0–100 scale, temperature 0.3).

4.6 LLM-as-a-Judge

We evaluated three LLMs as translation quality judges on the same 1,000 sentence pairs: Gemma 3 27B (Gemma Team, 2025), Gemini 3 Flash, and GPT-5.4.[4] Each model scored all pairs on the 0–100 holistic scale. Gemini 3 Flash and GPT-5.4 additionally produced fluency and adequacy scores for all 1,000 pairs.

Table 4 reports the Spearman correlations between LLM scores and averaged expert holistic scores, with both variables converted to z-scores and means computed from raw scores. Gemini 3 Flash ($r = 0.821$) and GPT-5.4 ($r = 0.814$) slightly outperform all dedicated QE models in their baseline configuration from the prior study ($r = 0.747$). Notably, the Gemma 3 27B model yields consistent results across prompts, with $r = 0.747$ reported in (Chaplynskyi and Zakharov, 2025) and $r = 0.722$ in the current study. This suggests that a more detailed prompt with an explicit evaluation rubric (LLM-as-a-Judge) does not improve performance, indicating that a structured scoring prompt does not necessarily benefit smaller models.

GPT-5.4 is the best-calibrated model, with a mean score (68.1) closest to the expert average (70.9) and a score distribution that closely matches the expert distribution. Gemini and Gemma both exhibit the leniency bias observed across all LLM evaluators.

Fluency and adequacy raw scores from all three LLMs (1,000 pairs each) replicate the key patterns observed in human evaluation observed with

[4] All models were prompted with the same holistic scoring rubric used for human evaluators, with temperature set to 0.3.

Spearman correlations. LLM adequacy correlates strongly with expert holistic scores (Gemini: $r = 0.772$; GPT-5.4: $r = 0.759$; Gemma: $r = 0.704$) and expert adequacy (Gemini: $r = 0.809$; GPT-5.4: $r = 0.799$; Gemma: $r = 0.694$), while LLM fluency correlates weakly with expert holistic scores ($r = 0.227$–0.326) and shows low correlation with expert adequacy ($r = 0.174$–0.250). The internal fluency–adequacy correlations for Gemini ($r = 0.263$) and GPT-5.4 ($r = 0.288$) are comparable to human experts ($r = 0.267$), confirming that the models separate these dimensions similarly to professionals. Gemma shows weaker separation ($r = 0.447$). GPT-5.4 is also the best-calibrated model on the adequacy scale (mean 75.2 vs. expert 73.3).

Pairwise agreement among LLMs holistic scores is high (Gemini–GPT: $r = 0.861$; Gemma–GPT: $r = 0.798$; Gemini–Gemma: $r = 0.749$), indicating substantial convergence despite different architectures and training data.

4.7 Text Complexity and Score Variance

Hypothesis 4: Text complexity is positively correlated with score variance. We find no evidence for this hypothesis. Correlations between text complexity measures (readability indices, lexical diversity, sentence length) and the standard deviation of expert scores across evaluators are weak and inconsistent across evaluator groups. The only notable observation is that longer source texts are associated with lower fluency ratings, possibly because longer sentences provide more opportunities for grammatical or stylistic issues.

Hypothesis 7: Text complexity affects human scores more than machine scores. There is no strong evidence for differential impact. Text complexity measures show similarly weak relationships with score variance for both human and machine evaluators.

4.8 Evaluation Dynamics

Hypothesis 5: Experts assign higher scores over time. A linear mixed-effects model predicting the holistic score from evaluation order (with random intercepts for sentence pair and evaluator; $N = 5{,}871$ evaluations from 7 experts) reveals a small but significant positive effect of order on scores ($\beta = 0.006$, $SE = 0.002$, $t = 3.94$, $p < 0.001$). Over the course of approximately 1,000 evaluations, this corresponds to a cumulative increase of roughly 6 points on the 0–100 scale. This leniency drift is consistent with the positive biases observed in the test–retest analysis (Table 2) and suggests that experts become more lenient as they progress through the evaluation.

Hypothesis 6: Evaluation duration decreases over time. A mixed-effects model predicting evaluation duration from order ($N = 9{,}846$ evaluations from 12 evaluators) confirms a significant learning effect ($\beta = -0.012$, $SE = 0.001$, $t = -12.79$, $p < 0.001$). Experts become faster at the evaluation task as they gain experience, with the duration reduction amounting to approximately 12 seconds over 1,000 evaluations. This pattern is consistent across both evaluation conditions and indicates genuine task learning—experts develop more efficient evaluation strategies over time.

The combination of increasing speed and increasing leniency suggests a form of evaluator fatigue or habituation: as the task becomes more routine, experts spend less time on each pair and default to higher scores, potentially reflecting reduced attention to translation errors.

5 Discussion

The present study extends Chaplynskyi and Zakharov (2025) by introducing professional translators as evaluators and comparing holistic and multidimensional evaluation paradigms for English-Ukrainian translation quality assessment. Our findings offer several insights with implications for both MT evaluation methodology and practical corpus filtering.

Expertise matters, but method matters more. Professional translators using a holistic scale achieve substantially higher inter-rater reliability than linguistics students evaluating the same translations with the same interface and instructions. This confirms that the moderate agreement reported in our prior work was partly an artifact of evaluator inexperience. However, the improvement is specific to the holistic condition: professionals using separate fluency and adequacy scales do not clearly outperform students. This suggests that evaluation reliability depends on the interaction between evaluator expertise and task design, not on expertise alone.

The fluency-adequacy decomposition does not help. Our most surprising finding is that the multidimensional evaluation paradigm, despite its the-

oretical appeal and widespread use in MT evaluation, does not improve inter-rater agreement over simple holistic scoring. One possible explanation is that fluency and adequacy are not orthogonal for web-crawled parallel data of the type in our sample: most low-quality translations fail on both dimensions simultaneously (garbled text is neither fluent nor adequate), while most high-quality translations succeed on both. The decomposition may add value primarily for translations where fluency and adequacy diverge—a relatively rare case in naturally occurring parallel corpora. The fluency evaluation also carries the additional cognitive burden of evaluating a text without its source, which may introduce uncertainty.

Adequacy dominates quality perception. When translators assign a single quality score, their judgment aligns strongly with adequacy (meaning preservation) and only weakly with fluency (linguistic surface quality). This finding has practical implications for corpus filtering: if the goal is to predict human quality judgments, adequacy-oriented metrics may be more informative than fluency-oriented ones. It also suggests that the fluency dimension, at least as measured with the scale descriptors used here, captures something that human evaluators do not strongly weight when making holistic quality judgments about parallel corpus data.

Evaluator drift is real and should be monitored. The significant leniency drift we observe—experts assigning progressively higher scores over the evaluation session—is a practical concern for any large-scale human evaluation campaign. The concurrent decrease in evaluation time suggests that the drift reflects reduced engagement rather than genuine recalibration. Future evaluation protocols should consider randomizing the presentation order more aggressively, inserting calibration anchors throughout the session, or normalizing scores within evaluation blocks to mitigate this effect.

LLMs as viable replacements for QE models. The two largest LLM judges—Gemini 3 Flash ($r = 0.821$) and GPT-5.4 ($r = 0.814$)—outperform all dedicated QE models in correlation with expert holistic scores, both individually and as an averaged ensemble ($r \leq 0.810$). Gemma 3 27B achieves slightly lower performance as an LLM-as-a-Judge ($r = 0.722$) and in its QE configuration ($r = 0.747$), suggesting that structured prompting does not benefit smaller models. GPT-5.4 achieves the closest calibration to human experts (mean 68.1 vs. 70.9), while Gemini achieves the highest correlation. The LLMs also replicate the adequacy-dominance pattern: their QE scores correlate strongly with expert adequacy ($r = 0.69$–0.79) but weakly with expert fluency ($r = 0.22$–0.26), confirming that holistic quality perception—whether by humans or LLMs—is primarily driven by meaning preservation. When prompted for separate fluency and adequacy scores, Gemini and GPT-5.4 achieve fluency–adequacy separations ($r = 0.26$–0.29) comparable to human experts ($r = 0.26$), indicating that LLMs can meaningfully decompose translation quality into independent dimensions.

Implications for corpus filtering. The non-linear relationship between QE model scores and expert judgments persists when moving from student to professional evaluators, confirming that this is a genuine feature of the QE-human relationship rather than an artifact of student evaluation quality. For practical corpus filtering, this reinforces the recommendation from Chaplynskyi and Zakharov (2025) to use non-linear ensemble models rather than raw QE scores when estimating human quality perception.

6 Conclusion

We presented a systematic comparison of professional human evaluation and automatic quality estimation for English-Ukrainian machine translation, testing seven hypotheses. Our key findings are:

- Professional translators using holistic scoring achieve significantly higher inter-rater reliability than linguistics students, confirming that evaluator expertise improves the quality of human reference data for MT evaluation.
- Contrary to expectations, holistic evaluation outperforms the fluency-adequacy decomposition in terms of inter-rater agreement, suggesting that simpler evaluation protocols may be preferable for corpus-level quality assessment.
- Adequacy strongly predicts holistic quality judgments, while fluency is a largely independent dimension—indicating that meaning preservation is the dominant factor in how translators perceive overall translation quality.

- Experts exhibit a significant leniency drift (higher scores over time) coupled with faster evaluation, pointing to habituation effects that should be accounted for in evaluation design.

- LLM-as-a-Judge evaluation with Gemini 3 Flash ($r = 0.821$) and GPT-5.4 ($r = 0.814$) outperforms all dedicated QE models in correlation with expert scores. All three LLMs replicate the adequacy-dominance and fluency-independence patterns observed in human evaluation.

For future work, we plan to: (1) evaluate the downstream impact of corpus filtering on NMT model performance; (2) investigate whether LLM-as-a-Judge models can fully replace human evaluation for corpus quality assessment at scale; and (3) apply the framework to other language pairs. We release the full sample of 1,000 evaluated English–Ukrainian sentence pairs together with all individual expert and LLM ratings[5] and the analysis code[6] to support reproducibility and external validation.

Limitations

- The study focuses on a single language pair (English-Ukrainian), and results may not generalize to other pairs with different morphological or resource characteristics.

- The sample of 1,000 sentence pairs, while sufficient for statistical analysis, represents a small fraction of the 55 million pairs in the full corpus.

- Professional translators may still exhibit domain-specific biases (e.g., technical vs. literary).

- One evaluator completed insufficient evaluations and was excluded; another (Expert 4) showed poor test–retest reliability (ICC = 0.21), suggesting that professional status alone does not guarantee evaluation quality.

- The fluency-adequacy evaluation inherently takes longer per pair (two ratings), which may introduce differential fatigue effects not present in the holistic condition.

- We acknowledge that the sequential fluency-then-adequacy design may introduce an anchoring effect, potentially influencing the independence of the two judgments. However, the observed low correlation suggests that any such bias did not artificially inflate agreement and is unlikely to have driven the main results.

- The evaluated sample is drawn from the publicly available OPUS corpus, which predates the training cutoffs of the LLMs we use as judges. We cannot rule out that some sentence pairs were seen during LLM pre-training, and a degree of memorization-driven inflation of LLM-as-a-Judge scores is therefore possible. The modest gap between the best LLM judge and the strongest dedicated QE model ($\Delta r \leq 0.07$ in Spearman correlation) makes contamination unlikely to be the sole driver of the observed advantage, but the effect cannot be quantified from the data available here.

Ethical Considerations

This study involves human evaluation of translation quality by professional translators who were compensated for their work. All evaluators participated voluntarily and were informed about the purpose of the study.

Parts of the codebase (data processing, analysis scripts, and the crowdsourcing platform plugins) were developed with the assistance of Claude Code (Anthropic), an AI-based coding tool. Claude Code was also used as a writing aid during the preparation of this manuscript. All AI-generated content was reviewed and edited by the authors. The LLM-as-a-Judge evaluation prompts are provided in Appendix A for reproducibility.

Acknowledgments

We thank the professional translators who contributed their time and expertise to this evaluation: Anton Shpigunov, Vladislav Demyanov, Kristina Zayka, Anatolii Zhylavyi, Oleksandra Vankevych, Faina Zholobak, Olena Pantsyr, and Olga Prokopchuk.

References

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

[5] https://huggingface.co/datasets/lang-uk/qe-vs-human

[6] https://github.com/Amice13/translation-quality

Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017. Is neural machine translation the new state of the art? In *The Prague Bulletin of Mathematical Linguistics*, volume 108, pages 109–120.

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

Domenic V. Cicchetti. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4):284–290.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation*, pages 46–68. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774. Association for Computational Linguistics.

Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Natasa Gisev, J Simon Bell, and Timothy F Chen. 2013. Interrater agreement and interrater reliability: key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, 9(3):330–338.

Attila Görög. 2014. Quantifying and benchmarking quality: The TAUS dynamic quality framework. *Tradumatica: Tecnologies de la Traducció*, (12):443–454.

Yvette Graham, Timothy Baldwin, Meghan Dowling, Maria Eskevich, Teresa Lynn, and Lamia Tounsi. 2016. Is all that glitters in machine translation quality estimation really gold? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3124–3134, Osaka, Japan. The COLING 2016 Organizing Committee.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203. European Association for Machine Translation.

Terry K. Koo and Mae Y. Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163.

Arle Richard Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumatica: Tecnologies de la Traducció*, (12):455–463.

Robert G Marx, Alia Menezes, Lois Horovitz, Edward C Jones, and Russell F Warren. 2003. A comparison of two time intervals for test-retest reliability of health status instruments. *Journal of clinical epidemiology*, 56(8):730–735.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference*

on Empirical Methods in Natural Language Processing (EMNLP), pages 2685–2702. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 634–645. Association for Computational Linguistics.

Patrick E. Shrout and Joseph L. Fleiss. 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428.

Jörg Tiedemann. 2016. OPUS – parallel corpora for everyone. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation (EAMT)*, page 384.

Jaume Zaragoza-Bernabeu, Gema Ramírez-Sánchez, Marta Bañón, and Sergio Ortiz Rojas. 2022. Bicleaner AI: Bicleaner goes neural. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 824–831. European Language Resources Association.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*.

A LLM-as-a-Judge Prompts

Holistic QE prompt. All three LLMs received the following system prompt for holistic quality estimation:

> You are a professional English-to-Ukrainian translation evaluator. Rate the translation quality on a 0–100 scale using these descriptors: 0–10 Incorrect translation; 11–29 A few correct keywords, but the meaning is different; 30–50 Major mistakes in translation; 51–69 Understandable but contains typos or grammatical errors; 70–90 Preserves semantics closely; 91–100 Perfect translation. Return only a JSON object with a single "score" field.

Fluency and adequacy prompt. For the two-stage evaluation, the following prompt was used:

> You are a professional English-to-Ukrainian translation evaluator. First, evaluate the fluency of the Ukrainian text (0–100): 0–25 Incomprehensible; 25–50 Disfluent; 50–75 Good; 75–100 Flawless. Then, evaluate how much of the meaning from the English source is preserved in the Ukrainian translation (0–100): 0–25 None; 25–50 Little; 50–75 Most; 75–100 All. Return only a JSON object with "fluency" and "adequacy" fields.

All models were called with temperature 0.3. The English source text and Ukrainian translation were provided as user input.

Belief Propagation in LLM World Models: Measuring Strategic Information Bias with Prediction Markets

Mykola Khandoga[1], **Yevhen Kostiuk**[1,2], **Anton Polishko**[1], **Yurii Filipchuk**[1], **Kostiantyn Kozlov**[1], **Dmytro Zamriy**[1], **Artur Kiulian**[1]
[1]Future Principle [2]Aarhus University

Abstract

Every information ecosystem produces beliefs that shape strategic decisions. Both human analysts and AI systems inherit the blind spots of their information sources. We show that LLMs, combined with prediction markets, function as a calibrated instrument for measuring how far ecosystem-induced beliefs deviate from an external reference: LLMs extract the beliefs a text corpus implies, and prediction market price trajectories – anchored at resolution by realised outcomes – provide the calibration reference against which to quantify the deviation.

We isolate the bias contribution of specific text through ablation: varying information context while holding the model fixed, with a contaminated model that knows actual outcomes as control. Applied to 111 Ukraine-related prediction markets (~93,000 predictions, four models), we find that English news context systematically biases territorial predictions, wrong 64–72% of the time ($p < 10^{-6}$). A contaminated model that knows actual outcomes shows the same error rate, indicating the bias originates primarily in the text. Supplementing with Ukrainian military-analytical sources reduces the bias for all clean models; absolute-error gains are partial and model-dependent.

We show that the distortion originates primarily in the sources, not the models. Consistent across four architectures, it will persist in any system that processes them and propagate into downstream decisions.

1 Introduction

The beliefs propagated by news coverage about ongoing events have major consequences for policy, public opinion, and resource allocation. Yet there exists no method to quantify how close they are to the public consensus. Existing approaches either detect framing properties of text without measuring their downstream cost (Ali and Hassan, 2022; Otmakhova et al., 2024), or evaluate LLM forecasting accuracy without analyzing the information diet that drives it (Karger et al., 2025; Halawi et al., 2024).

Closing this gap requires solving two problems. First, we need a model that internalizes discourse framing – not classifying frames from the outside, but absorbing them so its output reflects the belief the text induces. Second, we need a grounded scale against which to measure that belief – an external reference anchored by realised outcomes, not another model's opinion.

LLMs solve the first problem. In-context learning operates as implicit Bayesian inference over latent concepts in the input (Xie et al., 2022), mechanistically equivalent to gradient descent on internal representations (von Oswald et al., 2023). The model doesn't just read the text – it updates its beliefs toward what the text implies. The output probability is the induced belief.

Prediction markets solve the second. A belief without an external reference is just an opinion. Prediction markets (Wolfers and Zitzewitz, 2004, 2006) provide continuous, financially incentivized probability estimates that are eventually anchored by realised outcomes. We use them in two distinct roles: the binary resolution gives us a low-power but genuine ground truth (§4.1), while the continuous price trajectory serves as the calibration reference for all model comparisons. The delta between the LLM-induced belief and market price, measured in percentage points (pp), is our calibrated measure of framing cost.

Our goal is to measure the bias that a specific text corpus induces – how many pp closer to or further from the market estimate does this text push the model's prediction? An LLM prediction reflects both parametric knowledge from pretraining and in-context beliefs induced by the prompt. To isolate each source's contribution, we construct an ablation ladder:

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 108–120

- **A** provides market overview and current price data – the minimal context from which the model reasons using only parametric knowledge.
- **B** adds a price chart: structured numerical signal without narrative framing.
- **C** adds English news articles: narrative without the rest of the context.
- **D** combines all English-language sources – chart, news, war map, trader comments – representing the full English information ecosystem.
- $\mathbf{D_{UA}}$ supplements D with Ukrainian military sources: General Staff reports, frontline bloggers, defense media.

Each transition is a measurement in pp: A→B measures the value of enriched price history signal, A→C the cost of English news narrative alone, A→D the cost of the full English ecosystem, D→D_{UA} the value of Ukrainian source diversification. A formal framework is given in Appendix A; Appendix H reports an exploratory decomposition of parametric and context-induced bias components.

Our contributions are:

1. A method for measuring the bias a text corpus induces via belief propagation in LLMs, in calibrated probability units, validated against prediction markets;
2. An ablation structure that isolates parametric from context-induced bias, revealing that the English information ecosystem systematically distorts LLM predictions on territorial questions – a finding confirmed by a contaminated model control and by linguistic analysis of reasoning traces (offense-dominant verb framing, asymmetric counterfactual reasoning) – and that supplementing with Ukrainian military-analytical sources partially counterbalances the bias;
3. A dataset of ∼93,000 predictions with full reasoning traces.[1]

2 Related Work

LLMs as forecasters. ForecastBench (Karger et al., 2025) shows LLMs now outperform non-expert crowds; agentic systems have reached superforecaster-level performance (Alur et al., 2025). We depart from this line entirely: rather than benchmarking accuracy, we exploit LLMs' sensitivity to linguistic framing as a measurement instrument. Our models need not be good forecasters – they need to faithfully reflect what the text implies.

Computational framing analysis. Media framing shapes public understanding of conflict (Entman, 2004), and NLP work on factuality and bias of news media is surveyed in Nakov et al. (2024). Methods range from codebook annotation (Card et al., 2015) to LLM-based classification of political bias in conflict coverage (Baly et al., 2020; Chandra et al., 2026). Offense-dominant framing in Western coverage of Ukraine has been documented qualitatively (Ojala et al., 2024) and through topic modeling (Ptaszek et al., 2024). These approaches classify frames – detecting what framing exists. Our method measures what that framing costs, in calibrated probability units.

LLM bias. Biases in LLMs are documented across demographic dimensions (Gallegos et al., 2024; Feng et al., 2023). These treat the model as the object of study. Our method measures how biased text propagates *through* models into calibrated beliefs – the model is the instrument, not the subject.

3 Data and Methods

3.1 Dataset

We collected 111 Ukraine-related prediction markets from Polymarket[2] (January 2025 – January 2026): 65 territorial ("Will side X control [city] by [date]?") and 46 diplomatic (negotiations, sanctions, Zelenskyy suit). For each market, we constructed rolling prediction points at ∼8 cutoff dates with a median gap of ∼16 days, creating instances where we know the current price, the actual price at each horizon (6h, 12h, 1d, 2d, 3d, 5d and 7d), and all information available up to the cutoff.

Each prediction is made under five information conditions: **A** (market overview + current price data only), **B** (A + price chart image), **C** (A + English news blocks), **D** (A + chart + English news + war map + Polymarket trader comments), and $\mathbf{D_{UA}}$ (D supplemented with Ukrainian-language military sources: General Staff casualty reports, Telegram military bloggers, Militarnyi). Conditions are identical across models; the contaminated model (Gemini 3.1 Pro Preview) runs all conditions except D_{UA}.

[1] https://huggingface.co/datasets/OpenBabylon/unlp-ukraine-forecasting

[2] https://polymarket.com

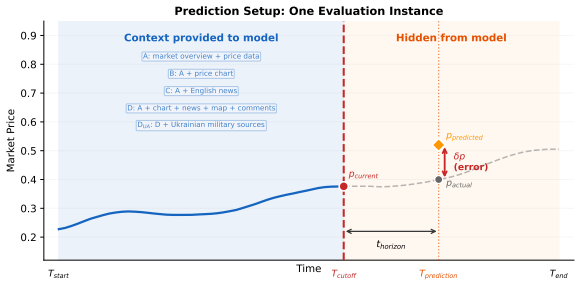


Figure 1: One evaluation instance. The model receives context up to T_{cutoff} under varying information conditions (A–D_{UA}); we compare its prediction at $T_{\text{prediction}}$ against the actual market price.

The English news corpus linked to our 111 benchmark markets comprises 16,457 articles from 2,217 domains (subset of a 122,290-article GDELT collection; Appendix F): 89% Western media, 3.4% think tanks (ISW), 2.1% Ukrainian English-language outlets, 2.8% Russian sources, and zero Ukrainian-language analytical sources. The D_{UA} corpus draws from DeepState frontline maps, General Staff daily loss reports, Ukrainian military bloggers, Militarnyi, Defense Express, and Ukrainian OSINT channels (Table 4).

3.2 Models

We evaluate three clean models – Gemini 2.5 Flash, Gemini 2.5 Pro, and GPT-5-mini – plus one contaminated control: Gemini 3.1 Pro Preview, whose training data extends past market resolution dates. The contaminated model's blind predictions show near-zero bias (+0.35 pp vs +2.0 pp clean average) and beat the no-change baseline by 10.4%, confirming knowledge of actual outcomes.

Flash and Pro share training data (Gemini 2.5 family) but differ in reasoning depth, enabling controlled comparison of processing effects. All models output structured predictions with full reasoning traces.

3.3 Statistical Framework

The market is the unit of independence throughout (N=65 territorial, N=46 diplomatic). Adjacent cutoffs have overlapping prediction windows (~44% overlap), so all tests use market-level aggregates with cluster-robust inference. We apply Bonferroni correction across 8 primary tests. With 65 clusters, we detect Cohen's $d \geq 0.35$ at 80% power. Our accuracy baseline is *no-change*: predict that the price at the horizon equals the current price. We use two distinct references: realised binary outcomes as the ground truth anchor in §4.1 (low-power but genuine), and the continuous market price trajectory as the calibration reference for all model comparisons in §4.2–4.4. Bias and MAE are therefore deviations from the market reference, not claims about reality. Full statistical details in Appendix B.

4 Results

4.1 The Benchmark Is Biased – But Useful

Polymarket overestimates Russian territorial capture by +3.5 pp relative to actual outcomes ($t(381) = 5.58$, $p < 10^{-7}$, $d = 0.29$; this test is point-level since it characterises a property of the benchmark itself, distinct from the market-clustered tests used for model comparisons in §4.2–4.4). Despite this bias, the market correlates strongly with outcomes ($r = 0.83$, $R^2 = 0.69$), and binary resolution provides insufficient power (3/65 territorial markets resolved YES). Note that Polymarket resolution for territorial markets relies on ISW and DeepState frontline updates, themselves expert assessments rather than direct observation; residual disagreements between these sources and on-the-ground reality are absorbed into the +3.5 pp figure. We use the continuous price trajectory as the calibration reference for all downstream comparisons.

4.2 English Context Destroys Signal

Blind models (condition A) show +2.0 pp average pro-capture bias – already present in training data, but *less* than Polymarket's +3.5 pp. Adding English news (condition D) pushes models to +3.4 pp, nearly matching the market's overestimation. English news does not add information – it replaces the model's more accurate prior with the discourse's less accurate framing. The shift is significant for Pro 2.5 (+2.4 pp, $p = 0.0001$) and GPT-5-mini (+1.4 pp, $p = 0.002$), both surviving Bonferroni correction. Flash shows a consistent but smaller effect (+0.5 pp, $p = 0.066$). All bias measurements below are relative to the market price trajectory; §4.1 establishes that this proxy is biased (+3.5 pp) but strongly correlated with outcomes ($r = 0.83$), meaning our estimates are conservative – the true distortion relative to reality is likely larger.

When context pushes predictions toward capture, those pushes are wrong 64–72% of the time across all clean models (binomial test vs 50%, $p < 10^{-6}$;

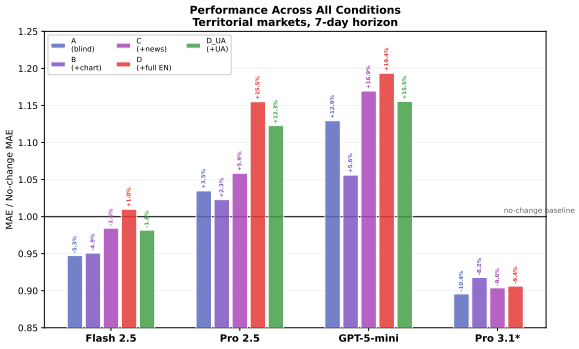


Figure 2: MAE vs no-change baseline across all information conditions. Adding English context (A→D) degrades predictions for all clean models; chart-only (B) outperforms full context (D). D_{UA} partially recovers accuracy. The contaminated model (Pro 3.1*) beats baseline under all conditions; D_{UA} was not run on it. Territorial markets, 7-day horizon.

Table 1). The bias is systematic and traceable: linguistic analysis of reasoning traces (Appendix L) shows Russia receives 2.3–3.7× more offensive verbs, "advances into" territory 77 times while Ukraine never does, and models reason about "if Russia succeeds" 60 times but "if Ukraine succeeds" zero times. Bias^2 accounts for only 2–9% of total error (Appendix G); we study it because it is directional, not because it dominates accuracy.

The full condition ladder (Figure 2; Appendix I) reveals that chart-only predictions (B) outperform full English context (D) for all clean models. Diplomatic markets serve as placebo: context does not damage predictions ($p = 0.03$ improvement for Flash), confirming a domain-specific mechanism (Appendix K). The bias is directionally asymmetric: downward predictions carry genuine signal, while upward predictions carry the discourse's offense-dominant distortion. Filtering out upward predictions converts all three models from losing to beating the no-change baseline (Appendix D).

4.3 Contaminated Model Ablation

The contaminated model (Pro 3.1*) knows actual outcomes – it beats no-change by 10.4%. Yet when English context pushes it toward capture, it is wrong at the same rate as clean models (Table 1).

The pro-capture push error rate is a property of the English news corpus, not of model ignorance (Appendix J). An inverted-question robustness check confirms this is not an artifact of question framing (Appendix C).

Model	Knows?	Push↑ acc.
Pro 3.1* (contam.)	Yes	27.9%
Flash 2.5	No	28.1%
GPT-5-mini	No	29.1%
Pro 2.5	No	36.3%

Table 1: When English context pushes predictions toward capture, accuracy is 28–36% across four models with different knowledge levels ($p < 10^{-6}$ for all, binomial test vs 50%). The error rate is a property of the corpus, not model ignorance.

4.4 Ukrainian Sources Reduce Bias; Accuracy Effects Are Model-Dependent

Supplementing English news with Ukrainian military sources (D → D_{UA}) reduces pro-capture bias across all three clean models (Figure 3). We use "correction" here in the sense of reducing deviation from the market reference; D_{UA} is source diversification, not fact-checking against a ground truth.

For Flash, D_{UA} eliminates the context-induced directional shift: bias drops from +0.4 pp ($p = 0.066$) to +0.1 pp ($p = 0.378$), indistinguishable from blind prediction. For Pro and GPT, significant pro-capture bias persists ($p < 0.01$). Ukrainian sources provide a comparable absolute correction across models, but the outcome differs because models accumulate different amounts of context damage: Flash takes little damage from English text, so the correction is sufficient; Pro amplifies the offense-dominant signal far beyond what source supplementation can repair (Appendices K, E, M).

Accuracy effects are more mixed. D_{UA} improves MAE on average across clean models, but not uniformly: the blind condition A remains competitive for some configurations, and D_{UA} hurts diplomatic predictions (Appendix K). The robust finding is bias reduction; absolute error improvements are partial and model-dependent.

5 Discussion and Conclusion

Our results indicate a measurable cost of information ecosystem misalignment with the market reference: English-language context induces a systematic pro-capture bias through both biased framing and source exclusion, and Ukrainian military sources partially counterbalance it. The bias is robust ($p < 10^{-6}$ for push accuracy), domain-specific (territorial but not diplomatic), and invariant to model knowledge (contaminated model shows same push error rate).

The practical implications are immediate. Sup-

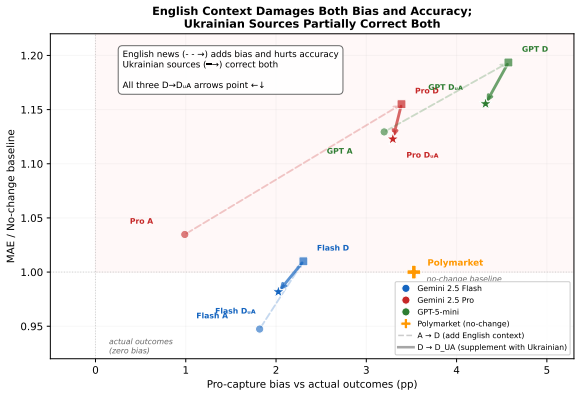


Figure 3: Bias vs accuracy for each model under conditions A (blind), D (English context), and D_{UA} (D supplemented with Ukrainian sources). Dashed arrows: A→D (English context damages). Solid arrows: D→D_{UA} (Ukrainian sources correct). All solid arrows point left and down. Territorial markets, 7-day horizon.

plementing English retrieval with Ukrainian sources through multilingual RAG reduces directional bias across all clean models; absolute-error improvements follow on average but are partial and model-dependent. Conservative reasoning (Flash) benefits most, while deeper reasoning (Pro) amplifies the offense-dominant signal beyond what source supplementation can repair. This makes model selection a critical lever: choosing conservative over deep reasoning can matter more than improving the information itself. The parametric bias already present in condition A reflects an English-centric information ecosystem and requires broadening source inclusion.

A case study illustrates the thesis in miniature: all clean models confidently predicted Zelenskyy would wear a suit to a papal funeral – logically sound but culturally blind (Appendix N). The method we introduce turns ecosystem misalignment from a qualitative concern into a grounded quantity.

Limitations

With 65 territorial market clusters, we achieve 80% power to detect Cohen's $d \geq 0.35$. Our MAE effects ($d = 0.25$–0.31) fall below this threshold, explaining non-significance after Bonferroni correction. The D_{UA} vs D improvement does not reach significance at the market-cluster level ($p = 0.10$–0.37). The Flash/Pro processing divergence, while controlled (same training data), is a single comparison. Polymarket's Ukraine markets may not generalize to other conflicts. While the case study uses Ukrainian sources, the methodology generalises to any conflict where one information ecosystem is suspected of systematic framing distortion relative to another.

Ethical Considerations

Our dataset spans 111 prediction markets about Ukraine – from whether specific cities will be captured and communities displaced, to diplomatic negotiations, international sanctions, and aid decisions. The territorial markets reduce the fate of real communities to price movements on a trading platform. We find this commodification of human catastrophe deeply troubling.

We use this data not to legitimize prediction markets, but because their structure inadvertently exposes something important: a measurable gap between what the English-language information ecosystem implies about events in Ukraine and what is actually happening. This gap – driven by both offense-dominant framing within included sources and exclusion of Ukrainian analytical ones – has consequences beyond prediction accuracy. Distorted understanding shapes international policy, humanitarian response, and the political will to support Ukraine.

The induced bias is not merely dovish or negotiation-oriented. It systematically pushes toward a low-agency, offense-dominant view of Ukraine in which Ukrainian leverage is discounted and territorial loss is treated as more inevitable than reality later shows. The worldview induced by the Anglo-American source ecosystem qualitatively resembles later concessionist policy rhetoric that downplays Ukrainian leverage. We leave this striking alignment for future study.

The bias we document is not abstract. When English-language AI systems systematically overestimate Russian territorial success, they reinforce a narrative of inevitable Ukrainian loss that Ukrainian soldiers, analysts, and journalists work daily to counter – through sources that English-language pipelines do not include.

This work used AI-based writing assistance tools for editing and formatting.

Acknowledgments

We gratefully acknowledge Amazon Web Services and DigitalOcean for the cloud compute credits and infrastructure that supported this work. Model inference, training, and evaluation pipelines were

run on AWS (EC2 GPU instances, SageMaker, S3); data collection and experiment tracking were hosted on DigitalOcean managed database and compute instances.

References

Mohammad Ali and Naeemul Hassan. 2022. A survey of computational framing analysis approaches. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9335–9348.

Rohan Alur, Bradly C Stadie, Daniel Kang, Ryan Chen, Matt McManus, Michael Rickert, Tyler Lee, Michael Federici, Richard Zhu, Dennis Fogerty, Hayley Williamson, Nina Lozinski, Aaron Linsky, and Jasjeet S Sekhon. 2025. AIA forecaster: Technical report. *arXiv preprint arXiv:2511.07678*.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4982–4991.

Dallas Card, Amber E Boydstun, Justin H Gross, Philip Resnik, and Noah A Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 438–444.

Rohitash Chandra, Haoyan Chen, Yaqing Zhang, Jiacheng Chen, and Yuting Wu. 2026. An evaluation of LLMs for political bias in Western media: Israel-Hamas and Ukraine-Russia wars. *arXiv preprint arXiv:2601.06132*.

Robert M Entman. 2004. *Projections of Power: Framing News, Public Opinion, and U.S. Foreign Policy*. University of Chicago Press.

Shangbin Feng, Chan Young Park, Yohan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3).

Danny Halawi, Fred Shi, Sebastian Borgeaud, Adam Lerer, Pieter Abbeel, and Jacob Steinhardt. 2024. Approaching human-level forecasting with language models. *arXiv preprint arXiv:2402.18563*.

Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E Tetlock. 2025. ForecastBench: A dynamic benchmark of AI forecasting capabilities. In *International Conference on Learning Representations*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*.

Preslav Nakov, Jisun An, Haewoon Kwak, Muhammad Arslan Mansurov, and Momin Mansurov. 2024. A survey on predicting the factuality and the bias of news media. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics.

Markus Ojala, Mervi Pantti, and Jenni Kangas. 2024. Framing the war in Ukraine: A comparative study of news coverage. *Journalism Studies*.

Yulia Otmakhova, Shima Khanehzar, and Lea Frermann. 2024. Media framing: A typology and survey of computational approaches across disciplines. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Grzegorz Ptaszek, Bogdan Yuskiv, and Serhii Khomych. 2024. War on frames: Text mining of conflict in Russian and Ukrainian news agency coverage on Telegram. *Media, War & Conflict*, 17(1):41–61.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174.

Justin Wolfers and Eric Zitzewitz. 2004. Prediction markets. *Journal of Economic Perspectives*, 18(2):107–126.

Justin Wolfers and Eric Zitzewitz. 2006. Interpreting prediction market prices as probabilities. NBER Working Paper 12200, National Bureau of Economic Research.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit Bayesian inference. In *International Conference on Learning Representations*.

A Theoretical Framework

We formalize the mechanism studied in this paper. The goal is not a general theory of conflict forecasting, but to make precise what we mean by *offense-dominant* beliefs under source exclusion.

Latent battlefield state. Let i index a market-event pair. Let $y_i \in \{0, 1\}$ denote the realized outcome, where $y_i = 1$ corresponds to offensive success on the target event. Each event has an unobserved latent state $z_i \in \mathbb{R}$ capturing the true degree of offensive feasibility.

Information ecosystems as biased signals. We consider two information ecosystems: E (English-language) and U (Ukrainian military-analytical). Each provides a noisy signal about the same latent state:

$$x_i^E = z_i + \beta_E + \varepsilon_i^E, \qquad x_i^U = z_i + \beta_U + \varepsilon_i^U,$$

where $\varepsilon_i^E, \varepsilon_i^U$ are zero-mean noise and $\beta_E, \beta_U \in \mathbb{R}$ are systematic shifts. An ecosystem is *more offense-dominant* when it shifts beliefs further toward offensive success. The English ecosystem is more offense-dominant than the Ukrainian one whenever $\beta_E > \beta_U$.

LLMs as belief elicitation instruments. For model m, let μ_m denote the model's effective baseline belief under condition A (minimal context: market overview and current price data), absorbing both pretraining priors and the model's processing of the price signal. Given a source set S, the model produces

$$\hat{p}_i^{(m)}(S) = \sigma\big(g_i^{(m)}(S)\big),$$

where $\sigma(\cdot)$ is the logistic sigmoid and $g_i^{(m)}(S)$ is a latent score formed by combining the model prior with available signals:

$$g_i^{(m)}(S) = \frac{\lambda_{m,0}\mu_m + \sum_{e\in S} \lambda_{m,e}\, x_i^e}{\lambda_{m,0} + \sum_{e\in S} \lambda_{m,e}}.$$

Here $\lambda_{m,0} \geq 0$ weights the prior and $\lambda_{m,e} \geq 0$ weights ecosystem e. We treat these as effective parameters that absorb presentation-order effects and model-specific context processing. This maps onto our three experimental conditions:

$$\begin{aligned}\hat{p}_i^{(m)}(A) &= \sigma(\mu_m),\\ \hat{p}_i^{(m)}(D) &= \hat{p}_i^{(m)}(\{E\}),\\ \hat{p}_i^{(m)}(D_{UA}) &= \hat{p}_i^{(m)}(\{E, U\}).\end{aligned}$$

Proposition 1 (Source exclusion shifts latent scores toward offense)**.** *Assume ecosystem signals satisfy $x_i^e = z_i + \beta_e + \varepsilon_i^e$ with $\mathbb{E}[\varepsilon_i^e] = 0$, and that the latent score $g_i^{(m)}(S)$ is a positively weighted average of the model prior and available signals. If $\beta_E > \beta_U$ and $\lambda_{m,U} > 0$, then for every event i:*

$$\mathbb{E}\Big[g_i^{(m)}(D)\Big] > \mathbb{E}\Big[g_i^{(m)}(D_{UA})\Big].$$

Excluding Ukrainian sources systematically increases the expected latent score toward offensive success.

Proof. Substituting $\mathbb{E}[x_i^E] = z_i + \beta_E$ and $\mathbb{E}[x_i^U] = z_i + \beta_U$:

$$\mathbb{E}\Big[g_i^{(m)}(D)\Big] = \frac{\lambda_{m,0}\mu_m + \lambda_{m,E}(z_i + \beta_E)}{\lambda_{m,0} + \lambda_{m,E}},$$

$$\begin{aligned}&\mathbb{E}\Big[g_i^{(m)}(D_{UA})\Big]\\ &= \frac{\lambda_{m,0}\mu_m + \lambda_{m,E}(z_i + \beta_E) + \lambda_{m,U}(z_i + \beta_U)}{\lambda_{m,0} + \lambda_{m,E} + \lambda_{m,U}}.\end{aligned}$$

Since $\lambda_{m,U} > 0$ and $\beta_U < \beta_E$, the additional term pulls the weighted average below the English-only value. □

From scores to probabilities. Since $\sigma(\cdot)$ is monotone increasing, the score ordering implies $\hat{p}_i^{(m)}(D) > \hat{p}_i^{(m)}(D_{UA})$ pointwise for each realization of the noise. The ordering $\mathbb{E}[\hat{p}_i^{(m)}(D)] > \mathbb{E}[\hat{p}_i^{(m)}(D_{UA})]$ then follows by taking expectations. However, the *magnitude* of the probability gap depends on the curvature of σ and the noise distribution, so the proposition is stated at the level of latent scores where the result is exact.

Operational quantities. The framework yields three empirically measurable quantities. The *harm of English context*: $H_m = \mathbb{E}_i[\hat{p}_i^{(m)}(D) - \hat{p}_i^{(m)}(A)]$. The *source exclusion cost*: $X_m = \mathbb{E}_i[\hat{p}_i^{(m)}(D) - \hat{p}_i^{(m)}(D_{UA})]$. The *directional bias*: $B_m(S) = \mathbb{E}_i[\hat{p}_i^{(m)}(S) - y_i]$. On territorial markets, positive $B_m(S)$ indicates systematic overprediction of offensive success. Our main empirical finding is $B_m(D) > B_m(D_{UA})$, with $H_m > 0$ and $X_m > 0$.

Limitations of the formalization. This framework is deliberately modest. It does not claim that LLM outputs recover latent beliefs perfectly, nor that ecosystems differ only along a single dimension. The weighted-average assumption is an idealization: actual LLMs process context sequentially, and effective weights may depend on presentation order, prompt structure, and reasoning depth. The formalization makes explicit the mechanism tested

– if one ecosystem is more offense-dominant, excluding the other should produce more offense-dominant predictions – without claiming more than the experimental design supports.

B Statistical Methodology

Clustering. Adjacent cutoffs within a market have overlapping 7-day prediction windows (~44% overlap at the median 16-day gap). Intra-cluster correlation ranges from 0.008 (Flash) to 0.141 (GPT-5-mini). All tests use market-level aggregates.

Multiple testing. We apply Bonferroni correction across 8 primary tests. Market bias (territorial and diplomatic), directional bias shift (3 models), and MAE damage (3 models). After correction: market bias ($p < 10^{-4}$), Pro 2.5 bias shift ($p = 6.8 \times 10^{-4}$), and GPT bias shift ($p = 0.018$) survive. MAE tests and Flash bias shift do not survive. Push accuracy tests ($p < 10^{-6}$) are excluded from this family.

Power. With N=65 clusters, 80% power at $\alpha = 0.05$ requires $d \geq 0.35$. Observed: Pro 2.5 $d = 0.50$ (powered), GPT $d = 0.36$ (marginal), Flash $d = 0.19$ (underpowered). MAE effects $d = 0.25$–0.31 (underpowered). Flash-scale detection requires ~215 clusters.

Permutation tests. Sign-randomization tests (10,000 iterations) confirm all parametric results with concordant p-values.

C Question Inversion (Anti-X)

We re-run territorial predictions with inverted questions: "Will Russia *fail to capture* X?" with inverted prices ($1-p$). This is analogous to semantic entropy (Kuhn et al., 2023) but diagnoses *context quality* rather than model confidence.

When contradictions occur under condition D, they are directionally asymmetric. For Pro 3.1* (contaminated), the asymmetry is 47:1 – original-down/anti-up vastly dominates. English military reporting is stronger evidence *against failure* than *for capture*.

D Selective Trust

A zero-parameter rule – trust the model only when it predicts downward movement – converts losing strategies to winning ones. Flash selective: −3.4% vs no-change ($p < 10^{-21}$). Pro 2.5 selective: −3.3%. This reveals that downward predictions carry genuine signal while upward (pro-capture) predictions carry systematic noise. D_{UA} + selective is the best combination for GPT-5-mini (−0.8%), finally converting it to a winning strategy.

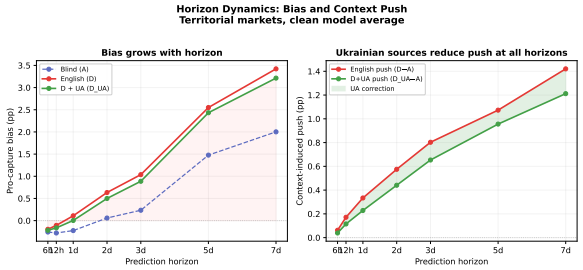


Figure 4: Left: pro-capture bias grows with prediction horizon for all conditions. Right: context-induced push (D−A) and Ukrainian-corrected push (D_{UA}−A) across horizons. D_{UA} reduces push at every horizon. Territorial markets, clean model average.

Hz	A	D	D_{UA}	D push	D_{UA} push
6h	−0.3	−0.2	−0.2	+0.1	+0.0
1d	−0.2	+0.1	+0.0	+0.3	+0.2
3d	+0.2	+1.0	+0.9	+0.8	+0.7
7d	+2.0	+3.4	+3.2	+1.4	+1.2

Table 2: Bias by horizon (pp, clean model average, territorial). Bias grows with horizon; D_{UA} reduces push at all horizons.

E Per-Horizon Analysis

Pro-capture bias grows with prediction horizon (Table 2, Figure 4). D_{UA} consistently produces less push than D at every horizon.

F Source Ecosystem and Utility

F.1 English Information Diet

The full GDELT collection spans all topics and comprises 122,290 articles from 6,232 domains. Of these, 16,457 articles from 2,217 domains are linked to our 111 Ukraine benchmark markets (§3). The composition statistics below describe the benchmark subset. The corpus is US-centric (63% of articles) and Western-oriented (Table 3).

The top domains by volume are Yahoo (5,864), marketscreener (1,142), freerepublic (781), Daily Mail (736), globalsecurity.org (659), ZeroHedge (595), ABC News (560), ISW (534), Independent (464), and Newsweek (462). The corpus reflects what GDELT surfaces as the English-language information diet about Ukraine – mainstream wire content, finance aggregators, and partisan outlets.

Category	Examples	%
Wire / mainstream	AP, Reuters, CNN, Guardian	36
Aggregators	Yahoo, AOL, bignewsnetwork	12
International	jpost, economictimes (India)	12
Conservative US	Fox, Breitbart, Epoch Times	5
Finance	marketscreener, fxstreet, Forbes	5
UK press	Daily Mail, Independent	5
Specialist / OSINT	ISW, globalsecurity.org	1
Other / regional	various	24

Table 3: English source composition by category. US sources account for 63% of articles. Ukrainian-language analytical sources: zero.

Zero Ukrainian-language analytical sources appear.

F.2 Ukrainian Information Diet

The D_{UA} condition supplements condition D with three source types (Table 4).

Source	Count	Type
Telegram bloggers	47,009 posts	16 channels
Militarnyi.com	2,352 articles	Defense news
General Staff losses	1,468 records	Daily casualty data

Table 4: Ukrainian sources added in D_{UA}.

The 16 Telegram channels represent the core of the Ukrainian military information ecosystem: 5 active/reserve military (including brigade commanders Magyar, Zhorin, Shtefan, Fedorenko, Krotevych), 4 journalists (Tsaplienko, Butusov, Kazanskyi), 4 analysts (Berezovets, Kovalenko, Mashovets, Zhdanov), 1 official channel (General Staff ZSU), 1 tech specialist (Beskrestnov), and 1 commentator. All channels represent Ukrainian perspectives – no Russian, neutral, or Western-OSINT channels are included. This is by design: D_{UA} tests whether adding the Ukrainian military-analytical ecosystem improves prediction accuracy, not whether balanced sourcing does.

F.3 Source Utility Analysis

We analyze which sources models actually cite in their reasoning traces (condition D, 7d horizon, clean models, $N = 2{,}130$ predictions) and whether citing a source correlates with better or worse predictions. The patterns below are correlational: ISW-cited predictions performing worse does not establish that citing ISW *causes* worse predictions; both may reflect harder markets or more contested events. Relative performance compares the model's error when citing a source against the no-change baseline for those same markets.

Source	Freq	Rel. Perf	Dir% (cited)	Dir% (not)
Russian officials	43.0%	−1.4%	**61.9**	51.9
Ukrainian officials	36.9%	−5.4%	60.0	54.0
Price action	76.9%	−8.6%	56.7	54.5
ISW	51.5%	−11.3%	50.9	**61.8**
Russian milbloggers	6.7%	−11.9%	54.8	56.3
Polymarket traders	17.7%	−6.5%	55.6	56.3
DeepState	3.8%	−1.0%	56.1	56.2
UA General Staff	5.4%	−15.1%	41.5	57.1
OSINT	1.9%	−14.6%	41.5	56.5

Table 5: Source utility: frequency of citation in reasoning traces and correlation with prediction quality. Relative performance = (NC_error − model_error) / NC_error; negative = worse than no-change. ISW – the most-cited analytical source (51.5%) – correlates with *worse* directional accuracy (50.9%) than predictions not citing it (61.8%).

The ISW paradox. ISW is the dominant analytical source (51.5% citation rate). Yet predictions citing ISW show 50.9% directional accuracy – a coin flip – compared to 61.8% when ISW is not cited. This is consistent with ISW's offense-dominant analytical framing: detailed, authoritative coverage of offensive operations that models internalize as evidence for territorial change. The authority of the source amplifies the framing effect.

Russian officials as accidental signal. Russian officials are cited at 43.0% with the best directional accuracy (61.9%) among high-frequency sources. This is counterintuitive until one considers the epistemic framing from Appendix L: models treat Russian claims with skepticism ("Russia *claims*..."), and this discounting accidentally produces better-calibrated predictions than absorbing ISW's authoritative framing uncritically.

Contaminated model source shift. The contaminated model (Pro 3.1*), which knows outcomes, shifts its citation patterns: ISW drops from 51.5% to 35.4%, Ukrainian officials from 36.9% to 23.1%, while price action rises from 76.9% to 81.0%. A model with outcome knowledge relies less on narrative sources and more on the price signal – further evidence that narrative sources add framing, not information.

G Bias-Variance Decomposition

Bias^2 accounts for 2–9% of model MSE; variance dominates at 91–98%. Polymarket: 8%

Model	Param.	Context	UA corr.
Flash 2.5	+1.8 pp	+0.5 pp	−0.3 pp
Pro 2.5	+1.0 pp	+2.4 pp	−0.1 pp
GPT-5-mini	+3.2 pp	+1.4 pp	−0.3 pp

Table 6: Three-layer bias decomposition. Parametric bias (A) is in the weights. Context-induced bias (D−A) comes from English news. Ukrainian correction ($D-D_{UA}$) is the bias reduction from supplementing with Ukrainian sources – roughly constant across models, but against vastly different context damage.

Model	Bias recov.	Accuracy recov.
Flash 2.5	57%	45%
Pro 2.5	4%	27%
GPT-5-mini	18%	60%

Table 7: D_{UA} recovery of context-induced damage. All models improve on both dimensions.

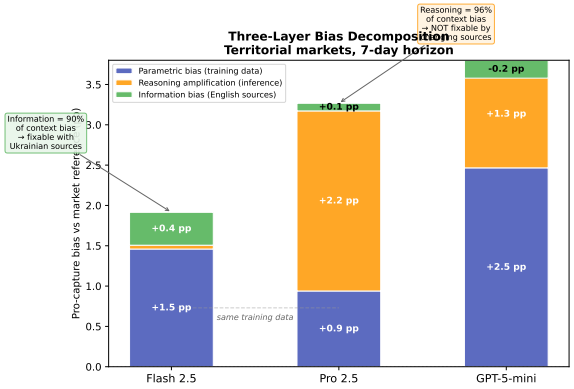


Figure 5: Bias decomposition. Flash and Pro share training data but differ 5× in context-induced bias. Ukrainian sources provide a similar absolute correction for all models, but it is swamped by Pro's context damage.

bias[2]. English context increases both components; Ukrainian sources reduce both. Variance reduction is the larger contributor to MAE improvement for bold models.

H Three-Layer Bias Decomposition

We present this decomposition as exploratory and correlational. With 65 territorial markets, per-model component estimates are noisy and we do not draw causal source-level conclusions from them. Our experimental conditions separate three components of pro-capture bias (Figure 5), measured on territorial markets at 7-day horizon.

Parametric bias (training data). Measured by condition A – no context, just the model's priors. Flash: +1.8 pp. Pro: +1.0 pp. GPT: +3.2 pp.

Context-induced bias (English news). Measured by D−A: the additional pro-capture shift from English context. This is where models diverge dramatically. Flash: +0.5 pp. Pro: +2.4 pp. GPT: +1.4 pp. Pro accumulates 5× more context damage than Flash despite sharing training data – deeper reasoning amplifies the offense-dominant signal in English text.

Ukrainian source correction. Supplementing with Ukrainian sources provides a roughly constant absolute correction across models: Flash −0.3 pp, Pro −0.1 pp, GPT −0.3 pp (Table 6). What differs is not the correction but the denominator – how much context damage each model accumulates. For Flash, 0.3 pp corrects most of the 0.5 pp context damage (57%). For Pro, a similar correction is negligible against 2.4 pp of damage (4%).

This is a model selection result: Ukrainian sources help all models roughly equally in absolute terms, but their impact depends on how aggressively the model amplifies English context. Conservative reasoning (Flash) keeps context damage small, making the correction sufficient. Deep reasoning (Pro) amplifies English bias beyond what source supplementation can repair.

I Full Condition Comparison

Table 8 reports MAE relative to the no-change baseline (in %) for every model under each of the five information conditions on territorial markets at the 7-day horizon. Negative values mean the model beats no-change; positive values mean it loses to it. Two patterns stand out. First, chart-only predictions (B) outperform full English context (D) for all three clean models, indicating that the bulk of the damage from condition D comes from narrative content rather than from price or chart inputs. Second, only the contaminated model (Pro 3.1*) beats no-change under any condition. Figure 6 shows the same comparison at the per-market level for D vs D_{UA}: D_{UA} wins the majority of markets for Flash (37/65) and Pro (32/65), confirming that the bias reduction reported in §4.4 is not driven by a few outlier markets.

J Contaminated Model Details

Pro 3.1* shows near-zero blind bias (+0.35 pp vs +2.0 pp clean average) and beats no-change by

	A	B	C	D	D_{UA}
Flash	−5.3	−4.9	−1.6	+1.0	−1.8
Pro	+3.5	+2.3	+5.9	+15.5	+12.3
GPT	+12.9	+5.6	+16.9	+19.4	+15.5
3.1*	−10.4	−8.2	−9.6	−9.4	–

Table 8: MAE vs no-change (%), territorial 7d. Chart-only (B) outperforms full context (D) for all clean models.

Figure 6: Per-market MAE under D (English) vs D_{UA} (supplemented with Ukrainian sources). Each dot is one territorial market. Points below the diagonal indicate D_{UA} outperforms D. D_{UA} wins the majority of markets for Flash (37/65) and Pro (32/65). Territorial markets, 7-day horizon.

10.4% blind, 9.4% with context – the only model to beat baseline under any condition. Despite this knowledge, its pro-capture push accuracy (27.9%) matches clean models, demonstrating that the directional signal in English military reporting overrides even direct outcome knowledge.

K Diplomatic Markets (Placebo)

D_{UA} hurts diplomatic predictions for Flash (+5.9% vs D) and Pro (+5.8%), while GPT is unchanged (−0.1%). Ukrainian military sources contain no diplomatic signal; English coverage of negotiations and sanctions is more informative. This confirms domain specificity (Figure 7).

L Linguistic Analysis of Offense-Dominant Framing

The quantitative results in §4 stand independently of the analysis below. We provide linguistic analysis of reasoning traces as qualitative illustration of the mechanism behind the statistical findings.

We analyze 444 reasoning traces (111 markets × 4 models, condition D, 7d horizon, first cutoff per market; ∼159,000 words) using regex-based proximity matching with manual validation. The analysis reveals eight dimensions of offense-dominant framing, all pointing in the same direction across all models.

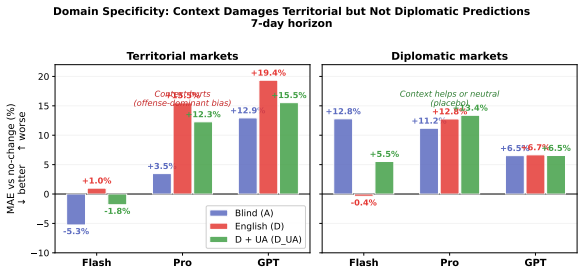


Figure 7: Domain specificity: English context damages territorial predictions (left) but not diplomatic ones (right). D_{UA} corrects territorial bias but hurts diplomatic predictions, confirming the intervention is domain-specific. 7-day horizon.

L.1 Agency: Who Acts

Russia is systematically framed as the agent. Across all models, Russia appears as grammatical subject 60–66% of the time (ratio 1.4–2.0× over Ukraine), is the first actor mentioned in 76–86% of texts, and receives 1.56–1.76× more total mentions. Ukraine's primary role is reactive.

L.2 Verb Semantics: What Each Side Does

Within 80 characters of each actor mention, 59–64% of verbs near Russia are offensive (advance, capture, assault, deploy) while Ukraine's verbs split between defensive (hold, resist, repel) and diplomatic. Russia receives 2.3–3.7× more offensive verbs than Ukraine across all four models.

L.3 Success/Failure Framing

The most extreme asymmetry. Russia's success-to-failure language ratio ranges from 6.7:1 to 14.1:1. Ukraine's ratio is 1.4–2.5:1. Russia is described in almost exclusively positive-outcome terms (advance, progress, gains, momentum) even when the model predicts those outcomes will not materialize. The contaminated model (Pro 3.1*) shows the highest ratio (14.1:1) despite knowing most territorial outcomes resolve against capture – framing and prediction are decoupled.

L.4 Territorial Lexicon

L.5 Conditional Erasure

Models reason about hypothetical Russian success but never about Ukrainian success:

Pro 3.1* produces 37 "if Russia succeeds" constructions – the most of any model – despite knowing most such outcomes do not occur.

Verb	Russia	Ukraine	Ratio
capture	143	16	8.9×
advance into	77	**0**	∞
seize	14	1	14×
encircle	15	2	7.5×
hold	2	3	0.7×

Table 9: Territorial verbs attributed to each actor (sum across 4 models, 444 traces). "Advance into" appears 77 times near Russia and **zero** times near Ukraine.

Pattern	GPT	Fl.	Pro	3.1*	Tot.
"If R succeeds…"	6	13	4	37	60
"If R fails…"	0	0	2	0	2
"If U succeeds…"	0	0	0	0	**0**
"If U fails…"	1	3	0	1	5

Table 10: Conditional framing in reasoning traces. Ukrainian success is never considered as a scenario.

L.6 Epistemic Framing

Russia "claims" (224 instances) while Ukraine "denies" (20 instances; Russia: 1). Ukraine's primary epistemic role is refuting Russian assertions rather than making its own. Russia is also "confirmed" 2× more than Ukraine, creating a paradox where Russian assertions are simultaneously more doubted and more validated.

L.7 Syntactic Subordination

"Despite Ukrainian resistance, Russia continues to advance" appears 30 times. The inverse –"Despite Russian challenges, Ukraine holds" – appears 6 times. Ukrainian action is systematically placed in concessive clauses; Russian action occupies the main clause. Ukrainian defense is framed as *overcome*; Russian offense as *persisting*.

L.8 Composite Scorecard

Every dimension – agency, verb semantics, success framing, conditional reasoning, epistemic credibility, syntactic structure – points in the same direction across all four models. The contaminated model, which *knows* most territorial outcomes resolve against offense, produces the most extreme framing on several dimensions. This decoupling of framing from knowledge confirms that the offense-dominant pattern is structural – embedded in how English-trained LLMs construct conflict narratives – rather than a reflection of model beliefs about outcomes.

Metric	GPT	Flash	Pro	3.1*
Mention ratio (R/U)	1.56	1.67	1.76	1.68
Offensive verb ratio	2.28	3.20	3.70	3.34
Russia S/F ratio	8.1	9.0	6.8	**14.1**
Ukraine S/F ratio	1.4	2.2	2.5	1.6
Subject ratio (R/U)	1.39	1.96	1.87	1.47
Russia first (%)	78.6	79.6	84.0	**85.9**
"If R succeeds"	6	13	4	**37**
"If U succeeds"	0	0	0	0

Table 11: Composite framing scorecard across all four models. Every metric shows the same direction. The contaminated model (3.1*) amplifies framing despite knowing outcomes.

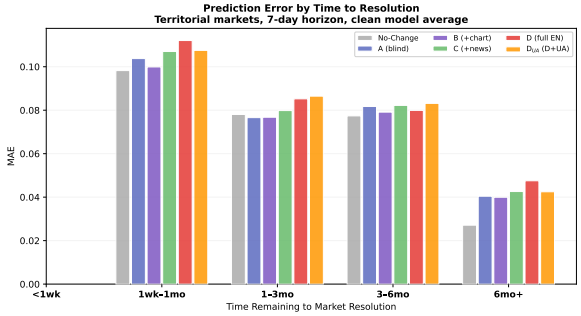


Figure 8: MAE by time remaining to market resolution. Context damage is strongest in the 1wk–1mo window where news flow is densest. Territorial markets, 7-day horizon, clean model average.

M Prediction Error by Time to Resolution

Figure 8 shows prediction error stratified by time remaining until market resolution. Context damage (D exceeding no-change) is concentrated in the 1-week-to-1-month window, where markets are most active and English news coverage is densest. At longer horizons (6+ months), all conditions converge as markets are less liquid and predictions are dominated by the prior.

N Case Study: The Zelenskyy Suit

Consider the Polymarket question "Will Zelenskyy wear a suit before June?" All three clean models predicted YES with high confidence (0.91–0.95), reasoning that a papal funeral demands formal attire – a logically sound inference from generic diplomatic norms. Any Ukrainian would have predicted differently. Zelenskyy has not worn a suit since February 24, 2022; the wartime military clothing is a deliberate political statement, not a wardrobe constraint. Returning to a suit would signal a fundamental shift in how Ukraine frames its wartime posture. The contaminated model, which

knows the outcome, predicted 0.33. The gap between 0.93 and 0.33 is not a reasoning failure – the logic is valid. It is an information ecosystem failure: the English-language corpus encodes "heads of state wear suits to funerals" but not "this particular head of state has made not wearing a suit a defining act of wartime leadership." The models reason fluently from the wrong world model.

Toward a Gold-Standard Benchmark for Evaluating Ukrainian Language Proficiency in LLMs

Svitlana Galeshchuk[2,3], Yuliia Maksymiuk[4], Yuliia Chernobrov[1], Oleksandra Antoniv[5], Nina Stankevych[5], Nataliia Faryna[5], Oksana Popkova[6]

[1]National Commission for State Language Standards, [2]BNP Paribas, [3]West Ukrainian National University, [4]Independent researcher, [5]Ivan Franko National University of Lviv, [6]Kherson State University

Correspondence: y.chernobrov@mova.gov.ua

Abstract

The paper presents an expert-curated benchmark for assessing Ukrainian proficiency in LLMs, focusing on grammar, lexical norms, and orthography as core components of language competence. Prepared by professional linguists, the proposed gold-standard dataset is designed to test normative Ukrainian usage.

The benchmark is further used to evaluate a range of LLMs, including Ukrainian-focused, multilingual, and large-scale models, under zero-shot and few-shot prompting in Ukrainian and English. Across these settings, smaller models achieve no more than 42.1% accuracy, while large-scale LLMs reach up to 59.6%. These results show that standard Ukrainian remains challenging for current LLMs and highlight the need for stronger language-specific evaluation and adaptation.

1 Introduction

Large language models (LLMs) are increasingly trained on multilingual datasets, but the majority of pre-training data usually consists of English texts. As a result, the ability of LLMs to process underrepresented languages such as Ukrainian remains difficult to assess reliably. This issue is especially important given the growing adoption of Artificial Intelligence (AI) systems employing language models by Ukrainian users. Can we claim that these or other closed- or open-source models are truly fluent in Ukrainian and capable of generating grammatically correct sentences?

In human language assessment, proficiency is typically evaluated through standardized examinations comprising multiple tests designed to measure different aspects of language competence. We adopt a similar perspective for evaluating LLMs and frame the task as a benchmark-based assessment.

Specifically, we present an expert-curated dataset of 347 multiple-choice questions designed to evaluate Ukrainian language proficiency in LLMs. The benchmark is developed by professional linguists and focuses on grammar (with particular emphasis on morphology and syntax), vocabulary, and orthography. Using this benchmark, we test a range of widely used closed- and open-source models, including Mistral Medium and Small, GPT-OSS-120B, Gemma 3 (4B and 12B), the Llama family of models, as well as models further pre-trained for Ukrainian, namely MamayLM and Lapa. Our evaluation reveals systematic weaknesses across multiple linguistic patterns and persistent challenges in Ukrainian language processing.

The paper is organized as follows. Section 2 describes the benchmark and its development, including its linguistic coverage and construction principles. Section 3 reviews related work on Ukrainian language benchmarks. Section 4 presents the experimental setup, including evaluation metrics, models, and prompting design. Section 5 reports the main results and discusses performance differences across models. Section 6 outlines the limitations of the current study and suggests directions for future work. Section 7 summarises the main findings and argues for the importance of the benchmark.

2 Benchmark Development

The grammar test tasks proposed by the authors aim to assess language competence (Latin *competens* – proper, appropriate), meaning knowledge of the norms of the modern standard language and the ability to use them skillfully in context. The benchmark targets, in particular, grammatical and orthographic competence at the level expected of native speakers. The benchmark was created by professional philologists and linguists with 10 to 30 years of experience teaching Ukrainian language courses at higher education institutions. In addition, each question was reviewed by at least one other expert for correctness and clarity. Test items found to be ambiguous or insufficiently comprehensive

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 121–135
May 29-30, 2026

were revised or removed.

The Ukrainian grammar test items were compiled from a broad range of normative and reference sources, including (Matsiuk and Stankevych, 2017), (Serbenska, 2019), (Voloshchak, 2007), (Maznichenko et al., 2019). These sources were selected because they provide extensive coverage of grammatical phenomena, reflect the norms of standard Ukrainian, and document their historical development.

The presented benchmark contains 347 multiple-choice questions with either four or five answer options. Each question targets a specific grammatical, lexical, or orthographic phenomenon and includes one correct answer consistent with the norms of standard Ukrainian together with plausible distractors but non-normative alternatives.

2.1 Benchmark Overview

This subsection describes the main properties of the proposed benchmark and the linguistic phenomena it covers. The grammar tests pay special attention to forms in which native speakers of Ukrainian often make mistakes due to negative interference (examples are provided where differences at the morphological and syntactic levels exist among various Slavic languages). LLMs are expected to clearly distinguish these languages and recognize the inherent features of Ukrainian. Each LLM is asked to analyze individual words as well as word combinations or sentences.

The tasks are formulated as questions that specify the essence of the possible error. A positive feature of the test tasks is the use of both appellative and onymic vocabulary, as well as the identification of the most common grammatical mistakes. Figure 1 shows the distribution of correct answer positions across the benchmark. The answers are relatively balanced overall, although the fifth appears less frequently because only a subset of items has five answer choices.

Figure 2 shows the distribution of question lengths, measured in number of words per question.

The questions cover three broad linguistic categories: grammar (morphology, syntax), lexical norms, and orthography. **Grammatical** competence refers to the knowledge and ability to use the grammatical resources of the Ukrainian language, including word-formation units, methods of word formation, morphological units, categories and forms, as well as syntactic units and categories. This competence is necessary for understanding

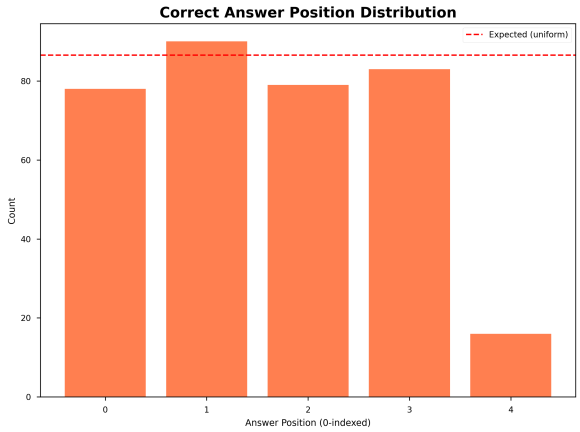


Figure 1: Distribution of correct answer positions across the five answer choices (A corresponds to 0, while the Ukrainian letter Д (equivalent to D) corresponds to 4). The red dashed line indicates the expected count under a uniform distribution. Most positions are close to the expected frequency, while position 4 appears somewhat underrepresented.

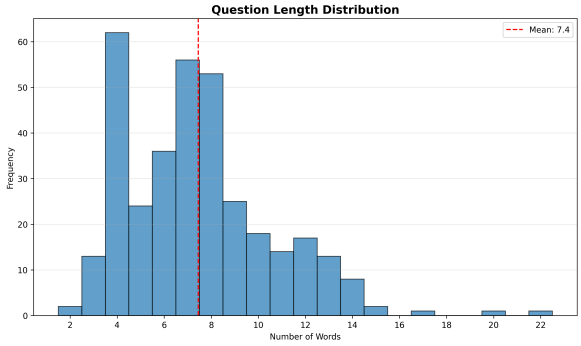


Figure 2: Distribution of question lengths in the benchmark, measured by number of words per question. The red dashed line marks the mean question length.

and producing texts in various fields of professional activity. It includes:

1. Morphology. This category includes nouns, adjectives, numerals, pronouns, and verbs. The benchmark covers such phenomena as genitive case endings *-a (-ia) / -u (-iu) (Ukrainian:* -а(-я)/-у(-ю)*) in masculine singular nouns*, instrumental and vocative case endings, genitive plural endings, singular and plural forms of nouns, gender forms of invariable foreign origin nouns and abbreviations, comparative and superlative forms of adjectives, the combination of numerals with nouns, case forms of numerals, the use of numerals to indicate time, normative use of pronouns, personal verb forms, future tense forms, imperative forms, and standard verb forms such as participles and gerunds.

2. Syntax. This category includes both word-level combinations and sentence-level syntax. At the word-combination level, the benchmark covers prepositional government, nonprepositional government, and complex cases of agreement. At the sentence level, it includes detached adverbial modifiers expressed by participial phrases, series of homogeneous sentence parts, agreement of the subject with the predicate, and norms for constructing complex sentences.

Vocabulary. This category covers lexical norms and normative word usage. It includes questions on the semantic compatibility of words in word combinations, distinctions between near-synonyms, and the selection of words appropriate to the meaning intended in context. These items target cases where incorrect usage often arises from semantic interference, calques, or confusion between similar lexical items.

Orthography. This category includes questions targeting the norms of standard Ukrainian spelling and related orthographic conventions.

Representative examples for all major categories are provided in Appendix A.

3 Related Work

Currently, there is no universally accepted standard for evaluating Ukrainian language proficiency in large language models. Existing Ukrainian-language benchmarks can be broadly categorized into three types: (i) resources translated from other languages and subsequently validated for annotation errors (Saini et al., 2024); (ii) synthetically generated benchmarks (Bondarenko et al., 2023); and (iii) corpora curated by human annotators and released as silver- or gold-standard resources.

Our benchmark belongs to the gold-standard category, as it has been carefully constructed and verified by domain experts following a rigorous annotation protocol. In this work, we therefore focus primarily on resources that are methodologically and qualitatively comparable to ours, namely UA-GEC and ZNO.

UA-GEC (Syvokon et al., 2023) is designed to support research in grammatical error correction and related tasks. It is an annotated corpus of texts containing grammatical errors and fluency issues, compiled from writings produced by both native and non-native speakers of Ukrainian. In total, the corpus comprises 1,872 documents that have been professionally corrected and annotated by expert linguists.

The ZNO benchmark (Romanyshyn et al., 2024) is a multiple-choice question resource derived from the Ukrainian External Independent Evaluation (Zovnishnie nezalezhne otsiniuvannia, ZNO), the standardized exam used for university admissions in Ukraine. It contains machine-readable questions and correct answer labels across two subject areas: Ukrainian language and literature, and History of Ukraine. The resource is provided in `.jsonl` format, where each entry consists of a question prompt, a set of answer options (A–D/E), the correct answer, and a subject label. It comprises around 3,800 question-answer pairs from exams administered between 2006 and 2023.

Compared with ZNO, our benchmark is narrower in scope but more focused on normative grammatical competence. Some items in our benchmark, especially morphology tasks, overlap with aspects covered by ZNO, while syntax-oriented questions extend the evaluation toward phenomena examined in greater detail in higher education Ukrainian language courses. Compared with UA-GEC, which is centered on error correction in running text, our benchmark provides a controlled multiple-choice format for targeted evaluation of grammatical and orthographic knowledge.

The methodological value of our benchmark lies in its role as a concentrated expert-curated evaluation resource of normative grammatical forms and in its focus on the most challenging areas of Ukrainian grammar.

4 Experimental Setup

4.1 Task Definition

Each benchmark item is formulated as a multiple-choice question with four or five candidate options and exactly one correct answer. The task for the model is to identify the option that conforms to the norms of standard Ukrainian grammar or orthography.

We treat this benchmark as a multiple-choice classification task. For each question q_i, the model is given a finite set of candidate answers $\mathcal{A}_i = \{a_{i1}, \ldots, a_{iK_i}\}$, where $K_i \in \{4, 5\}$, and must select the correct option a_i^*. Unlike standard classification tasks such as sentiment analysis, the answer labels do not carry fixed semantic meaning across items; the task is therefore to choose the correct option from the alternatives provided.

4.2 Evaluation Metrics

We evaluate model performance in two settings. In the first, the model scores the available answer options and the highest-scoring option is selected. In the second, the model generates an answer in text form. In both settings, we use accuracy as the main evaluation metric.

Let N be the total number of questions, let q_i be the i-th question, let $\mathcal{A}_i = \{a_{i1}, \ldots, a_{iK_i}\}$ be the set of answer options for that question, and let a_i^* be the correct answer.

4.2.1 Log-Likelihood Accuracy

In the log-likelihood setting, the model assigns a score to each answer option given the question. The predicted answer is the option with the highest score:

$$\hat{a}_i^{\mathrm{LL}} = \arg\max_{a \in \mathcal{A}_i} \log P_\theta(a \mid q_i).$$

Log-likelihood accuracy is then defined as

$$\mathrm{Acc}_{\mathrm{LL}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[\hat{a}_i^{\mathrm{LL}} = a_i^*\right],$$

where $\mathbf{1}[\cdot]$ equals 1 if the predicted answer is correct and 0 otherwise. This metric shows whether the model assigns the highest score to the correct answer.

4.2.2 Exact Match

In the generative setting, the model produces a text answer g_i. Exact Match counts a prediction as correct only if the normalized output exactly matches the correct answer:

$$\mathrm{EM} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[\mathrm{norm}(g_i) = a_i^*\right],$$

where $\mathrm{norm}(\cdot)$ denotes simple normalization of the generated answer.

4.2.3 Extractive Match

Models do not always return only the answer itself and may generate additional text. To account for this, we also report Extractive Match. This metric first extracts the predicted answer from the generated output and then compares it with the correct answer:

$$\mathrm{XM} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[\mathrm{ext}(g_i) = a_i^*\right],$$

where $\mathrm{ext}(\cdot)$ is a rule-based extraction function implemented with regular expressions.

4.3 Hardware

The experiments are conducted on a GPU server using a single 96 GB GPU (`gpu_1x_96gb`). vLLM v0.10.1.1 serves as the inference backend and LightEval v0.13.1.dev0 as the evaluation framework, running on Python 3.13.12.

4.4 Models

We evaluate models from four groups; the full list is provided in Appendix D.

Ukrainian-focused models. **MamayLM** (4B and 12B) (Yukhymenko et al., 2025) are Gemma 3 models further pre-trained and fine-tuned on Ukrainian datasets. **Lapa-12B** (lap) is also a Ukrainian-adapted Gemma 3 model, tested in both base and instruction-tuned variants.

Smaller multilingual models. We include **Gemma 3** (1B, 4B, 12B), **Llama 3.1-8B**, **Phi-4-Mini** (3.8B), and **Qwen3-8B** in reasoning mode. The latter is scored with extractive match only, as its chain-of-thought outputs are incompatible with log-likelihood evaluation.

Pretrained base models. To isolate the contribution of instruction tuning, we additionally include **Gemma 3 4B PT**, **Gemma 3 12B PT**, and **Llama 3.1 8B PT**, alongside **Lapa 12B PT**.

Large-scale multilingual models. We also evaluated larger models with 24B parameters or more on the proposed benchmark. These models are typically stronger on complex tasks and longer-context settings. In this group, we include GPT-OSS-120B, Mistral-Medium-2508, Mistral-Small-2506, and Llama-3.3-70B. All of these models are evaluated using the setup described in Section 4.

4.5 Prompt Design

Recall from Section 1 that our goal is to assess core linguistic knowledge rather than conditioned behavior, hence, we restrict experimentation to a standard prompt and vary only the prompt language. Persona-based prompting (Tan et al., 2024) is also not used, as it is unlikely to meaningfully improve grammatical competence but instead can lead to a superficial stylistic imitation. Moreover, prior work suggests that shorter prompts often lead to better performance, potentially because highly detailed instructions can overconstrain model behavior (Wang et al., 2026). Research shows that some LLMs might follow instructions in English better in the

A. PROMPT_EN *You are answering a multiple-choice question in Ukrainian about Ukrainian grammar and orthography.* **Return:** `answer:` ONLY the letter of the correct option (e.g., “А”, “Б”, “В”, “Г”, “Д”). **Question:** `{question}` **Options:** `{choices}`.

B. PROMPT_UA Дай відповідь на тестове запитання українською мовою з української граматики та орфографії. Вкажи: Відповідь: ЛИШЕ літеру правильної відповіді (наприклад, “А”, “Б”, “В”, “Г”, “Д”). Запитання: `{question}` Варіанти: `{choices}`.

C. PROMPT_UA_TRANSLIT *Dai vidpovid na testove zapytannia ukrainskoiu movoiu z ukrainskoi hramatyky ta orfohrafii.* **Vkazhy:** `Vidpovid:` LYSHE literu pravylnoi vidpovidi (napryklad, “A”, “B”, “V”, “H”, “D”). **Zapytannia:** `{question}` **Varianty:** `{choices}`.

Figure 3: English, Ukrainian, and transliterated Ukrainian prompt templates used in the experiments.

tasks of emotion detection (Dementieva et al., 2025), (De Bruyne et al., 2022). We therefore compare only English and Ukrainian prompt wording for language proficiency as well. This comparison is not fully controlled in the few-shot setting, since the in-context examples are in Ukrainian.

We evaluate models in a zero-shot setting using only task instructions that require the model to generate a single-letter answer. We also conduct a few-shot evaluation to examine how performance changes when example items are included in the prompt. This is particularly important for pretrained models, since following task instructions may be more difficult for base models prior to fine-tuning; few-shot prompting therefore helps mitigate this bias.

However, sampling test items for few-shot evaluation is not straightforward, as little research provides clear guidance on best practices. Tang et al. (2025) investigate how adding examples helps reduce ambiguity. They conclude that 5–20 examples is a sweet spot for the models used in our experiments, particularly LLaMA-3.1-8B and Gemma-3-4B. Using more examples may lead to over-prompting and performance degradation, potentially due to difficulties in handling longer contexts. They also identify three dominant sampling strategies: random sampling, sampling similar questions with TF-IDF, and sampling similar questions with semantic embeddings (Tang et al., 2025). We chose the second strategy and therefore employ five examples consisting of almost identical questions, each with different answer options and a different correct answer. Figure 4 shows the five examples sampled from the initial dataset, which were subsequently excluded from the main LLM evaluation. We argue that additional few-shot examples are unnecessary, as the primary role of prompt examples in our setting is to reinforce the instructions and constrain the model’s output format. Unlike tasks involving domain-specific classes, our multiple-choice setting only requires the model to select among predefined answer options denoted by letters. Furthermore, adding more examples increases prompt length and may introduce long-context effects that negatively affect performance.

Ex1 Q: У родовому відмінку однини закінчення -у (-ю) мають усі іменники в рядку. А) ансамбль, фольклор, Дон, дуб; Б) Буг, роман, Кавказ, мішок; В) Сибір, зошит, дощ, оркестр; Г) сніг, ураган, біль, понеділок; Д) гнів, Амур, сюжет, полк. А: Д

Ex2 Q: У родовому відмінку однини закінчення -у (-ю) мають усі іменники в рядку. А) автобус, хокей, шовк, жаль; Б) вальс, центнер, народ, відсоток; В) сум, інститут, міст, будинок; Г) футбол, університет, розрив, апарат (президента); Д) гектар, кілограм, вітер, гриб. А: Г

Ex3 Q: У родовому відмінку однини закінчення -у (-ю) мають усі іменники в рядку. А) організм, легіт, рейд, цемент; Б) спосіб, ясен, поле, ліс; В) трактор, метр, колір, Лондон; Г) реалізм, лозунг, рис, космос; Д) нуль, ситець, літр, мільйон. А: А

Ex4 Q: У родовому відмінку однини закінчення -а (-я) мають усі іменники в рядку. А) вокзал, атом, атлас, мороз; Б) жираф, вересень, перпендикуляр, конус; В) патріарх, переліг, театр, краєвид; Г) ерудит, Острог, край, овес; Д) материк, мікроб, ведмідь, поклик. А: Б

Ex5 Q: У родовому відмінку однини закінчення -а (-я) мають усі іменники в рядку. А) медик, вечір, комп’ютер, Берлін; Б) тигр, понеділок, прапор, рейс; В) прямокутник, підручник, кілограм, барометр; Г) водоспад, Дунай, цирк, задум; Д) егоїст, десяток, катод, порох. А: В

Figure 4: Five few-shot examples used in the prompt. All examples follow the same multiple-choice format but differ in choices and the correct answer. English translation and transliteration are provided in Appendices B and C

5 Results

Table 2 reports results across all model groups. Overall, performance stays below 43% for most models, suggesting that standard Ukrainian morphology, syntax, and orthography are still challenging for current LLMs, including models adapted for Ukrainian. Among the larger models, Mistral-Medium-2508 reaches 46.7–49.1%, while GPT-OSS-120B is the strongest model overall, exceeding 54% in several settings.

Ukrainian-focused vs. general multilingual models. **Lapa 12B IT** is the strongest Ukrainian-focused model, reaching 42.1% XM in the 5-shot setting. **MamayLM** scores consistently lower despite sharing the same Gemma 3 base family. This

suggests that adaptation data and tuning strategy matter more than the underlying base model alone. In the multilingual group, **Gemma 3 12B IT** is the strongest model and comes close to Lapa 12B IT in several settings, even though it is not fine-tuned specifically for Ukrainian.

Results with large-scale models. The Ukrainian grammar proficiency of four instruction-tuned large-scale language models—GPT-OSS-120B, Llama-3.3-70B, Mistral-Medium-2508, and Mistral-Small-2506—is evaluated using Extractive Match and Exact Match under zero-shot and few-shot settings. Table 2 reports the results. GPT-OSS-120B achieves the best performance, with its highest accuracy obtained in the zero-shot setting with English instructions. Its accuracy decreases under few-shot prompting. We hypothesise that this effect may emerge from the multilingual nature of the prompt, where English instructions are combined with Ukrainian examples. The drop in accuracy is approximately 5 percentage points. However, the gains from few-shot prompting for other large models, such as Mistral-Medium-2508 and Mistral-Small-2506, are statistically insignificant. A similar pattern is observed across the other models we evaluate. This finding supports our hypothesis stated in the Experimental Setup: although prompting examples may help guide the model toward the expected output format, they add limited value in the MCQ setting, where the target classes (letters) do not carry semantic meaning.

Per-question results are further used to assign a hardness level to each test item. We categorize question hardness according to the number of correct generations out of four large-scale models as these models show the best overall accuracy. If no model answers a question it is labeled as hard, those answered correctly by exactly one model are labeled difficult, those answered correctly by two models medium, and those answered correctly by three or four models easy. This discrete definition avoids arbitrary thresholds and directly reflects inter-model agreement. The resulting distribution is shown in Table 1. We suggest that hard and difficult questions are challenging because they involve specific linguistic patterns or highly plausible distractors that require more fine-grained Ukrainian proficiency.

Effect of prompt language and few-shot examples. English and Ukrainian prompts lead to similar results for most models, usually within a few percentage points, but the effect of few-shot prompting is not consistent across model families. For example, under English prompting, **MamayLM 4B IT** improves from 30.1% to 35.5% XM, while **Lapa 12B IT** improves from 38.9% to 42.1%. The few-shot setting remains important for pretrained models as in-context examples help follow the required answer format.

Table 1: Distribution of question hardness in the dataset.

Level	Count	%
Hard	70	20.1
Difficult	90	26.0
Medium	88	25.3
Easy	99	28.6
Total	347	100.0

Reasoning model behavior. **Qwen3-8B (RSN)** shows a clear gap between Extractive Match and Exact Match. Under Ukrainian 5-shot prompting, it reaches 34.3% XM but only 6.4% EM. This suggests that the model often includes the correct answer in its output but does not follow the required answer format. For this reason, XM appears to be the more suitable metric for this model.

LLaMA: Sensitivity to Instruction Language. We observe that LLaMA-based model consistently performs better when the task instructions are written in English rather than Ukrainian, even though the questions and answer options are still in Ukrainian.

This does not necessarily mean that the model has stronger Ukrainian language ability. A more likely explanation is that it follows instructions more reliably in English: it is more likely to produce the required answer format and respect the multiple-choice setup.

6 Limitations and Future Directions

The following limitations and possible mitigation strategies constitute future directions for improving LLM evaluation on the proposed benchmark.

Prompt optimization. The current setup uses a fixed prompt and five few-shot examples to ensure consistency across model types. However, experiments with more prompts and their customization for each model might better showcase its strengths (see (Khattab et al., 2024)), since models are trained

Table 2: Evaluation results on the proposed Ukrainian language benchmark (↑, all values in %). **LL** = log-likelihood accuracy (zero-shot); **XM** = extractive match; **EM** = exact match. Generative results are reported in zero-shot (FS0) and 5-shot (FS5) settings, each under English (EN) and Ukrainian (UK) prompt language. Dashes indicate settings not evaluated for a given model. Highlighted cells mark the best result per column within each model group.

	LL (zero-shot)		**Extractive Match**				**Exact Match**	
			FS5		FS0		FS5	
Model	EN	UK	EN	UK	EN	UK	EN	UK
Ukrainian-focused models								
MamayLM 4B IT	29.6	30.5	35.5	34.4	30.1	31.6	35.5	34.4
MamayLM 12B IT	37.5	36.3	39.1	38.1	37.1	34.6	39.1	38.1
Lapa 12B IT	**38.7**	**37.6**	**42.1**	**41.1**	**38.9**	**38.4**	**42.1**	**41.1**
Lapa 12B PT	35.0	31.0	38.4	38.6	–	–	38.4	38.6
Smaller multilingual instruction-tuned models								
Gemma 3 1B IT	23.2	27.5	21.7	22.1	21.7	21.3	21.7	22.1
Gemma 3 4B IT	25.6	33.6	28.5	32.9	26.2	33.9	28.5	32.9
Gemma 3 12B IT	**36.2**	**35.8**	**36.3**	**37.0**	**37.6**	**36.2**	**36.3**	**37.0**
Phi-4-Mini IT	26.0	27.0	24.8	27.2	25.2	26.9	19.6	24.5
Llama 3.1 8B IT	29.9	25.4	30.1	29.6	31.6	27.3	30.1	29.6
Pretrained base models								
Gemma 3 4B PT	27.4	27.1	28.9	29.1	–	–	28.9	29.1
Gemma 3 12B PT	**29.8**	**32.3**	**35.8**	**37.8**	–	–	**35.8**	**37.8**
Llama 3.1 8B PT	25.3	21.6	31.9	25.0	–	–	31.9	25.0
Reasoning model								
Qwen3-8B (RSN)	–	–	**35.2**	**34.3**	**38.2**	**35.3**	**35.2**	6.4
Large-scale multilingual models								
gpt-oss-120b	–	–	**55.0**	**55.2**	**59.6**	**54.7**	**55.0**	**55.2**
Meta-Llama-3.3-70B-Instruct	–	–	36.9	34.9	36.3	33.7	36.9	34.9
mistral-medium-2508	–	–	48.5	49.1	46.7	47.2	48.5	49.1
mistral-small-2506	–	–	37.2	41.0	34.3	38.6	33.5	39.4

on different data and with different architectural parameters.

Reasoning models and log-likelihood evaluation. Reasoning models were not evaluated in the log-likelihood setup because their long thinking traces do not fit this scoring method well. Future work could explore adapted evaluation protocols that would allow more direct comparison with other model types.

Answer-label bias. Fixed answer labels such as A/B/C/D may introduce label and positional bias. Future work should test alternative label formats and full-answer generation to reduce this effect (Nowak et al., 2026).

Generation size calibration. In the generative setup, the maximum output length for standard models was limited to 15 tokens. This is longer than needed to produce a single answer letter, but it makes it possible to capture cases where the correct answer appears later in the output. At the same time, longer generations increase the risk of extraction errors, since regex-based metrics may recover a letter that does not reflect the model's final choice. Future work should study this trade-off more systematically and determine a more principled generation limit for this type of benchmark.

7 Discussion and Conclusion

We introduced an expert-curated benchmark for evaluating Ukrainian language proficiency in large language models. The benchmark contains 347 multiple-choice questions covering morphology, syntax, vocabulary, and orthography.

Our results show that the benchmark is challenging for current models. In most evaluated settings with smaller LLMs, performance remains below 43%, including for models adapted specifically for Ukrainian. This suggests that general multilingual ability does not guarantee reliable knowledge of Ukrainian grammar rules. Few-shot prompting helps mainly for pretrained models, and the choice of prompt language had little consistent effect. Large-scale models exhibit better performance (GPT-OSS-120B, Mistral-Medium-2508) due to their improved training and capacity to detect complex patterns. However, even the best-performing model demonstrates maximum accuracy of approximately 60%.

In conclusion, the benchmark offers a focused resource for evaluating how well language models handle the Ukrainian grammar and orthography and provides a basis for future work on Ukrainian evaluation and model development. More broadly, our findings also highlight the importance of expert-designed benchmarks for underrepresented languages. The dataset is also openly accessible to the community to advance further research in Ukrainian natural language processing (see ULP[1]).

Acknowledgments

We would like to thank the Kyivstar team for testing our dataset within the Kyivstar evaluation framework. We are also grateful to Denys Yurchenko for his helpful comments and suggestions.

References

Lapa llm. https://huggingface.co/lapa-llm/lapa-12b-pt. Hugging Face repository.

Maksym Bondarenko, Artem Yushko, Andrii Shportko, and Andrii Fedorych. 2023. Comparative study of models trained on synthetic data for ukrainian grammatical error correction. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 103–113.

Luna De Bruyne, Pranaydeep Singh, Orphée De Clercq, Els Lefever, and Véronique Hoste. 2022. How language-dependent is emotion detection? evidence from multilingual bert. In *Proceedings of the 2nd Workshop on Multi-lingual Representation Learning (MRL)*, pages 76–85.

Daryna Dementieva, Nikolay Babakov, and Alexander Fraser. 2025. Emobench-ua: A benchmark dataset for emotion detection in ukrainian. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Saiful Haq, Ashutosh Sharma, Thomas Joshi, Hanna Moazam, Heather Miller, and 1 others. 2024. Dspy: compiling declarative language model calls into state-of-the-art pipelines. In *International Conference on Learning Representations*, volume 2024, pages 54928–54958.

Zoriana Matsiuk and Nina Stankevych. 2017. Ukrainska mova profesiinoho spilkuvannia. *K.: Karavela*.

Ye. I. Maznichenko, V. Ye. Makedon, S. V. Sharabanova, and I. L. Yalovnycha. 2019. *Ykrainsky pravopus*. Kyiv: Instytut movoznavstva imeni O. O. Potebni Natsionalnoi akademii nauk Ukrainy.

Mateusz Nowak, Xavier Cadet, and Peter Chin. 2026. Abcd: All biases come disguised. *arXiv preprint arXiv:2602.17445*.

[1] https://huggingface.co/datasets/SGaleshchuk/ULP-Ukrainian-Language-Proficiency

Mariana Romanyshyn, Oleksiy Syvokon, and Roman Kyslyi. 2024. The unlp 2024 shared task on fine-tuning large language models for ukrainian. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP)@ LREC-COLING 2024*, pages 67–74.

Aman Saini, Artem Chernodub, Vipul Raheja, and Vivek Kulkarni. 2024. Spivavtor: An instruction tuned ukrainian text editing model. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP)@ LREC-COLING 2024*, pages 95–108.

Oleksandra Serbenska, editor. 2019. *Antysurzhyk. Vchymosia vvichlyvo povodytys i pravylno hovoryty*. Svit, Lviv. Navchalnyi posibnyk.

Oleksiy Syvokon, Olena Nahorna, Pavlo Kuchmiichuk, and Nastasiia Osidach. 2023. Ua-gec: Grammatical error correction and fluency corpus for the ukrainian language. In *Proceedings of the second Ukrainian natural language processing workshop (UNLP)*, pages 96–102.

Fiona Anting Tan, Gerard Christopher Yeo, Kokil Jaidka, Fanyou Wu, Weijie Xu, Vinija Jain, Aman Chadha, Yang Liu, and See-Kiong Ng. 2024. Phantom: Persona-based prompting has an effect on theory-of-mind reasoning in large language models. *arXiv preprint arXiv:2403.02246*.

Yongjian Tang, Doruk Tuncel, Christian Koerner, and Thomas Runkler. 2025. The few-shot dilemma: Over-prompting large language models. *arXiv preprint arXiv:2509.13196*.

Mariia Voloshchak. 2007. *Nepravylno–pravylno: dovidnyk z ukrainskoho slovovzhyvannia*, 2 edition. Prosvita and Ukrainska Vydavnycha Spilka, Kyiv.

Qile Wang, Prerana Khatiwada, Avinash Chouhan, Ashrey Mahesh, Joy Mwaria, Duy Duc Tran, Kenneth E Barner, and Matthew Louis Mauriello. 2026. " the explanation makes sense": An empirical study on llm performance in news classification and its influence on judgment in human-ai collaborative annotation. *arXiv preprint arXiv:2602.19690*.

Hanna Yukhymenko, Anton Alexandrov, and Martin Vechev. 2025. Mamaylm: An efficient state-of-the-art ukrainian llm. https://huggingface.co/blog/INSAIT-Institute/mamaylm.

A Representative Benchmark Items

This appendix presents details on representative benchmark questions in the original Ukrainian, together with transliteration and an English translation of each question stems.

A.1 Morphology

Nouns. Genitive case endings -а(я)/-у(ю) **in masculine singular nouns.**
У котрому рядку всі іменники в родовому відмінку однини мають закінчення -у (-ю)? – а) трамвай, бік, університет, Буг; б) Дніпро, вокзал, стан, садок; в) Рим, термін, гіпс, соняшник; г) міст, дах, Дністер, Сибір.
English translation: In which row do all masculine singular nouns take the genitive ending *-u (-iu)*?
Transliteration: U kotromu riadku vsi imennyky v rodovomu vidminku odnyny maiut zakinchennia -u (-iu)? – a) tramvai, bik, universytet, Buh; b) Dnipro, vokzal, stan, sadok; v) Rym, termin, hips, soniashnyk; h) mist, dakh, Dnister, Sybir.

Instrumental case endings.
У котрому рядку форми орудного відмінка іменника утворено правильно? – а) пишаюсь ім'ям; б) пливе Черемошом; в) поснідав кашою; г) милувався вежою.
English translation: In which row is the instrumental case form of the noun formed correctly?
Transliteration: U kotromu riadku formy orudnoho vidminka imennyka utvoreno pravylno? – a) pyshaius imiam; b) plyve Cheremoshom; v) posnidav kashoiu; h) myluvavsia vezhoiu.

Vocative case endings.
У котрому рядку усі форми звертання правильні? – а) Пані Оксано; дорога бабусю; шановний колего; б) Вельмишановні учасники, пане ректору, Ілле Пилиповичу; в) Олеже Андрійовичу, Настю, Гале; г) любий друже, товаришe, Саво Петровиче.
English translation: In which row are all vocative forms correct?
Transliteration: U kotromu riadku usi formy zvertannia pravylni? – a) Pani Oksano; doroha babusiu; shanovnyi koleho; b) Velmyshanovni uchasnyky, pane rektoru, Ille Pylypovychu; v) Olezhe Andriiovychu, Nastiu, Hale; h) liubyi druzhe, tovaryshe, Savo Petrovyche.

Genitive plural endings.
У котрому рядку форми родового відмінка множини іменника утворено правильно? – а) статтей; б) суддей; в) бур; г) узвишшів.
English translation: In which row is the genitive plural form of the noun formed correctly?
Transliteration: U kotromu riadku formy rodovoho vidminka mnozhyny imennyka utvoreno pravylno? – a) stattei; b) suddei; v) bur; h) uzvyshshiv.

Singular and plural forms of nouns.
У котрому рядку всі іменники мають форму однини і множини? – а) задум, маніпуляція, свідчення, доба; б) радість, задума, досвід, чистота; в) досягнення, команда, аспірин, Марія; г) вода, зима, азот, Херсон.
English translation: In which row do all nouns have both singular and plural forms?
Transliteration: U kotromu riadku vsi imennyky maiut formu odnyny i mnozhyny? – a) zadum, manipuliatsiia, svidchennia, doba; b) radist, zaduma, dosvid, chystota; v) dosiahnennia, komanda, aspiryn, Mariia; h) voda, zyma, azot, Kherson.

Gender forms of invariable foreign origin nouns and abbreviations.
Граматичну норму дотримано в рядку: а) великий НЛО, гарний Тбілісі; б) ВООЗ заборонила, молода івасі; в) чистий Онтаріо, УПА боролася; г) сильний сироко, гарне кенгуру.
English translation: In which row is the grammatical norm observed?
Transliteration: Hramatychnu normu dotrymano v riadku: a) velykyi NLO, harnyi Tbilisi; b) VOOZ zaboronyla, moloda ivasi; v) chystyi Ontario, UPA borolasia; h) sylnyi syroko, harne kenhuru.

Adjectives. Comparative and superlative forms.
У котрому рядку подано правильні форми вищого й найвищого ступеня порівняння прикметників? – а) більш дешевший; б) довгуватіший; в) довжелезний; г) якнайкращий.
English translation: In which row are the comparative and superlative forms of the adjectives correct?
Transliteration: U kotromu riadku podano pravylni formy vyshchoho y naivyshchoho stupenia porivniannia prykmetnykiv? – a) bilsh deshevshyi; b) dovhuvatishyi; v) dovzheleznyi; h) yaknaikrashchyi.

Numerals. Combination of numerals with nouns.
У котрому словосполученні правильно узгоджено числівник з іменником? – а) півтори літри; б) близько двадцяти гривнів; в) чотири з половиною дні; г) півтора дні.
English translation: In which phrase is the numeral correctly combined with the noun?
Transliteration: U kotromu slovospoluchenni pravylno uzghodzheno chyslivnyk z imennykom? – a) pivtory litry; b) blyzko dvadtsiaty hryvniv; v) chotyry z polovynoiu dni; h) pivtora dni.

Case forms of numerals.
У котрому варіанті подано правильні відмінкові форми числівників? – а) трьохстам учасникам; б) ста мовами; в) восьмидесяти років; г) із восьмистами студентами.
English translation: In which option are the case forms of the numerals correct?
Transliteration: U kotromu varianti podano pravylni vidminkovi formy chyslivnykiv? – a) trokhstam uchasnykam; b) sta movamy; v) vosmydesiaty rokiv; h) iz vosmystamy studentamy.

Use of numerals to indicate time.
У котрому рядку допущено помилку у відповіді на питання «Котра година?» – а) сім хвилин на шосту; б) вісім годин; в) тридцять хвилин по першій; г) чверть на дванадцяту.
English translation: In which row is there an error in answering the question “What time is it?”
Transliteration: U kotromu riadku dopushcheno pomylku u vidpovidi na pytannia “Kotra hodyna?” – a) sim khvylyn na shostu; b) visim hodyn; v) trydtsiat khvylyn po pershii; h) chvert na dvanadtsiatu.

Pronouns. Normative use of pronouns.
Де займенник вжито правильно? – а) їх робота; б) їхня робота; в) чиєгось місця; г) на мойому місці.
English translation: In which option is the pronoun used correctly?
Transliteration: De zaimennyk vzhyto pravylno? – a) yikh robota; b) yikhnia robota; v) chyiehos mistsia; h) na moiomu mistsi.

Verbs. Personal verb forms.
У котрому рядку дієслово має правильне особове закінчення? – а) мелють каву; б) мелять каву; в) ревлять турбіни; г) купляють взуття.
English translation: In which row does the verb have the correct personal ending?
Transliteration: U kotromu riadku diieslovo maie pravylne osobove zakinchennia? – a) meliut kavu; b) meliat kavu; v) revliat turbiny; h) kupliaiut vzuttia.

Forms of the future tense.
У котрому рядку подано правильні форми майбутнього часу дієслова? – а) продаш мені книжку; б) з’їсиш борщ; в) розповіси казку; г) буде читав казку.

English translation: In which row are the future-tense forms of the verb correct?
Transliteration: U kotromu riadku podano pravylni formy maibutnoho chasu diieslova? – a) prodash meni knyzhku; b) zisysh borshch; v) rozpovisy kazku; h) bude chytav kazku.

Forms of the imperative mood.
У котрому рядку подано правильні форми наказового способу дієслова? – а) ходімте з нами; б) давай зустрінемось; в) розповіси нам історію; г) берімось до роботи.
English translation: In which row are the imperative forms of the verb correct?
Transliteration: U kotromu riadku podano pravylni formy nakazovoho sposobu diieslova? – a) khodimte z namy; b) davai zustrinemos; v) rozpovisy nam istoriiu; h) berimos do roboty.

Standard verb forms: participles and gerunds.
Котре словосполучення відповідає нормі? – а) працююче населення; б) захоплюючий фільм; в) тремтячий голос; г) керуючий банком.
English translation: Which phrase conforms to the standard norm?
Transliteration: Kotre slovospoluchennia vidpovidaie normi? – a) pratsiuiuche naselennia; b) zakhopliuiuchyi film; v) tremtiachyi holos; h) keruiuchyi bankom.

A.2 Syntax

Word-level combinations. Prepositional government.
У котрому словосполученні правильно вжито прийменник при? – а) виявилось при дослідженні; б) було при Б. Хмельницькому; в) говорити при свідках; г) при допомозі ліків.
English translation: In which phrase is the preposition *pry* used correctly?
Transliteration: U kotromu slovospoluchenni pravylno vzhyto pryimennyk pry? – a) vyiavylos pry doslidzhenni; b) bulo pry B. Khmelnytskomu; v) hovoryty pry svidkakh; h) pry dopomozi likiv.

Nonprepositional government.
Котре словосполучення відповідає нормі? – а) хворіє грипом; б) говорить на англійській; в) навчається музики; г) оплачує за проїзд.
English translation: Which phrase conforms to the standard norm?
Transliteration: Kotre slovospoluchennia vidpovidaie normi? – a) khvoriie hrypom; b) hovoryt na anhliiskii; v) navchaietsia muzyky; h) oplachuie za proizd.

Complex cases of agreement in word combinations.
У котрому рядку є приклади порушення норм узгодження? – а) на станції Бахмут, на вулиці Хрещатику; б) у штаті Вірджинія, на вулиці Зеленій; в) у місті Броди, нова стаття-дослідження; г) новий допис-оприлюднення, у Карпатах.
English translation: In which row are there examples of violations of agreement norms?
Transliteration: U kotromu riadku ye pryklady porushennia norm uzghodzhennia? – a) na stantsii Bakhmut, na vulytsi Khreshchatyku; b) u shtati Virdzhyniia, na vulytsi Zelenii; v) u misti Brody, nova stattia-doslidzhennia; h) novyi dopys-opryliudnennia, u Karpatakh.

Sentence-level syntax. Detached adverbial modifiers expressed by participial phrases.
Неправильно побудовано речення з дієприслівником: а) Слухаючи доповідь лектора, не забувайте робити нотатки; б) Створюючи проєкт, він виявився дуже цікавим; в) Прочитавши лекцію, професор вийшов; г) Ще не навчаючись в університеті, я вивчав географічні карти.
English translation: Which sentence with a verbal adverb is constructed incorrectly?
Transliteration: Nepravylno pobudovano rechennia z diiepryslivnykom: a) Slukhaiuchy dopovid lektora, ne zabuvaite robyty notatky; b) Stvoriuiuchy proiekt, vin vyiavyvsia duzhe tsikavym; v) Prochytavshy lektsiiu, profesor vyishov; h) Shche ne navchaiuchys v universyteti, ya vyvchav heohrafichni karty.

Series of homogeneous sentence parts.
Порушено норми побудови рядів однорідних членів речення у рядку: а) Треба вивчати іноземні мови і спілкуватися ними, щоб знати; б) Застосувати цю технологію можна на різних майданчиках, сценах і локаціях міста; в) Будь-які пари – лекції чи семінари – важливі; г) Усі викладачі та студенти взяли участь у конференції.
English translation: In which row are the norms for constructing series of homogeneous sentence parts violated?
Transliteration: Porusheno normy pobudovy riadiv odnoridnykh chleniv rechennia u riadku: a)

Treba vyvchaty inozemni movy i spilkuvatysia nymy, shchob znaty; b) Zastosuvaty tsiu tekhnolohiiu mozhna na riznykh maidanchykakh, stsenakh i lokatsiiakh mista; v) Bud-yaki pary – lektsii chy seminary – vazhlyvi; h) Usi vykladachi ta studenty vzialy uchast u konferentsii.

Agreement of the subject with the predicate.
Правильно узгоджено підмет із присудком у рядку: а) Більшість прийшли на модуль з математики; б) Багато днів минуло з того часу; в) Дехто з присутніх не вивчили матеріалу; г) Багато викладачів та студентів взяло участь у конференції.
English translation: In which row is the subject correctly agreed with the predicate?
Transliteration: Pravylno uzghodzheno pidmet iz prysudkom u riadku: a) Bilshist pryishly na modul z matematyky; b) Bahato dniv mynulo z toho chasu; v) Dekhto z prysutnikh ne vyvchyly materialu; h) Bahato vykladachiv ta studentiv vzialo uchast u konferentsii.

Norms for constructing complex sentences.
Яке речення збудоване без граматичних помилок? – а) Ми не прийшли, так як не мали змоги; б) Сьогодні представлять алгоритм дій, який створили географи, які були на конференції, яка була вчора; в) Андрій попросив колегу переглянути свою доповідь; г) Котрий з двох студентів, які брали участь у змаганні, посів призове місце?
English translation: Which sentence is constructed without grammatical errors?
Transliteration: Yake rechennia zbudovane bez hramatychnykh pomylok? – a) My ne pryishly, tak yak ne maly zmohy; b) Sohodni predstavliat alhorytm dii, yakyi stvoryly heohrafy, yaki buly na konferentsii, yaka bula vchora; v) Andrii poprosyv kolehu perehlianuty svoiu dopovid; h) Kotryi z dvokh studentiv, yaki braly uchast u zmahanni, posiv pryzove mistse?

A.3 Vocabulary

Word usage. Correctness of word combinations by meaning.
Котре словосполучення семантично правильне? – а) перевернути сторінку; б) перевернути стілець; в) носити назву; г) приступати до роботи.
English translation: Which phrase is semantically correct?
Transliteration: Kotre slovospoluchennia semantychno pravylne? – a) perevernuty storinku; b) perevernuty stilets; v) nosyty nazvu; h) prystupaty do roboty.

Distinguishing word meaning.
Котре слово є синонімом до безпідставний? – а) безуспішний; б) голослівний; в) даремний; г) неузгоджений.
English translation: Which word is a synonym of *bezpidstavnyi* (“groundless / unfounded”)?
Transliteration: Kotre slovo ye synonimom do bezpidstavnyi? – a) bezuspishnyi; b) holoslivnyi; v) daremnyi; h) neuzghodzhenyi.

A.4 Orthography

Counting violations of orthographic norms.
Скільки порушень мовних норм у реченні: Українські вчені докладають багато зусилль, щоб врятувати еко-систему. – а) 2; б) 3; в) 4; г) 5. Відповідь: Б.
English translation: How many violations of language norms are there in the sentence: “Ukrainski vcheni dokladaiut bahato zusyll, shchob vriatuvaty eko-systemu.” *Answer:* B.
Transliteration: Skilky porushen movnykh norm u rechenni: Ukrainski vcheni dokladaiut bahato zusyll, shchob vriatuvaty eko-systemu. – a) 2; b) 3; v) 4; h) 5. *Vidpovid:* B.

B Few-Shot Examples English Translation

1. *Question:* In the genitive singular, all nouns in the row take the ending *-u (-iu)*.

   A) ensemble, folklore, Don, oak
   B) Buh, novel, Caucasus, sack
   V) Siberia, notebook, rain, orchestra
   H) snow, hurricane, pain, Monday
   D) anger, Amur, plot, regiment

   Correct answer: D

2. *Question:* In the genitive singular, all nouns in the row take the ending *-u (-iu)*.

   A) bus, hockey, silk, sorrow
   B) waltz, centner, people, percent
   V) sadness, institute, bridge, building
   H) football, university, rupture, apparatus (of the president)
   D) hectare, kilogram, wind, mushroom

   Correct answer: H

3. *Question:* In the genitive singular, all nouns in the row take the ending *-u (-iu)*.

   A) organism, breeze, raid, cement
   B) manner, ash tree, field, forest
   V) tractor, meter, color, London
   H) realism, slogan, rice, cosmos
   D) zero, chintz, liter, million

   Correct answer: A

4. *Question:* In the genitive singular, all nouns in the row take the ending *-a (-ia)*.

   A) station, atom, atlas, frost
   B) giraffe, September, perpendicular, cone
   V) patriarch, fallow land, theater, landscape
   H) erudite, Ostroh, region, oats
   D) mainland, microbe, bear, call

   Correct answer: B

5. *Question:* In the genitive singular, all nouns in the row take the ending *-a (-ia)*.

   A) medic, evening, computer, Berlin
   B) tiger, Monday, flag, voyage
   V) rectangle, textbook, kilogram, barometer
   H) waterfall, Danube, circus, intention
   D) egoist, ten-item set, cathode, gunpowder

   Correct answer: V

Figure 5: English version of the five few-shot examples used in the prompt.

C Few-Shot Examples Transliteration

1. *Question:* U rodovomu vidminku odnyny zakinchennia -u (-iu) maiut usi imennyky v riadku

   A) ansambl, folklor, Don, dub
   B) Buh, roman, Kavkaz, mishok
   V) Sybir, zoshyt, doshch, orkestr
   H) snih, urahan, bil, ponedilok
   D) hniv, Amur, siuzhet, polk

   Correct answer: D

2. *Question:* U rodovomu vidminku odnyny zakinchennia -u (-iu) maiut usi imennyky v riadku

   A) avtobus, khokei, shovk, zhal
   B) vals, tsentner, narod, vidsotok
   V) sum, instytut, mist, budynok
   H) futbol, universytet, rozryv, aparat (prezydenta)
   D) hektar, kilohram, viter, hryb

   Correct answer: H

3. *Question:* U rodovomu vidminku odnyny zakinchennia -u (-iu) maiut usi imennyky v riadku

   A) orhanizm, lehit, reid, tsement
   B) sposib, yasen, pole, lis
   V) traktor, metr, kolir, London
   H) realizm, lozunh, rys, kosmos
   D) nul, sytets, litr, milion

   Correct answer: A

4. *Question:* U rodovomu vidminku odnyny zakinchennia -a (-ia) maiut usi imennyky v riadku

   A) vokzal, atom, atlas, moroz
   B) zhyraf, veresen, perpendykuliar, konus
   V) patriarkh, perelih, teatr, kraievyd
   H) erudyt, Ostroh, krai, oves
   D) materyk, mikrob, vedmid, poklyk

   Correct answer: B

5. *Question:* U rodovomu vidminku odnyny zakinchennia -a (-ia) maiut usi imennyky v riadku

   A) medyk, vechir, komp'iuter, Berlin
   B) tyhr, ponedilok, prapor, reis
   V) priamokutnyk, pidruchnyk, kilohram, barometr
   H) vodospad, Dunai, tsyrk, zadum
   D) ehoist, desiatok, katod, porokh

   Correct answer: V

Figure 6: Transliterated version of the five few-shot examples used in the prompt.

D Models Information

Model	Variant	Setting	Category
MamayLM	4B	IT	Ukrainian-focused
MamayLM	12B	IT	Ukrainian-focused
Lapa	12B	IT	Ukrainian-focused
Lapa	12B	PT	Ukrainian-focused
Gemma 3	1B	IT	Multilingual
Gemma 3	4B	IT	Multilingual
Gemma 3	12B	IT	Multilingual
Phi-4-Mini	3.8B	IT	Multilingual
Llama 3.1	8B	IT	Multilingual
Gemma 3	4B	PT	Base model
Gemma 3	12B	PT	Base model
Llama 3.1	8B	PT	Base model
Qwen3	8B	RSN	Reasoning model
GPT-OSS	120B	IT	Large models
Mistral-medium	2508	IT	Large models
Mistral-small	2506	IT	Large models
Llama 3.3	70B	IT	Large models

Table 3: Models evaluated in this work. IT denotes instruction-tuned, PT denotes pretrained, and RSN denotes reasoning.

How Far Can Prompting Go for Minimal-Edit Ukrainian Grammatical Error Correction?

Kateryna Karpo[υ,σ] **Artem Chernodub**[ζ]
[υ]Ukrainian Catholic University [σ]YouScan [ζ]Zendesk

Abstract

Fine-tuned Large Language Models (LLMs) dominate in Ukrainian grammatical error correction (GEC), while API-accessed LLMs remain nearly untested on minimal-edit benchmarks. We evaluate 11 commercial LLMs from four providers and one open-source Ukrainian model on the UNLP 2023 Shared Task benchmark, GEC-only track, comparing zero-shot, few-shot, minimal-edits, and LLM-assisted prompt optimization strategies. Our best configuration (Gemini 3.1-Pro) reaches $F_{0.5}$ = 69.22, closing over 90% of the gap to fine-tuned SOTA ($F_{0.5}$ = 73.14). For zero-shot prompts, only Claude models benefit from Ukrainian instructions. However, the best overall results for all models use Ukrainian minimal-edits prompts, whose language-specific rules require Ukrainian to express precisely. LLM-assisted prompt optimization on top of minimal-edits + few-shot achieves the highest score. Detailed minimal-edits instructions yield the largest gains for punctuation and case errors but cause the model to abandon several low-frequency categories. Delving into error analysis, we identify five recurring overcorrection patterns tied to Ukrainian-specific linguistic phenomena. Code, prompts, and outputs are publicly available.[1][2][3]

1 Introduction and Related Work

Grammatical error correction (GEC) systems operate under two paradigms (Bryant et al., 2023). Minimal-edit correction targets only clear grammatical, spelling, and punctuation errors, preserving the author's wording. Fluency-oriented correction additionally permits lexical substitutions, syntactic restructuring, and stylistic improvements. The minimal-edit setting is especially relevant for educational tools, where feedback should pinpoint errors rather than rewrite learner text, and for writing assistants that must preserve authorial voice.

For English, a high-resource language with decades of GEC research, this distinction is well established. Minimal-edit evaluation is the standard in shared tasks such as CoNLL-2014 (Ng et al., 2014) and BEA-2019 (Bryant et al., 2019), while JFLEG (Napoles et al., 2017) targets fluency. Staruch et al. (2025) recently achieved state-of-the-art single-model minimal-edit results on BEA-2019 by adapting a decoder-only LLM. The MultiGEC-2025 shared task (Masciolini et al., 2025) extended the two-track paradigm to twelve European languages, confirming it as a cross-lingual standard.

For Ukrainian, GEC infrastructure has only recently begun to emerge. The UNLP 2023 Shared Task (Syvokon and Romanyshyn, 2023) introduced the first benchmark with two parallel tracks: GEC-only (minimal-edit) and GEC+Fluency, both evaluated with span-based $F_{0.5}$. Since then, research has shifted toward fluency (Saini et al., 2024), with Luhtaru et al. (2024) pushing GEC+Fluency SOTA to $F_{0.5}$ = 74.09 with a fine-tuned Llama 2 model, surpassing the original winner ($F_{0.5}$ = 68.17; Bondarenko et al., 2023). The GEC-only track, however, has seen no new results. Ukrainian was also included in MultiGEC-2025, where the winning team's fine-tuned Gemma 2 scored GLEU = 79.55 on minimal edits vs. GLEU = 68.03 for a one-shot Llama 3.1 baseline. Most recently, Kovalchuk et al. (2025) introduced silver-standard GEC corpora for multiple languages including Ukrainian and fine-tuned multilingual models on them; however, their work centers on training data creation and finetuning rather than prompting strategies for API-accessed LLMs.

To the best of our knowledge, most of Ukrainian GEC systems available to date rely on fine-tuned models that require dedicated GPU infras-

[1]Correspondence: a.chernodub@gmail.com
[2]https://github.com/katerynkarpo/gec_unlp_2026
[3]This work was conducted as part of Kateryna Karpo's M.Sc. thesis at the Ukrainian Catholic University, Faculty of Applied Sciences.

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 136–154

tructure. Commercial API-accessed LLMs offer a lightweight alternative, yet remain nearly untested for Ukrainian minimal-edit GEC. The only published result is from Katinskaia and Yangarber (2024), who evaluated GPT-3.5 (specifically, `gpt-3.5-turbo-0613`) in a zero-shot setting on the UNLP 2023 Shared Task test set, GEC-only track (henceforth UNLP-2023-test (GEC only)), and obtained $F_{0.5} = 27.4$, far below the fine-tuned SOTA of 73.14.

This paper is the first to test newer API-accessed models on this benchmark and to explore whether better prompting strategies can close the gap with fine-tuned systems.

2 Experimental Setup

Data. We use the GEC-only track of the UNLP 2023 Shared Task (Syvokon and Romanyshyn, 2023), which is built on the UA-GEC corpus. We adopt UA-GEC's own train and valid splits as our training and validation sets (31,038 and 1,422 sentences) and report all final numbers on the shared task's test set (1,274 sentences), whose gold annotations are held out from participants and never inspected during prompt development. We use the training and validation sets for prompt development (few-shot exemplar selection and prompt engineering) and report only on the test set.

Models. We evaluate commercial, API-accessed LLMs from four providers and one open-source Ukrainian model, Lapa v0.1.2 (Paniv et al., 2025).[4] We report exact snapshot identifiers for reproducibility. From OpenAI, we use GPT-4.1 (`gpt-4.1-2025-04-14`), GPT-4.1-mini (`gpt-4.1-mini-2025-04-14`), GPT-5.1 (`gpt-5.1-2025-11-13`), GPT-5.2 (`gpt-5.2-2025-12-11`), and GPT-5.4 (`gpt-5.4-2026-03-05`). From Moonshot, we use Kimi-K2 (`kimi-k2-0905-preview`; `0905` denotes a dated preview build). The Google and Anthropic APIs do not expose dated snapshot identifiers; we use Gemini 3-Flash (`gemini-3-flash-preview`), Gemini 3-Pro (`gemini-3-pro-preview`), Gemini 3.1-Pro (`gemini-3.1-pro-preview`), Claude Sonnet 4.6 (`claude-sonnet-4.6`), and Claude Opus 4.6 (`claude-opus-4.6`).[5]

[4]Provider documentation: OpenAI `https://platform.openai.com`; Anthropic `https://docs.anthropic.com`; Google `https://ai.google.dev`; Moonshot `https://platform.moonshot.ai`.

[5]Inference-time parameters differ: GPT-4.1 and Kimi use temperature/top-p; Claude and Gemini use either temperature or a reasoning effort budget; GPT-5.x uses an effort level (`low`/`medium`/`high`). We set temperature 0 where available, default effort for Claude and Gemini, and `medium` for GPT-5.x.

2.1 Research Questions

We address the following four research questions:

RQ1: What is the zero-shot minimal-edit GEC performance of current LLMs on Ukrainian relative to fine-tuned SOTA, and how sensitive is it to prompt language? We systematically compare 2025–2026 commercial LLMs on UNLP-2023-test (GEC only) against the fine-tuned SOTA of $F_{0.5} = 73.14$, and test both English and Ukrainian prompt variants to assess whether instruction language affects correction quality for a morphologically rich, low-resource language. To the best of our knowledge, the only published LLM baseline for Ukrainian GEC is the GPT-3.5 zero-shot result (English) from Katinskaia and Yangarber (2024), which we include for reference.

RQ2: Can prompting strategies reduce overcorrection compared to zero-shot baselines? We evaluate how each of the four prompting strategies affects the precision–recall trade-off: (1) zero-shot, (2) few-shot, (3) minimal-edits + zero-shot, and (4) minimal-edits + few-shot.

RQ3: Can LLM-assisted prompt optimization improve over manually crafted prompts? We apply an LLM-assisted prompt optimization pipeline built on Claude Code skills, an agentic system that iteratively generates, evaluates, and refines GEC prompts using the full evaluation loop as feedback.

RQ4: Where do minimal-edits instructions help and where do they fail? We compare per-error-type performance between a standard zero-shot prompt and our best optimized prompt using ERRANT category breakdowns, identifying which error types benefit most from detailed minimal-edits instructions and which remain resistant to prompt-based improvement.

Prompting strategies. We compare four manually engineered prompting configurations that vary in prompt detail (general vs. minimal-edits) and use of examples (zero-shot vs. few-shot).

1. **Zero-shot** (A.1): a general system prompt that instructs the model to correct grammatical and spelling errors and return the original sentence if no errors are found. No examples are provided.

2. **Few-shot** (A.2): the zero-shot prompt augmented with source–target correction pairs from the training set, covering spelling, punctuation, and morphological errors as well as already-correct sentences.

3. **Minimal-edits + zero-shot** (A.3): a detailed system prompt enumerating which error types to correct, Ukrainian-specific conventions (e.g., dash vs. hyphen in dialogue, у/в 'u/v' alternation, vocative case in forms of address, etc.), and categories of changes to avoid. No examples are provided.

4. **Minimal-edits + few-shot** (A.4): combines the detailed minimal-edits system prompt with few-shot correction examples from the training set, providing both rule-based guidance and concrete demonstrations.

Strategies (1)–(4) are tested with both English (EN) and Ukrainian (UA) prompt text to address RQ1. For subsequent experiments (RQ2–RQ4), we use EN for zero-shot and few-shot prompts (where it performs best for most models; see RQ1) and UA for minimal-edits variants. This was an intentional design choice: the minimal-edits rules reference specific Ukrainian word forms, morphological categories, and language-specific conventions (e.g., vocative case paradigms, euphonic preposition alternation) that cannot be adequately expressed in English. Because the prompt language and strategy are tied together in this comparison, we treat them as a single design decision. To test whether an LLM can improve over these handcrafted prompts, we also apply LLM-assisted prompt optimization, inspired by automatic prompt optimization methods (see Ramnath et al., 2025, for a survey), on top of the best minimal-edits + few-shot prompt (Appendix A.5.2; RQ3).

Evaluation. We use the official UNLP-2023 evaluation pipeline (GEC only), which computes span-level Precision (P), Recall (R), and $F_{0.5}$ using a Ukrainian adaptation of ERRANT. Per-error-type scores are extracted from the ERRANT alignment for the per-error-type analysis (RQ4).

3 Prompt Design

Zero-shot. We use a single-sentence system prompt (Appendix A.1), adapted from Loem et al. (2023):

> `Reply with a corrected version of the sentence with all grammatical and spelling errors fixed. If there are no errors, reply with a copy of the original sentence. Input sentence: {sentence}. Corrected sentence:`

The Ukrainian version is its direct translation:

> Надай виправлену версію речення з виправленими всіма граматичними та орфографічними помилками. Якщо помилок немає, надай копію оригінального речення. Вхідне речення: {sentence}. Виправлене речення:

Few-shot. The few-shot prompt extends the zero-shot instruction with source–target correction pairs from the training set (Appendix A.2), e.g.:

> `[Same header as zero-shot prompt]`
> `Input:` Так само потерпає Україна і сьогодні від того що насправді талановитим людям заважають працювати...
> `Output:` Так само потерпає Україна і сьогодні від того, що насправді талановитим людям заважають працювати...
> *(Input: 'Ukraine suffers the same today from the fact that truly talented people are prevented from working...'*
> *Output: 'Ukraine suffers the same today from the fact, that truly talented people are prevented from working...')*

Exemplars are drawn from the UA-GEC training split because it is the only publicly available Ukrainian GEC corpus at the required scale and annotation quality. Since this split is public, it may have been seen by commercial LLMs during pretraining; drawing exemplars from an independent Ukrainian GEC corpus would be a cleaner control, but no comparable dataset currently exists. We therefore treat our numbers as establishing prompting baselines on this benchmark and revisit this risk in the Limitations section.

Minimal-edits. The zero-shot and few-shot prompts give only a generic correction instruction ("fix all grammatical and spelling errors"), which provides no guidance on correction scope. In practice, this leads LLMs to overcorrect: rephrasing sentences, substituting synonyms, or "improving" stylistically acceptable constructions. Since the ERRANT-based $F_{0.5}$ metric penalizes unnecessary edits, such overcorrection directly hurts precision.

The minimal-edits prompt addresses this with a two-part structure (Appendix A.3). The first part explicitly declares the minimal-edit constraint: "correct only clear-cut errors while preserving the original wording". The second part provides a

specific taxonomy of 16 Ukrainian GEC error categories (spelling, punctuation, case, gender, number, aspect, tense, etc.), followed by language-specific conventions (e.g., у/в 'u/v' alternation before consonants/vowels, em-dash in dialogue) and strict rules on what not to change (no synonym substitution, no quote style normalization, no changes when in doubt):

> ...
> Виправляй ЛИШЕ такі типи помилок:
> *('Fix ONLY the following types of errors:')*
> 1. Орфографія: явні орфографічні помилки
> *('1. Spelling: obvious spelling errors')*
> 2. Пунктуація: пропущені або зайві коми, крапки...
> *('2. Punctuation: missing or extra commas, periods...')*
> 3. G/Case: некоректне вживання відмінкової форми
> *('3. G/Case: incorrect use of case form')*
> ...

The minimal-edits + few-shot variant (Appendix A.4) combines this detailed system prompt with few-shot examples.

LLM-assisted prompt optimization. Inspired by automatic prompt optimization methods (Ramnath et al., 2025), we develop a semi-automatic approach in which an LLM proposes prompt edits but a human reviews and accepts them. Our method borrows ideas from several automatic prompt optimization papers: like ProTeGi (Pryzant et al., 2023), we use LLM-generated "textual gradients" derived from error analysis to guide prompt edits; following PromptAgent (Wang et al., 2024), we cluster prediction–reference mismatches into recurring linguistic patterns (e.g., "unnecessary dash normalization", "missed comma before subordinate conjunction") to produce domain-expert-style prompt sections; and as in OPRO (Yang et al., 2024), we maintain an optimization history of previous candidates and their scores to inform each iteration.

LLM-assisted prompt optimization design. We implemented this pipeline as a Claude Code skill powered by Claude Opus 4.6, which acts as both error analyst and prompt engineer. Starting from the best manual prompt (usually minimal-edits + few-shot), the agent iteratively: (1) evaluates the candidate on the validation set, recording span-level TP/FP/FN; (2) clusters mismatches into linguistic patterns ranked by frequency; (3) modifies the prompt via rule insertion (an explicit prohibition in the "do not change" section) or example insertion (a targeted input–output pair, including "no-change" examples); (4) accepts the change only if $F_{0.5}$ improves, otherwise reverts. The cycle repeats until gains plateau.

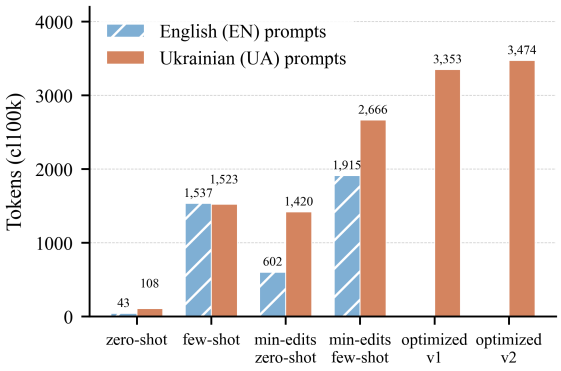


Figure 1: Prompt length (in tokens, cl100k_base tokenizer[6]) across prompting strategies for English (EN) and Ukrainian (UA) prompts. Optimized variants are available in UA only. The minimal-edits + few-shot + optimized-v2 prompt (A.5.2) is ~32× longer than the zero-shot UA baseline (A.1.2; 3,474 vs. 108 tokens).

Optimization setup. Due to cost and time constraints, we could not run the optimization loop separately for every model. Instead, we selected the best-performing manual prompt, minimal-edits + few-shot (A.4), and optimized it in two rounds: first on GPT-4.1-mini, producing minimal-edits + few-shot + optimized-v1 (A.5.1), and then starting from that result on Gemini 3-Flash, producing minimal-edits + few-shot + optimized-v2 (A.5.2). We then transferred these prompts to the remaining models without further tuning: GPT and Claude models are evaluated with optimized-v1, Gemini models with optimized-v2 (Table 3). We acknowledge that per-model optimization would give a more complete picture; we report these preliminary results as a useful reference point.

Prompt length. Figure 1 shows how prompt length grows across strategies, from 43 tokens for zero-shot (EN, A.1.1) to 3,474 tokens for minimal-edits + few-shot + optimized-v2 (UA, A.5.2).

4 Experimental Results

RQ1: Zero-shot performance and prompt language (Table 1). We start with zero-shot prompts, the simplest and most widely used setup for LLM-based GEC, and test whether prompting in Ukrainian rather than English improves results.

[6] https://github.com/openai/tiktoken

Model	Lang.	Prec.	Rec.	$F_{0.5}$	UA?
Baseline: fine-tuned					
mBART50-large†	–	78.52	50.60	70.71	
mT5-large‡	–	**76.81**	**61.39**	**73.14**	
Baseline: LLM zero-shot					
GPT-3.5*	EN	25.80	36.20	27.40	
API-accessed (zero-shot)					
GPT-4.1	EN	37.17	55.54	39.80	
GPT-4.1	UA	30.24	58.41	33.47	
GPT-4.1-mini	EN	39.06	51.92	41.09	
GPT-4.1-mini	UA	36.73	53.03	39.13	
GPT-5.1 (medium)	EN	36.06	60.15	39.20	
GPT-5.1 (medium)	UA	32.35	62.61	35.82	
GPT-5.2 (medium)	EN	34.04	66.31	37.71	
GPT-5.2 (medium)	UA	29.89	66.90	33.60	
GPT-5.4 (medium)	EN	36.87	62.96	40.20	
GPT-5.4 (medium)	UA	32.73	65.35	36.36	
Claude Sonnet 4.6	EN	41.79	45.83	42.54	
Claude Sonnet 4.6	UA	42.70	47.76	43.63	+
Claude Opus 4.6	EN	47.60	46.20	47.30	
Claude Opus 4.6	UA	**49.20**	51.60	**49.70**	+
Gemini 3-Flash	EN	39.09	64.85	42.46	
Gemini 3-Flash	UA	35.91	**67.38**	39.61	
Gemini 3-Pro	EN	37.89	60.54	40.96	
Gemini 3-Pro	UA	36.30	63.58	39.71	
Gemini 3.1-Pro	EN	41.50	58.03	44.01	
Gemini 3.1-Pro	UA	39.16	62.01	42.28	
Kimi-K2	EN	34.24	52.74	36.82	
Kimi-K2	UA	29.03	60.11	32.38	
Open-source (zero-shot)					
Lapa v0.1.2§	EN	24.24	23.52	24.09	
Lapa v0.1.2§	UA	28.57	32.38	29.26	+

Table 1: RQ1: Zero-shot GEC performance on UNLP-2023-test (GEC only). Lang.: prompt language, Ukrainian (UA) or English (EN). Bold marks the best result per column among API-accessed models. UA?: + indicates that the UA prompt yields a higher $F_{0.5}$ than the EN prompt for the same model. †Syvokon and Romanyshyn (2023); ‡Gomez et al. (2023); *Katinskaia and Yangarber (2024), EN; §Paniv et al. (2025), Ukrainian open-source LLM.

Zero-shot performance. All current models dramatically outperform the GPT-3.5 baseline of $F_{0.5}$ = 27.4 reported by Katinskaia and Yangarber (2024). The best zero-shot system, Claude Opus 4.6 with a UA prompt, reaches $F_{0.5}$ = 49.70, nearly doubling the GPT-3.5 score. Notably, GPT-5.x reasoning models do not outperform the older GPT-4.1-mini ($F_{0.5}$ = 41.09), despite their stronger general benchmarks. We attribute this to over-correction: reasoning-optimized models tend to over-interpret the correction task, producing more extensive rewrites that ERRANT penalizes as false positives.

Even the best zero-shot result remains 23.4 $F_{0.5}$ points below the fine-tuned SOTA of 73.14. Comparing the best zero-shot system (Claude Opus 4.6 UA: P=49.20, R=51.60) against the fine-tuned reference (P=76.81, R=61.39), the gap is primarily driven by precision (27.6-point difference) rather than recall (9.8-point difference). Without task-specific fine-tuning, zero-shot models lack the calibration to suppress spurious corrections.

Models differ in their precision–recall profiles. Gemini variants are high-recall correctors (R=60–67%) with low precision (P=36–42%), flagging many candidates, many of which are spurious. Claude models show a more balanced profile with precision and recall within 5 points of each other, which is favorable under $F_{0.5}$'s precision weighting. GPT-5.x models lean toward high recall and low precision, similar to Gemini.

Prompt language. Surprisingly, Claude is the only family where UA prompts improve performance: Claude Opus 4.6 gains 2.4 points (47.30 → 49.70) and Claude Sonnet 4.6 gains 1.1 points (42.54 → 43.63), with both precision and recall improving simultaneously. For all other models, UA prompts degrade $F_{0.5}$ by 1–6 points, consistently trading precision for recall and amplifying overcorrection. For reference, we also include Lapa v0.1.2[7], an open-source Ukrainian LLM, which scores $F_{0.5}$ = 29.26 with a UA prompt, above GPT-3.5 but well below the API-accessed models.

RQ2: Prompting strategies (Table 2). We select the six best-performing models from RQ1 (one to two per provider, excluding lower-scoring variants) and test whether few-shot examples and minimal-edits constraints can reduce the overcorrection observed in RQ1.

Best results and gap to SOTA. Combining few-shot examples with minimal-edits instructions, minimal-edits + few-shot yields the best or near-best $F_{0.5}$ for every model. Claude Opus 4.6 ($F_{0.5}$ = 63.73) and Gemini 3.1-Pro (63.68) effectively tie despite different zero-shot starting points. GPT-5.4 shows the largest absolute gain (+19.5

[7]Lapa (Paniv et al., 2025) is an open-source Ukrainian LLM fine-tuned with GEC-style prompts different from ours, and the only non-API-accessed model in our evaluation. We include it as a reference point for open-weight Ukrainian-centric models, though a full comparison with open-source alternatives is beyond the scope of this work.

Model	Prompting Strategy	Lang.	Prec.	Rec.	$F_{0.5}$
Baseline: fine-tuned SOTA					
mT5-large‡	–	–	76.81	61.39	73.14
GPT-4.1-mini	zero-shot (A.1.1)	EN	39.06	51.92	41.09
	few-shot (A.2.1)	EN	44.75	59.73	47.11
	minimal-edits + zero-shot (A.3.1)	UA	46.66	51.52	47.56
	minimal-edits + few-shot (A.4.1)	UA	47.16	51.48	47.97
GPT-5.4	zero-shot (A.1.1)	EN	36.87	62.96	40.20
	few-shot (A.2.1)	EN	46.24	66.67	49.26
	minimal-edits + zero-shot (A.3.1)	UA	57.05	63.18	58.18
	minimal-edits + few-shot (A.4.1)	UA	60.02	58.40	59.69
Claude Sonnet 4.6	zero-shot (A.1.1)	EN	41.79	45.83	42.54
	few-shot (A.2.1)	EN	52.52	56.26	53.23
	minimal-edits + zero-shot (A.3.1)	UA	62.44	48.55	59.06
	minimal-edits + few-shot (A.4.1)	UA	61.96	49.10	58.88
Claude Opus 4.6	zero-shot (A.1.1)	EN	47.60	46.20	47.30
	few-shot (A.2.1)	EN	56.83	55.93	56.65
	minimal-edits + zero-shot (A.3.1)	UA	67.63	47.08	62.20
	minimal-edits + few-shot (A.4.1)	UA	**68.54**	49.75	63.73
Gemini 3-Flash	zero-shot (A.1.1)	EN	39.09	64.85	42.46
	few-shot (A.2.1)	EN	48.48	**72.01**	51.87
	minimal-edits + zero-shot (A.3.1)	UA	53.29	66.40	55.48
	minimal-edits + few-shot (A.4.1)	UA	60.28	66.18	61.38
Gemini 3.1-Pro	zero-shot (A.1.1)	EN	41.50	58.03	44.01
	few-shot (A.2.1)	EN	54.92	66.25	56.86
	minimal-edits + zero-shot (A.3.1)	UA	60.49	65.24	61.38
	minimal-edits + few-shot (A.4.1)	UA	63.76	63.35	**63.68**

Table 2: RQ2: Effect of prompting strategies on UNLP-2023-test (GEC only). Zero-shot and few-shot use EN prompts; minimal-edits variants use UA prompts (see Section 3 for rationale). Lang.: prompt language, Ukrainian (UA) or English (EN). For each model, we report the best configuration per strategy. Bold marks the best result per column among API-accessed models. ‡Gomez et al. (2023).

points), recovering from the weakest zero-shot result to a competitive 59.69.

However, even the best prompted result falls 9.4 points below the fine-tuned SOTA of 73.14, with the gap concentrated in precision (P=76.81 vs. 68.54). The minimal-edits instruction suppresses the most egregious false positives, but a long tail of borderline corrections remains that likely requires task-specific fine-tuning.

Few-shot gains. Adding few-shot examples to the zero-shot prompt produces moderate but reliable gains of +6–13 $F_{0.5}$ points, driven by improvements in both precision and recall. The gains are largest for Gemini 3.1-Pro (+12.85) and smallest for GPT-4.1-mini (+6.02).

Minimal-edits instructions matter most. The minimal-edits constraint has a larger effect than few-shot examples. Switching from a generic EN prompt to a UA minimal-edits instruction, even without few-shot examples, already matches or exceeds few-shot-only performance for five out of six models. The most striking case is GPT-5.4: minimal-edits + zero-shot alone yields $F_{0.5}$ = 58.18, a full 9 points above its few-shot score of 49.26. The mechanism is a sharp precision increase (+8–21 points across models) with modest recall change, meaning the constraint reduces unnecessary edits without hurting the model's ability to catch real errors.

Overall trends. Across all six models, $F_{0.5}$ improves consistently along the progression: zero-shot < few-shot < minimal-edits + zero-shot ≤ minimal-edits + few-shot, with one notable exception: Claude Sonnet 4.6 peaks at minimal-edits + zero-shot (59.06) and slightly drops with the addition of few-shot examples (58.88). As discussed in Section 3, the minimal-edits prompts are written in Ukrainian by design, so we cannot fully separate the effect of prompt language from the effect of the prompting strategy itself.

RQ3: LLM-assisted prompt optimization (Table 3).

Model	Prompting Strategy	Lang.	Prec.	Rec.	$F_{0.5}$
Baseline: fine-tuned SOTA					
mT5-large‡	–	–	76.81	61.39	73.14
GPT-4.1-mini	minimal-edits + few-shot (A.4.1)	UA	47.16	51.48	47.97
	minimal-edits + few-shot + optimized-v1 (A.5.1)	UA	55.75	51.49	54.84
GPT-5.4	minimal-edits + few-shot (A.4.1)	UA	60.02	58.40	59.69
	minimal-edits + few-shot + optimized-v1 (A.5.1)	UA	63.94	51.75	61.07
Claude Sonnet 4.6	minimal-edits + zero-shot[8] (A.3.1)	UA	62.44	48.55	59.06
	minimal-edits + few-shot + optimized-v1 (A.5.1)	UA	66.73	36.58	57.29
Claude Opus 4.6	minimal-edits + few-shot (A.4.1)	UA	68.54	49.75	63.73
	minimal-edits + few-shot + optimized-v1 (A.5.1)	UA	66.54	38.27	57.98
Gemini 3-Flash	minimal-edits + few-shot (A.4.1)	UA	60.28	**66.18**	61.38
	minimal-edits + few-shot + optimized-v2 (A.5.2)	UA	69.98	62.87	68.43
Gemini 3.1-Pro	minimal-edits + few-shot (A.4.1)	UA	63.76	63.35	63.68
	minimal-edits + few-shot + optimized-v2 (A.5.2)	UA	**70.77**	63.63	**69.22**

Table 3: RQ3: Effect of LLM-assisted prompt optimization, evaluated on UNLP-2023-test (GEC only). Optimized-v1 was tuned on GPT-4.1-mini, optimized-v2 on Gemini 3-Flash; both were then transferred to other models. Lang.: prompt language, Ukrainian (UA) or English (EN). For each model, the first row shows the best manual prompt result (from Table 2); the second row shows the best optimized result. Bold marks the best result per column among API-accessed models. ‡Gomez et al. (2023).

Best results and gap to SOTA. The best optimized result is Gemini 3.1-Pro with minimal-edits + few-shot + optimized-v2 (A.5.2; $F_{0.5}$ = 69.22), followed closely by Gemini 3-Flash with the same prompt (68.43). This narrows the gap to fine-tuned SOTA from 9.5 to 3.9 points. The gain on the target model is precision-driven: precision rises from 63.76 to 70.77 (+7.0) while recall remains nearly unchanged. The remaining 3.9-point gap is concentrated in precision (70.77 vs. 76.81), suggesting that closing it likely requires more focused instructions.

Improvement within Gemini. Since the minimal-edits + few-shot + optimized-v2 prompt (A.5.2) was tuned directly on Gemini 3-Flash, its strong gain on that model ($F_{0.5}$: 61.38 → 68.43, +7.05) is expected. More notably, the same prompt transfers successfully to Gemini 3.1-Pro, which achieves the overall best result ($F_{0.5}$ = 69.22), indicating that optimization on a smaller model within the family can benefit larger variants.

Improvement on GPT. GPT models show mixed results. GPT-4.1-mini gains +6.87 points (47.97 → 54.84), a substantial improvement but still the weakest absolute result. GPT-5.4 gains only +1.38 (59.69 → 61.07), suggesting that the stronger model already captures most correction patterns encoded in the optimized prompt.

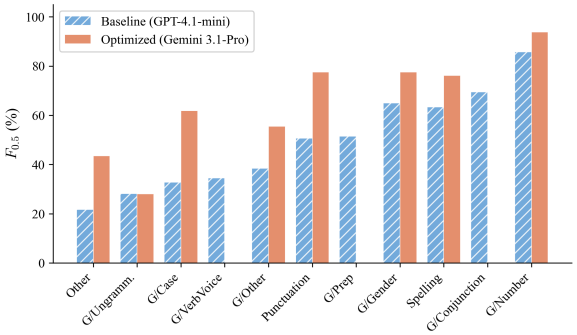


Figure 2: RQ4: Per-error-type $F_{0.5}$ on UNLP-2023-test (GEC only). Baseline: GPT-4.1-mini (zero-shot (A.1.1), EN); Optimized: Gemini 3.1-Pro (minimal-edits + few-shot + optimized-v2 (A.5.2), UA). Error types sorted as in Table 4 (by baseline overcorrection ratio, descending).

No improvement on Claude. The optimized prompt degrades both Claude models: Opus drops by −5.75 points (63.73 → 57.98) and Sonnet by −1.77 points (59.06 → 57.29), driven largely by a recall collapse (Opus: 49.75 → 38.27; Sonnet: 48.55 → 36.58). Claude appears to interpret the optimized rules too conservatively, suppressing genuine corrections alongside false positives. This finding shows that optimization on one model family does not necessarily guarantee improvement on another.

RQ4: Where do minimal-edits instructions help and where do they fail? (Figure 2, Table 4). We compare per-error-type $F_{0.5}$ between a standard

Error Type	GPT-4.1-mini	Gemini 3.1-Pro
Other	3.67	0.50
G/Ungramm.	2.50	2.00
G/Case	1.88	0.39
G/VerbVoice	1.50	–
G/Other	1.00	0.00
Punctuation	0.95	0.26
G/Prep	0.86	–
G/Gender	0.60	0.22
Spelling	0.59	0.25
G/Conjunction	0.40	∞
G/Number	0.17	0.00

Table 4: RQ4: Overcorrection ratio (FP/TP) per error type; lower is better. The two columns compare the weakest (GPT-4.1-mini, zero-shot (A.1.1), EN) and strongest (Gemini 3.1-Pro, minimal-edits + few-shot + optimized-v2 (A.5.2), UA) configurations from Tables 2–3; note that both model and prompt differ. Rows sorted by GPT-4.1-mini ratio (descending); overcorrection (more false positives than true positives) occurs above 1.0. "–" indicates no predictions for that type; ∞ indicates only false positives.

zero-shot prompt (GPT-4.1-mini) and our best optimized prompt (Gemini 3.1-Pro, minimal-edits + few-shot + optimized-v2; A.5.2) to identify which error categories benefit most from detailed minimal-edits instructions. Note that this comparison reflects the combined effect of model choice, prompt strategy, and prompt language; we select these two configurations as the weakest and strongest endpoints of our evaluation pipeline.

Where minimal-edit instructions help. The largest $F_{0.5}$ gains appear in categories amenable to explicit rules. Punctuation improves from 50.67 to 77.56 (+26.9), G/Case from 32.82 to 61.83 (+29.0), and G/Gender from 64.94 to 77.59 (+12.7). The overcorrection ratios in Table 4 confirm the mechanism: FP/TP drops from 0.95 to 0.26 for Punctuation, from 1.88 to 0.39 for G/Case, and from 0.60 to 0.22 for G/Gender. Spelling and G/Number are reliable under both configurations.

Where minimal-edit instructions fail. Three categories drop to $F_{0.5} = 0$ under the optimized prompt: G/Prep, G/VerbVoice, and G/Conjunction. The baseline achieves non-trivial $F_{0.5}$ scores on these types (51.47, 34.48, and 69.44), but the detailed minimal-edits rules cause the model to avoid these corrections entirely. G/UngrammaticalStructure remains persistently overcorrected in both settings ($F_{0.5}$: 28.17 → 28.04), indicating a structural difficulty that instructions cannot resolve.

Ukrainian-specific overcorrection patterns. Error analysis reveals five recurring patterns specific to Ukrainian, driven by the interaction between English-calibrated correction heuristics and Ukrainian linguistic norms. Below, we report false positive counts out of 1,274 test sentences.

En-dash over-normalization (– → —; ~49 FP, 3.8% of sentences). Em-dashes are obligatory in direct speech but not elsewhere; the model generalizes the rule indiscriminately.

Dialogue reformatting (~40 FP, 3.1% of sentences). The model converts acceptable quote-style dialogue («текст», — сказав ' "text," — said') to dash-style, applying a real Ukrainian norm where none was required. A single prohibition rule was sufficient to suppress this pattern.

Synonym and register substitution (~30 FP, 2.4% of sentences). Acceptable words are replaced with literary alternatives (знаходиться 'is located' → перебуває 'is situated'), violating the minimal-edit constraint.

Euphonic preposition alternation (в/у 'v/u', з/із/зі 'z/iz/zi'; ~17 FP, 1.3% of sentences). Ukrainian preposition choice is phonetically conditioned; the model both over- and under-corrects within the same category.

Collapse of morphological variants. Ukrainian admits multiple grammatically correct surface forms (навчались/навчалися 'studied', їх/їхній 'their'); the model collapses these to a single preferred form. This space of acceptable alternations is open-ended and cannot be exhaustively covered by prompt examples.

These patterns share a common cause: the model's correction heuristics are calibrated to English, where most of these alternations do not exist. Overall, our strongest prompt (Gemini 3.1-Pro, minimal-edits + few-shot + optimized-v2 (A.5.2), UA) significantly reduces overcorrection for high-frequency, rule-based categories (Table 4), but at the cost of the model becoming too conservative on low-frequency grammatical categories.

5 Conclusion

We presented the first systematic evaluation of prompting strategies for minimal-edit Ukrainian GEC using API-accessed LLMs. While fine-tuned models currently dominate GEC benchmarks, we show that prompting alone can be competitive. On the UNLP-2023-test (GEC only), our best configuration (Gemini 3.1-Pro with LLM-assisted opti-

mization) reaches $F_{0.5} = 69.22$, closing over 90% of the gap between the previous API-accessed result of Katinskaia and Yangarber (2024) (GPT-3.5, $F_{0.5} = 27.4$) and the fine-tuned SOTA of Gomez et al. (2023) (mT5-large, $F_{0.5} = 73.14$).

Our findings yield four takeaways. First, for zero-shot and few-shot prompts, English is sufficient for most models; only Claude benefits from Ukrainian prompts (RQ1). Our best overall results, however, use Ukrainian minimal-edits prompts, as the language-specific rules they encode require Ukrainian to express precisely. Second, the minimal-edits strategy provides the largest gains, outperforming both zero-shot and few-shot baselines across all models (RQ2). Third, LLM-assisted prompt optimization yields further improvements on the model family it was optimized for, but does not transfer reliably across families (RQ3). Fourth, minimal-edits instructions yield the largest per-category gains for punctuation and case errors, but cause the model to abandon several low-frequency grammatical categories entirely, revealing a precision-recall tradeoff inherent to detailed prompting (RQ4).

Limitations

Our study has several limitations:

1. We evaluate on a single benchmark (UNLP-2023-test (GEC only)); results may not generalize to other Ukrainian GEC datasets or domains.

2. Although we include a single open-source model (Lapa v0.1.2) as a reference point, we do not systematically compare against open-weight models (e.g., Llama 3, Mistral, Lapa, MamayLM) that could be prompted or fine-tuned without API costs, leaving this as future work.

3. API-accessed models are opaque and subject to unannounced updates, making exact reproducibility difficult.

4. Although the UNLP-2023-test (GEC only) gold annotations are held out, the upstream UA-GEC train and valid splits are publicly available. Since UA-GEC train is the source of our few-shot exemplars, it is plausible that commercial LLMs encountered similar sentences and annotation patterns during pretraining, which could inflate recall on a corpus sharing the same annotation conventions. A cleaner control would draw exemplars from an independent Ukrainian GEC corpus, but no comparable dataset currently exists, so we treat our results as establishing initial prompting baselines; prior API-accessed results for Ukrainian GEC at this scale are essentially absent.

5. We run each configuration only once; since LLM outputs are not fully deterministic, reproduced scores may differ/cos slightly.

6. Our LLM-assisted prompt optimization pipeline optimizes $F_{0.5}$ on the validation set, which may overfit to its error distribution.

7. All our experiments are evaluated with span-based $F_{0.5}$ computed by ERRANT, the official metric of the UNLP 2023 Shared Task, GEC-only track (Syvokon and Romanyshyn, 2023); this differs from GLEU used in MultiGEC-2025 (Masciolini et al., 2025), so scores are not directly comparable across benchmarks.

8. Optimized prompts are substantially longer (roughly 32× the zero-shot baseline; see Figure 1), which may increase token cost and latency. Prompt caching, now widely supported by providers, amortizes much of this overhead, making the net cost hard to estimate.

Ethical Considerations

In accordance with the conference policy on AI-based writing assistance, we disclose that ChatGPT, Claude, Gemini, and Grammarly were used for drafting, editing, and proofreading. All AI-generated text was reviewed by the authors, who take full responsibility for the final content.

Acknowledgments

We are deeply grateful to YouScan for fostering an inspiring environment that encourages both research and professional development. We also express our appreciation to the Faculty of Applied Sciences at the Ukrainian Catholic University for supporting this work as part of an M.Sc. thesis program. We gratefully acknowledge Mariana Romanyshyn and Oleksiy Syvokon for their assistance with the UNLP 2023 Shared Task. Finally, we extend our sincere gratitude to the anonymous reviewers for their insightful feedback and dedicated efforts in refining this manuscript.

References

Maksym Bondarenko, Artem Yushko, and Andrii Shportko. 2023. Comparative study of models trained on synthetic data for Ukrainian grammatical error correction. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 103–113, Dubrovnik, Croatia. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, 49(3):643–701.

Frank Palma Gomez, Alla Rozovskaya, and Dan Roth. 2023. A low-resource approach to the grammatical error correction of Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 114–119, Dubrovnik, Croatia. Association for Computational Linguistics.

Anisia Katinskaia and Roman Yangarber. 2024. GPT-3.5 for grammatical error correction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7831–7843, Torino, Italia. ELRA and ICCL.

Roman Kovalchuk, Mariana Romanyshyn, and Petro Ivaniuk. 2025. Introducing OmniGEC: A silver multilingual dataset for grammatical error correction. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP)*, pages 162–178, Vienna, Austria. Association for Computational Linguistics.

Mengsay Loem, Masahiro Kaneko, Sho Takase, and Naoaki Okazaki. 2023. Exploring effectiveness of GPT-3 in grammatical error correction: A study on performance and controllability in prompt-based methods. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 205–219, Toronto, Canada. Association for Computational Linguistics.

Agnes Luhtaru, Taido Purason, Martin Vainikko, and Maali Helin Del. 2024. To err is human, but llamas can learn it too. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP)*, Torino, Italia. Association for Computational Linguistics.

Arianna Masciolini, Andrew Caines, Orphée De Clercq, Joni Kruijsbergen, Murathan Kurfalı, Ricardo Muñoz Sánchez, Elena Volodina, and Robert Östling. 2025. The MultiGEC-2025 shared task on multilingual grammatical error correction at NLP4CALL. In *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2025)*, pages 1–33, Tallinn, Estonia. University of Tartu Library.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 229–234, Valencia, Spain. Association for Computational Linguistics.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. Association for Computational Linguistics.

Yurii Paniv, Bohdan Didenko, Mykola Haltiuk, Vladyslav Humennyy, Andrian Kravchenko, Roman Kyslyi, Viktoriia Makovska, Artem Orlovskyi, Bohdan Ruban, Maksym-Yurii Rudko, Anastasiia Senyk, Nazarii Drushchak, Dmytro Chaplynskyi, and Mariana Romanyshyn. 2025. Lapa LLM v0.1.2 — the most efficient Ukrainian open-source language model.

Reid Pryzant, Dan Iter, Jerry Li, Yin Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with "gradient descent" and beam search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7957–7968, Singapore. Association for Computational Linguistics.

Kiran Ramnath, Kang Zhou, Sheng Guan, Soumya Smruti Mishra, Xuan Qi, Zhengyuan Shen, Shuai Wang, Sangmin Woo, Sullam Jeoung, Yawei Wang, Haozhu Wang, Han Ding, Yuzhe Lu, Zhichao Xu, Yun Zhou, Balasubramaniam Srinivasan, Qiaojing Yan, Yueyan Chen, Haibo Ding, and 2 others. 2025. A systematic survey of automatic prompt optimization techniques. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33078–33110, Suzhou, China. Association for Computational Linguistics.

Aman Saini, Artem Chernodub, Vipul Raheja, and Vivek Kulkarni. 2024. Spivavtor: An instruction tuned Ukrainian text editing model. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP)*, pages 95–108, Torino, Italia. ELRA and ICCL.

Ryszard Staruch, Filip Graliński, and Daniel Dzienisiewicz. 2025. Adapting LLMs for minimal-edit grammatical error correction. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 118–128, Vienna, Austria. Association for Computational Linguistics.

Oleksiy Syvokon and Mariana Romanyshyn. 2023. The UNLP 2023 shared task on grammatical error correction for Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 120–131, Dubrovnik, Croatia. Association for Computational Linguistics.

Xinyuan Wang, Chenxi Li, Zhen Wang, Fan Bai, Haotian Luo, Jiayou Zhang, Nebojsa Jojic, Eric P. Xing, and Zhiting Hu. 2024. PromptAgent: Strategic planning with language models enables expert-level prompt optimization. In *The Twelfth International Conference on Learning Representations*, Vienna, Austria. OpenReview.net.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2024. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*.

A Prompts

Below we list all prompts in both English (EN) and Ukrainian (UA) versions. The placeholder `{sentence}` is replaced with the input sentence at inference time.

A.1 Zero-shot prompts

The baseline prompt provides only a task description with no examples.

A.1.1 Zero-shot prompt (EN)

```
Reply with a corrected version of the sentence with all grammatical and spelling errors fixed.
If there are no errors, reply with a copy of the original sentence.
Input sentence: <input_text>
Corrected sentence:
```

A.1.2 Zero-shot prompt (UA)

Надай виправлену версію речення з виправленими всіма граматичними та орфографічними помилками.
Якщо помилок немає, надай копію оригінального речення.
Вхідне речення: <input_text>
Виправлене речення:

(English: 'Provide a corrected version of the sentence with all grammatical and spelling errors fixed. If there are no errors, provide a copy of the original sentence. Input sentence: <input_text>. Corrected sentence:')

A.2 Few-shot prompts

The few-shot prompt prepends source–target pairs selected from the training set to cover spelling, punctuation, and morphological error types.

A.2.1 Few-shot prompt (EN)

```
[Same header as Prompt 1]

Examples:

Input: Так само потерпає Україна і сьогодні від того що насправді талановитим людям заважають працювати усілякі посередності "у руля".
Output: Так само потерпає Україна і сьогодні від того, що насправді талановитим людям заважають працювати усілякі посередності "у руля".

Input: Це пов'язано з тим, що такі колективні рухи молекул води сильно збільшують характерні часи процесів які відбуваються в системі.
Output: Це пов'язано з тим, що такі колективні рухи молекул води сильно збільшують характерні часи процесів, які відбуваються в системі.

Input: Я ніколи не навчався у медичному коледжі, кажу, - Я лиш підійшов як звичайна людина подивитись чи можу бути чимось корисний.
Output: Я ніколи не навчався у медичному коледжі, кажу, - Я лиш підійшов як звичайна людина подивитись, чи можу бути чимось корисний.

Input: Це у місті швидка приїжджає, забирає хворого і везе у лікарню; якщо ж до лікарні кілька годин льоту то хворого можна і не довезти.
Output: Це у місті швидка приїжджає, забирає хворого і везе у лікарню; якщо ж до лікарні кілька годин льоту, то хворого можна і не довезти.

Input: Найбільше він любив тримати в руках старанно орнаментовані стародавні шолом і меч, котрі своїм золотом, здається гріли старого.
Output: Найбільше він любив тримати в руках старанно орнаментовані стародавні шолом і меч, котрі своїм золотом, здається, гріли старого.

Input: - Я часто казав тобі, що ти дурненька, - сказав він.
Output: — Я часто казав тобі, що ти дурненька, — сказав він.

Input: Така традиція також походить з Візантії, прикладом є зображення Андроніка II Палеолога.
Output: Така традиція також походить із Візантії, прикладом є зображення Андроніка II Палеолога.

Input: Як і більшість ділових людей, він не знав напамять жодного вірша і не памятав жодної казки, а тому щоразу мусив імпровізувати.
Output: Як і більшість ділових людей, він не знав напам'ять жодного вірша і не памятав жодної казки, а тому щоразу мусив імпровізувати.

Input: Настя, привіт! я хотіла уточнити про завдання Олі.
Output: Насте, привіт! Я хотіла уточнити про завдання Олі.

Input: Я принесу твої улюблені солодощі та обніму тебе міцно-міцно.
Output: Я принесу твої улюблені солодощі та обніму тебе міцно-міцно.

Input: Смакота ще та, скажу я вам))
Output: Смакота ще та, скажу я вам))

Input: У той час глибокий сенс народної мудрості нам був ще недоступним через брак досвіду.
Output: У той час глибокий сенс народної мудрості нам був ще недоступним через брак досвіду.

Input: Каналізація в будинках еволюціонує повільно, але все ж таки змінюється.
Output: Каналізація в будинках еволюціонує повільно, але все ж таки змінюється.

Input sentence: {input_text}
Corrected sentence:
```

A.2.2 Few-shot prompt (UA)

[Той самий заголовок, що і в Prompt 1]
(English: '[Same header as Prompt 1]')

Приклади: *(‘Examples:’)*

Вхід: Так само потерпає Україна і сьогодні від того що насправді талановитим людям заважають працювати усілякі посередності "у руля".
Вихід: Так само потерпає Україна і сьогодні від того, що насправді талановитим людям заважають працювати усілякі посередності "у руля".

Вхід: Я ніколи не навчався у медичному коледжі, кажу, - Я лиш підійшов як звичайна людина подивитись чи можу бути чимось корисний.
Вихід: Я ніколи не навчався у медичному коледжі, кажу, - Я лиш підійшов як звичайна людина подивитись, чи можу бути чимось корисний.

Вхід: Це у місті швидка приїжджає, забирає хворого і везе у лікарню; якщо ж до лікарні кілька годин льоту то хворого можна і не довезти.
Вихід: Це у місті швидка приїжджає, забирає хворого і везе у лікарню; якщо ж до лікарні кілька годин льоту, то хворого можна і не довезти.

Вхід: Найбільше він любив тримати в руках старанно орнаментовані стародавні шолом і меч, котрі своїм золотом, здається гріли старого.
Вихід: Найбільше він любив тримати в руках старанно орнаментовані стародавні шолом і меч, котрі своїм золотом, здається, гріли старого.

Вхід: - Я часто казав тобі, що ти дурненька, - сказав він.
Вихід: — Я часто казав тобі, що ти дурненька, — сказав він.

Вхід: Така традиція також походить з Візантії, прикладом є зображення Андроніка II Палеолога.
Вихід: Така традиція також походить із Візантії, прикладом є зображення Андроніка II Палеолога.

Вхід: Як і більшість ділових людей, він не знав напамять жодного вірша і не памятав жодної казки, а тому щоразу мусив імпровізувати.
Вихід: Як і більшість ділових людей, він не знав напам’ять жодного вірша і не памятав жодної казки, а тому щоразу мусив імпровізувати.

Вхід: Настя, привіт! я хотіла уточнити про завдання Олі.
Вихід: Насте, привіт! Я хотіла уточнити про завдання Олі.

Вхід: Я принесу твої улюблені солодощі та обніму тебе міцно-міцно.
Вихід: Я принесу твої улюблені солодощі та обніму тебе міцно-міцно.

Вхід: Смакота ще та, скажу я вам))
Вихід: Смакота ще та, скажу я вам))

Вхід: У той час глибокий сенс народної мудрості нам був ще недоступним через брак досвіду.
Вихід: У той час глибокий сенс народної мудрості нам був ще недоступним через брак досвіду.

Вхід: Каналізація в будинках еволюціонує повільно, але все ж таки змінюється.
Вихід: Каналізація в будинках еволюціонує повільно, але все ж таки змінюється.

Вхідне речення: {input_text}
Виправлене речення:

(English: Вхід/Вихід = *‘Input’/‘Output’;* Вхідне речення = *‘Input sentence’;* Виправлене речення = *‘Corrected sentence’. The few-shot examples are the same Ukrainian GEC sentence pairs as in the EN variant (A.2.1), with Ukrainian keywords.)*

A.3 Minimal-edits zero-shot prompts

This prompt replaces the generic system prompt with a detailed minimal-edit instruction containing Ukrainian-specific grammar rules. No few-shot examples are included.

A.3.1 Minimal-edits zero-shot prompt (UA)

Ти — система виправлення українських граматичних помилок. Внось МІНІМАЛЬНІ зміни, щоб виправити ЛИШЕ явні граматичні, орфографічні та пунктуаційні помилки. НЕ переписуй, не перефразовуй і не замінюй слова синонімами. Точно зберігай оригінальне формулювання.

Виправляй ЛИШЕ такі типи помилок:

1. Орфографія: явні орфографічні помилки (друкарські помилки, неправильні літери).
2. Пунктуація: пропущені або зайві коми, крапки, знаки питання; використання тире (—) замість дефіса (-) у діалогах та вставних конструкціях.
3. G/Case: некоректне вживання відмінкової форми (зокрема кличний відмінок при звертаннях).
4. G/Gender: некоректне вживання форми роду.
5. G/Number: некоректне вживання форми числа.
6. G/Aspect: некоректне вживання форми виду дієслова.
7. G/Tense: некоректне вживання часової

форми дієслова.
8. G/VerbVoice: некоректне вживання форми стану дієслова.
9. G/PartVoice: некоректне вживання форми стану дієприкметника.
10. G/VerbAForm: некоректне вживання аналітичної форми дієслова.
11. G/Prep: некоректне вживання прийменника.
12. G/Participle: некоректне вживання дієприслівника.
13. G/UngrammaticalStructure: порушення граматичних норм у синтаксичних конструкціях.
14. G/Comparison: некоректна форма ступенів порівняння.
15. G/Conjunction: некоректне вживання сполучників.
16. G/Other: інші граматичні помилки.

ВАЖЛИВІ ПРАВИЛА УКРАЇНСЬКОЇ МОВИ:
- Прийменник «у» вживається перед приголосними (у школі, у місті, у готелі), «в» — перед голосними та на початку речення.
- Прийменник «об» вживається перед голосними (об одинадцятій), «о» — перед приголосними.
- У діалогах вживається тире (—), а не дефіс (-): «Текст», — сказав він. — Текст далі.
- Вставні слова (може, мабуть, звичайно, здається) виділяються комами з обох боків.
- Кличний відмінок при звертаннях: Настя → Насте, Олег → Олеже, мама → мамо.

СУВОРІ ПРАВИЛА:
- Виправляй ЛИШЕ явні помилки, перелічені вище
- Для кожної помилки вноси НАЙМЕНШУ можливу зміну
- НІКОЛИ не замінюй слова синонімами і не перефразовуй (залишай «буду йти», НЕ змінюй на «піду»)
- НІКОЛИ не змінюй слово на інше, якщо воно не написане з помилкою
- Зберігай оригінальний стиль лапок ("або «»), НЕ перетворюй один тип лапок на інший
- Зберігай оригінальне використання великих/малих літер, якщо це не явна помилка
- Якщо граматична форма є прийнятною, залишай її, навіть якщо можлива й інша форма
- Якщо є сумнів, НЕ змінюй
- Якщо помилок немає, повертай оригінальний текст БЕЗ ЗМІН
- Поверни ЛИШЕ виправлений текст

English translation of the above prompt:

You are a Ukrainian grammatical error correction system. Make MINIMAL changes to fix ONLY obvious grammatical, spelling, and punctuation errors. Do NOT rewrite, rephrase, or substitute synonyms. Preserve the original wording exactly.

Fix ONLY the following types of errors:
1. Spelling: obvious spelling errors (typos, wrong letters).
2. Punctuation: missing or extra commas, periods, question marks; use of em-dash (—) instead of hyphen (-) in dialogues and parenthetical constructions.
3. G/Case: incorrect case form (especially vocative case in forms of address).
4. G/Gender: incorrect gender form.
5. G/Number: incorrect number form.
6. G/Aspect: incorrect verb aspect form.
7. G/Tense: incorrect verb tense form.
8. G/VerbVoice: incorrect verb voice form.
9. G/PartVoice: incorrect participle voice form.
10. G/VerbAForm: incorrect analytical verb form.
11. G/Prep: incorrect preposition usage.
12. G/Participle: incorrect adverbial participle usage.
13. G/UngrammaticalStructure: grammatical norm violations in syntactic constructions.
14. G/Comparison: incorrect comparative/superlative form.
15. G/Conjunction: incorrect conjunction usage.
16. G/Other: other grammatical errors.

IMPORTANT RULES OF UKRAINIAN:
- Preposition “u” is used before consonants (u shkoli, u misti), “v” before vowels and at sentence start.
- Preposition “ob” is used before vowels (ob odynadtsiatii), “o” before consonants.
- Em-dash (—) is used in dialogues, not hyphen (-).
- Parenthetical words (maybe, probably, of course, it seems) are set off by commas on both sides.
- Vocative case in forms of address: Nastia → Naste, Oleh → Olezhe, mama → mamo.

STRICT RULES:
- Fix ONLY obvious errors listed above
- For each error, make the SMALLEST possible change
- NEVER substitute synonyms or rephrase
- NEVER change a word to another unless it is misspelled
- Preserve original quote style, do NOT convert one type to another
- Preserve original capitalization unless it is a clear error
- If a grammatical form is acceptable, leave it even if another form is possible
- If in doubt, do NOT change
- If there are no errors, return the original text WITHOUT CHANGES
- Return ONLY the corrected text

A.4 Minimal-edits few-shot prompts

This prompt combines the minimal-edits system prompt (Appendix A.3) with few-shot examples from the training set.

A.4.1 Minimal-edits few-shot prompt (UA)

Ти — система виправлення українських граматичних помилок. Внось МІНІМАЛЬНІ зміни, щоб виправити ЛИШЕ явні граматичні, орфографічні та пунктуаційні помилки. НЕ переписуй, не перефразовуй і не замінюй слова синонімами. Точно зберігай оригінальне

формулювання.

Виправляй ЛИШЕ такі типи помилок:

1. Орфографія: явні орфографічні помилки (друкарські помилки, неправильні літери).
2. Пунктуація: пропущені або зайві коми, крапки, знаки питання; використання тире (—) замість дефіса (-) у діалогах та вставних конструкціях.
3. G/Case: некоректне вживання відмінкової форми (зокрема кличний відмінок при звертаннях).
4. G/Gender: некоректне вживання форми роду.
5. G/Number: некоректне вживання форми числа.
6. G/Aspect: некоректне вживання форми виду дієслова.
7. G/Tense: некоректне вживання часової форми дієслова.
8. G/VerbVoice: некоректне вживання форми стану дієслова.
9. G/PartVoice: некоректне вживання форми стану дієприкметника.
10. G/VerbAForm: некоректне вживання аналітичної форми дієслова.
11. G/Prep: некоректне вживання прийменника.
12. G/Participle: некоректне вживання дієприслівника.
13. G/UngrammaticalStructure: порушення граматичних норм у синтаксичних конструкціях.
14. G/Comparison: некоректна форма ступенів порівняння.
15. G/Conjunction: некоректне вживання сполучників.
16. G/Other: інші граматичні помилки.

ВАЖЛИВІ ПРАВИЛА УКРАЇНСЬКОЇ МОВИ:
- Прийменник «у» вживається перед приголосними (у школі, у місті, у готелі), «в» — перед голосними та на початку речення.
- Прийменник «об» вживається перед голосними (об одинадцятій), «о» — перед приголосними.
- У діалогах вживається тире (—), а не дефіс (-): «Текст», — сказав він. — Текст далі.
- Вставні слова (може, мабуть, звичайно, здається) виділяються комами з обох боків.
- Кличний відмінок при звертаннях: Настя → Насте, Олег → Олеже, мама → мамо.

СУВОРІ ПРАВИЛА:
- Виправляй ЛИШЕ явні помилки, перелічені вище
- Для кожної помилки внось НАЙМЕНШУ можливу зміну
- НІКОЛИ не замінюй слова синонімами і не перефразовуй (залишай «буду йти», НЕ змінюй на «піду»)
- НІКОЛИ не змінюй слово на інше, якщо воно не аписане з помилкою
- Зберігай оригінальний стиль лапок ("або «»), НЕ перетворюй один тип лапок на інший
- Зберігай оригінальне використання великих/малих літер, якщо це не явна помилка
- Якщо граматична форма є прийнятною, залишай її, навіть якщо можлива й інша форма
- Якщо є сумнів, НЕ змінюй
- Якщо помилок немає, повертай оригінальний текст БЕЗ ЗМІН

Приклади:

Вхід: Так само потерпає Україна і сьогодні від того що насправді талановитим людям заважають працювати усілякі посередності "у руля".
Вихід: Так само потерпає Україна і сьогодні від того, що насправді талановитим людям заважають працювати усілякі посередності "у руля".

Вхід: Я ніколи не навчався у медичному коледжі, кажу, - Я лиш підійшов як звичайна людина подивитись чи можу бути чимось корисний.
Вихід: Я ніколи не навчався у медичному коледжі, кажу, - Я лиш підійшов як звичайна людина подивитись, чи можу бути чимось корисний.

Вхід: Це у місті швидка приїжджає, забирає хворого і везе у лікарню; якщо ж до лікарні кілька годин льоту то хворого можна і не довезти.
Вихід: Це у місті швидка приїжджає, забирає хворого і везе у лікарню; якщо ж до лікарні кілька годин льоту, то хворого можна і не довезти.

Вхід: Найбільше він любив тримати в руках старанно орнаментовані стародавні шолом і меч, котрі своїм золотом, здається гріли старого.
Вихід: Найбільше він любив тримати в руках старанно орнаментовані стародавні шолом і меч, котрі своїм золотом, здається, гріли старого.

Вхід: - Я часто казав тобі, що ти дурненька, - сказав він.
Вихід: — Я часто казав тобі, що ти дурненька, — сказав він.

Вхід: Така традиція також походить з Візантії, прикладом є зображення Андроніка II Палеолога.
Вихід: Така традиція також походить із Візантії, прикладом є зображення Андроніка II Палеолога.

Вхід: Як і більшість ділових людей, він не знав напамять жодного вірша і не памятав жодної казки, а тому щоразу мусив імпровізувати.
Вихід: Як і більшість ділових людей, він не знав напам'ять жодного вірша і не памятав жодної казки, а тому щоразу мусив імпровізувати.

Вхід: Настя, привіт! я хотіла уточнити про завдання Олі.
Вихід: Насте, привіт! Я хотіла уточнити про завдання Олі.

Вхід: Я принесу твої улюблені солодощі та обніму тебе міцно-міцно.
Вихід: Я принесу твої улюблені солодощі та обніму тебе міцно-міцно.

Вхід: Смакота ще та, скажу я вам))
Вихід: Смакота ще та, скажу я вам))

Вхід: У той час глибокий сенс народної мудрості нам був ще недоступним через брак досвіду.
Вихід: У той час глибокий сенс народної мудрості нам був ще недоступним через брак досвіду.

Вхід: Каналізація в будинках еволюціонує повільно, але все ж таки змінюється.
Вихід: Каналізація в будинках еволюціонує повільно, але все ж таки змінюється.

Вхідне речення: {input_text}
Виправлене речення:

English: The system prompt is identical to the minimal-edits zero-shot prompt (A.3.1); see the English translation there. The few-shot examples are the same Ukrainian GEC sentence pairs as in the few-shot EN variant (A.2.1). Keywords: Приклади = *'Examples';* Вхід/Вихід = *'Input'/'Output'.*

A.5 Optimized prompts

The following prompts were produced by LLM-assisted prompt optimization (Section 3). Each was derived from its parent prompt via iterative refinement on the validation set (see Section 3 for the optimization procedure).

A.5.1 Minimal-edits + few-shot + optimized-v1 (UA): optimized on GPT-4.1-mini

Ти — система виправлення українських граматичних помилок. Внось МІНІМАЛЬНІ зміни, щоб виправити ЛИШЕ безсумнівні граматичні, орфографічні та пунктуаційні помилки. НЕ переписуй, не перефразовуй і не замінюй слова синонімами.

Виправляй:
- Орфографічні помилки (друкарські помилки, пропущені/зайві літери, неправильне написання разом/окремо: незважати → не зважати, буд-якому → будь-якому).
- Пунктуацію: пропущені коми перед підрядними сполучниками (що, який, бо, чи, коли, щоб, де, поки), при звертаннях, при вставних словах (мабуть, може, звичайно). У прямій мові дефіс (-) замінюй на тире (—): "текст сказав → "текст — сказав.
- Відмінкові помилки (зокрема кличний відмінок у звертаннях: Привіт Настя → Привіт, Насте).
- Узгодження роду, числа, відмінка в словосполученнях.
- Прийменники: о/об (об одинадцятій, о третій); з → зі/із перед збігом приголосних (з зображенням → зі зображенням, з Візантії → із Візантії).

НЕ змінюй:
- НІКОЛИ не замінюй слова синонімами і не змінюй форми слів на альтернативні (виказував, кучею, достойною, вивести — залишай як є).
- НЕ переформатовуй діалоги: якщо діалог оформлений лапками ("/«»), зберігай лапки, НЕ замінюй їх на тире.
- НЕ змінюй граматичні форми, які є допустимими варіантами: відмінкові форми прикметників (недоступним/недоступний), варіанти дієслів (навчались/навчалися), форми займенників (їх/їхній) — якщо форма граматично допустима, залишай її.
- Стиль та тон тексту: неформальний текст (чати, смс) залишай як є — не додавай крапки в кінці, не прибирай смайлики)).
- Порядок слів у реченні.
- Великі/малі літери, крім початку речення після крапки.
- Лапки: зберігай оригінальний стиль.
- Розділові знаки кінця речення: НЕ змінюй . на ? або навпаки.
- НЕ додавай тире (—) там, де його не було в оригіналі, окрім прямої мови.
- Дефіс у складених словах та повторах (міцно-міцно, дере-дере-дере, все-таки, все ж таки — залишай як є).
- Якщо сумніваєшся — НЕ змінюй.

[Ті самі приклади, що і в Prompt 2] *('[Same examples as in Prompt 2]')*

Поверни ЛИШЕ виправлений текст. *('Return ONLY the corrected text.')*

English translation of the instruction part:

You are a Ukrainian grammatical error correction system. Make MINIMAL changes to fix ONLY unambiguous grammatical, spelling, and punctuation errors. Do NOT rewrite, rephrase, or substitute synonyms.

Fix:

- Spelling errors (typos, missing/extra letters, incorrect joined/separate writing: nezvazhaty → ne zvazhaty, bud-yakomu → bud'-yakomu).
- Punctuation: missing commas before subordinate conjunctions (shcho, yakyi, bo, chy, koly, shchob, de, poky), in forms of address, with parenthetical words (mabut', mozhe, zvychaino). In direct speech, replace hyphen (-) with em-dash (—).
- Case errors (especially vocative in address: Pryvit Nastia → Pryvit, Naste).
- Gender, number, case agreement in phrases.
- Prepositions: o/ob; z → zi/iz before consonant clusters.

Do NOT change:
- NEVER substitute synonyms or change word forms to alternatives — leave as is.
- Do NOT reformat dialogues: if dialogue uses quotes, keep quotes, do NOT replace with dashes.
- Do NOT change grammatically acceptable variant forms.
- Text style and tone: leave informal text (chats, SMS) as is.
- Word order, capitalization (except after period), quote style, sentence-final punctuation.
- Do NOT add em-dashes where there were none, except in direct speech.
- Hyphens in compound words and repetitions — leave as is.
- If in doubt — do NOT change.

A.5.2 Minimal-edits + few-shot + optimized-v2 (UA): optimized on Gemini 3-Flash

The prompt below was derived from minimal-edits + few-shot + optimized-v1 (A.5.1) by further LLM-assisted optimization on Gemini 3-Flash over 5 iterations on the validation set. It includes additional Ukrainian-specific rules discovered during optimization.

Ти — система виправлення українських граматичних помилок. Внось МІНІМАЛЬНІ зміни, щоб виправити ЛИШЕ безсумнівні граматичні, орфографічні та пунктуаційні помилки. НЕ переписуй, не перефразовуй і не замінюй слова синонімами.

Виправляй:
- Орфографічні помилки (друкарські помилки, пропущені/зайві літери, неправильне написання разом/окремо: незважати → не зважати, буд-якому → будь-якому).
- Пунктуацію: пропущені коми перед підрядними сполучниками (що, який, бо, чи, коли, щоб, де, поки), при звертаннях, при вставних словах (мабуть, може, звичайно). У прямій мові дефіс (-) замінюй на тире (—): "текст сказав → "текст — сказав.
- Відмінкові помилки (зокрема кличний відмінок у звертаннях: Привіт Настя → Привіт, Насте).
- Узгодження роду, числа, відмінка в словосполученнях.
- Прийменники: о/об (об одинадцятій, о третій); з → зі/із перед збігом приголосних (з зображенням → зі зображенням, з Візантії → із Візантії).

НЕ змінюй:
- НІКОЛИ не замінюй слова синонімами і не змінюй форми слів на альтернативні (виказував, кучею, достойною, вивести — залишай як є).
- НЕ переформатовуй діалоги: якщо діалог оформлений лапками ("/«»), зберігай лапки, НЕ замінюй їх на тире.
- НЕ змінюй граматичні форми, які є допустимими варіантами: відмінкові форми прикметників (недоступним/недоступний), варіанти дієслів (навчались/навчалися), форми займенників (їх/їхній) — якщо форма граматично допустима, залишай її.
- Стиль та тон тексту: неформальний текст (чати, смс) залишай як є — не додавай крапки в кінці, не прибирай смайлики)).
- Порядок слів у реченні.
- Великі/малі літери, крім початку речення після крапки.
- Лапки: зберігай оригінальний стиль.
- Розділові знаки кінця речення: НЕ змінюй . на ? або навпаки.
- НЕ додавай тире (—) там, де його не було в оригіналі, окрім прямої мови.
- Дефіс у складених словах та повторах (міцно-міцно, дере-дере-дере, все-таки, все ж таки — залишай як є).
- Якщо сумніваєшся — НЕ змінюй.

Приклади:

Вхід: Так само потерпає Україна і сьогодні від того що насправді талановитим людям заважають працювати усілякі посередності "у руля".
Вихід: Так само потерпає Україна і сьогодні від того, що насправді талановитим людям заважають працювати усілякі посередності "у руля".

Вхід: Це пов'язано з тим, що такі колективні рухи молекул води сильно збільшують характерні часи процесів які відбуваються в системі.
Вихід: Це пов'язано з тим, що такі колективні рухи молекул води сильно збільшують характерні часи процесів, які відбуваються в системі.

Вхід: Я ніколи не навчався у медичному коледжі, кажу, - Я лиш підійшов як звичайна людина подивитись чи можу бути чимось корисний.
Вихід: Я ніколи не навчався у медичному коледжі, кажу, - Я лиш підійшов як звичайна людина подивитись, чи можу бути чимось корисний.

Вхід: Це у місті швидка приїжджає, забирає хворого і везе у лікарню; якщо ж

до лікарні кілька годин льоту то хворого можна і не довезти.
Вихід: Це у місті швидка приїжджає, забирає хворого і везе у лікарню; якщо ж до лікарні кілька годин льоту, то хворого можна і не довезти.

Вхід: Найбільше він любив тримати в руках старанно орнаментовані стародавні шолом і меч, котрі своїм золотом, здається гріли старого.
Вихід: Найбільше він любив тримати в руках старанно орнаментовані стародавні шолом і меч, котрі своїм золотом, здається, гріли старого.

Вхід: - Я часто казав тобі, що ти дурненька, - сказав він.
Вихід: — Я часто казав тобі, що ти дурненька, — сказав він.

Вхід: Така традиція також походить з Візантії, прикладом є зображення Андроніка II Палеолога.
Вихід: Така традиція також походить із Візантії, прикладом є зображення Андроніка II Палеолога.

Вхід: Як і більшість ділових людей, він не знав напамять жодного вірша і не памятав жодної казки, а тому щоразу мусив імпровізувати.
Вихід: Як і більшість ділових людей, він не знав напам'ять жодного вірша і не памятав жодної казки, а тому щоразу мусив імпровізувати.

Вхід: Настя, привіт! я хотіла уточнити про завдання Олі.
Вихід: Насте, привіт! Я хотіла уточнити про завдання Олі.

Вхід: Я принесу твої улюблені солодощі та обніму тебе міцно-міцно.
Вихід: Я принесу твої улюблені солодощі та обніму тебе міцно-міцно.

Вхід: Смакота ще та, скажу я вам))
Вихід: Смакота ще та, скажу я вам))

Вхід: У той час глибокий сенс народної мудрості нам був ще недоступним через брак досвіду.
Вихід: У той час глибокий сенс народної мудрості нам був ще недоступним через брак досвіду.

Вхід: Каналізація в будинках еволюціонує повільно, але все ж таки змінюється.
Вихід: Каналізація в будинках еволюціонує повільно, але все ж таки змінюється.

Вхід: "Досить добре вийшов з Весту, чи не так? спитав міліціонер.
Вихід: "Досить добре вийшов з Весту, чи не так?"— спитав міліціонер.

Вхід: Не знаю як у інших, а у мене в житті траплялось не так багато див.
Вихід: Не знаю, як у інших, а у мене в житті траплялось не так багато див.

Вхід: Хочеться закінчити у дусі книг з самодопомоги.
Вихід: Хочеться закінчити у дусі книг із самодопомоги.

Вхід: Наступного дня тим же автобаном повернулися назад, звідти – ще півтори години літаком.
Вихід: Наступного дня тим же автобаном повернулися назад, звідти — ще півтори години літаком.

Вхід: От читаєш такі новини і жалкуєш, що населенню України Господь дав все, крім совісті і мозгів.
Вихід: От читаєш такі новини і жалкуєш, що населенню України Господь дав усе, крім совісті і мозгів.

Поверни ЛИШЕ виправлений текст. *('Return ONLY the corrected text.')*

English: The instruction structure is the same as optimized-v1 (A.5.1); see the translation there. This version includes additional few-shot examples (lines 6–16 above) and was further refined on Gemini 3-Flash. Keywords: Виправляй = *'Fix';* НЕ змінюй = *'Do NOT change';* Приклади = *'Examples';* Вхід/Вихід = *'Input'/'Output'.*

B Inference Pipeline and Structured Output

Each input sentence is processed independently through a single LLM call, with no cross-sentence batching. Before any model invocation, the pipeline applies a lightweight passthrough rule: lines matching the document-marker pattern `# <digits>` (e.g. `# 0001`) are emitted verbatim and never sent to the LLM. These markers delimit document boundaries in the UA-GEC corpus and carry no correctable content, so routing them around the model both saves tokens and prevents spurious edits.

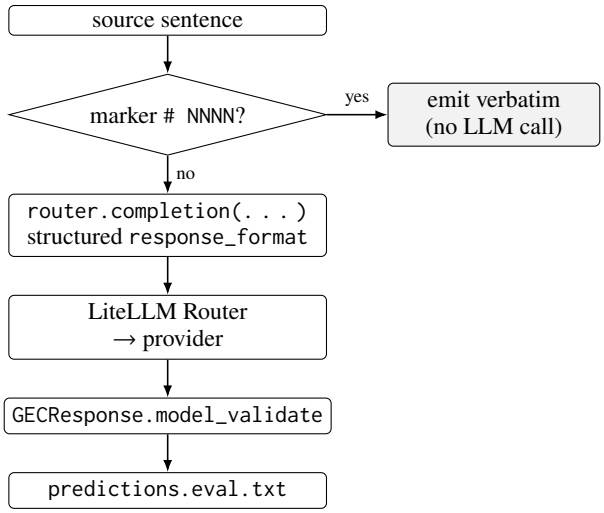
For all remaining sentences, the agent issues one chat completion through a LiteLLM router that abstracts over the underlying provider (OpenAI, Anthropic, Google, Moonshot). The system message is the configured prompt template; the user message is the raw source sentence. To eliminate free-form post-processing of model output, we constrain the response with a JSON schema derived from a Pydantic model and attached to the request as a strict `response_format`. Field descriptions declared on the Pydantic model propagate into the schema and act as in-band instructions to the model.

Schema definition. The response contract is declared once as a Pydantic class:

```
class GECResponse(BaseModel):
    corrected_sentence: str = Field(
        ..., description="Corrected version of
            the input sentence")
```

Generated JSON schema. At call time, the class is converted to JSON Schema, all object nodes are closed with `additionalProperties: false`, and the result is wrapped into the provider-agnostic `response_format` envelope:

```
{
  "type": "json_schema",
  "json_schema": {
    "name": "GECResponse",
    "strict": true,
    "schema": {
      "type": "object",
      "additionalProperties": false,
      "required": ["corrected_sentence"],
      "properties": {
        "corrected_sentence": {
          "type": "string",
          "description": "Corrected version of
              the input sentence"
        }
      }
    }
  }
}
```



Parameter	Value (from run YAML)
model	gpt-4.1-mini
temperature	0.0
top_p	0.1
reasoning_effort	None
timeout	90 s
response_format	json_schema(GECResponse)

Figure 3: Per-sentence inference flow. Document markers bypass the LLM; all other sentences go through one structured-output call. Decoding parameters are read verbatim from the run YAML, ensuring deterministic replay.

The decoding parameters in Figure 3 are read from the run YAML and the exact configuration file is copied into the output directory alongside `results.json`, so a run can be replayed bit-for-bit given the same provider model snapshot. The returned payload is validated with `model_validate`, so any schema violation is caught deterministically rather than being masked by string heuristics. If a provider rejects `json_schema`, the client transparently retries with `response_format = {"type":"json_object"}`; for the single-field case, a final recovery path extracts the corrected sentence from malformed JSON to keep evaluation aligned. This design ensures that every non-marker sentence yields exactly one validated correction, making the sentence-to-prediction mapping bijective and the run reproducible given a fixed configuration.

Data-Efficient Adaptation of Multilingual LLMs to Ukrainian

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

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

Correspondence: paniv@ucu.edu.ua

Abstract

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

1 Introduction

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

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

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

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

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

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

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 155–168
May 29-30, 2026

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

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

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

2 Related Work

2.1 Multilingual Language Models

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

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

2.2 Ukrainian NLP

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

2.3 Data Quality for Pretraining

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

3 Model Architecture and Tokenizer

3.1 Base Model Selection

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

3.2 Tokenizer

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4 Pretraining Data Collection and Filtering

4.1 Data Sources

Our pretraining data combines several sources.

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

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

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

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

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

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

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

4.2 Filtering Pipeline

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

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

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

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

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

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

Overall, our quality dimensions include:

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

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

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

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

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

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

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

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

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

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

5 Pretraining

5.1 Training Setup

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

5.2 Data Mixing Strategy

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

5.3 Results

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

6 Instruction-Tuning

6.1 Dataset Creation

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

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

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

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

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

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

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

Table 5: Pretraining hyperparameters for pretrained and instruct models.

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

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

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

6.2 Vision Instruction Data

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

6.3 Results

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

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

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

7 Additional experiments

7.1 Reasoning Capabilities

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

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

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

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

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

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

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

7.2 Persona Editing

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

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

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

8 Evaluation

8.1 Benchmark Suite

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

9 Discussion

9.1 Societal Impact

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

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

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

10 Conclusion

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

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

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

Limitations

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

Acknowledgements

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

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

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

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

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

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

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

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

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

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

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

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

A Tokenizer Replacement Statistics

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

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

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

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

A.1 Additional Tokenizer Efficiency Metrics

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

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

B Author Contributions

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

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

Dictionary-Based Speculative Decoding for Non-Latin-Script Languages

Oleksiy Syvokon
Lviv Polytechnic National University
oleksii.o.syvokon@lpnu.ua

Abstract

Large language models tokenize non-Latin-script languages inefficiently: a single word in Ukrainian or Crimean Tatar is split into two to three times as many tokens as its English equivalent. We propose *dictionary-based speculative decoding* (DictSpec), which accelerates inference by proposing draft continuations from a static n-gram lookup table built offline from an unlabeled corpus. The lookup table requires no trainable parameters or GPU resources, is inexpensive to construct, adds under 5 MB of memory overhead, and can be reused across models that share a tokenizer. We evaluate DictSpec on Ukrainian and Crimean Tatar (Cyrillic and Latin scripts), implementing a vLLM plugin to benchmark five models ranging from 3B to 70B parameters on consumer- and server-grade GPUs. In controlled emulation, DictSpec reduces verification steps by up to 1.65×, with gains correlating substantially with tokenizer fertility. In live vLLM serving, pure DictSpec gives modest speedups, while a hybrid with prompt-local n-gram speculation reaches up to 1.76×. We release our code and vLLM plugin as open source.[1]

1 Introduction

Inference cost for large language models is not uniform across languages. Tokenizers trained on English-dominated corpora assign most common English words a single token but fragment words in non-Latin scripts into many subword pieces. Each piece requires a separate forward pass through the model. For languages such as Ukrainian, Georgian, Hindi, and Crimean Tatar, this creates a *tokenization tax* that makes inference substantially slower (Rust et al., 2021). We target this latency gap without modifying the model or tokenizer.

Our main insight is that this inefficiency is not uniformly distributed across decoding steps. The model faces a genuinely difficult decision when it must choose which word or phrase to begin. But once the initial tokens of a word have been committed, the remaining tokens are often highly predictable. A model that has generated the Ukrainian token prefix "`_пер·сон·аль·ний _ком`" almost certainly intends to continue with "`п'·ют·ер`", completing *"персональний комп'ютер"* ("personal computer").

Standard LLM inference spends the same compute on a predictable token as on a genuinely hard one. *Speculative decoding* exploits this gap: a cheap draft mechanism proposes a multi-token continuation, and the target model verifies the entire proposal in a single forward pass, accepting correct tokens and falling back on mismatches.

We propose *dictionary-based speculative decoding* (DictSpec), which realizes this idea with a simple draft mechanism: a static lookup table built offline from a corpus in the target language. The table maps short token-sequence prefixes to their most probable continuations. For instance, the table maps the word prefix "*головнок*" ("commander-i…") to the continuation completing "*головнокомандувач*" ("commander-in-chief"). When more preceding context is available, even shorter prefixes become predictable: the tokens for "*верховний го*" ("supreme commander-i…") are enough to predict the full word "*головнокомандувач*", because "*верховний*" ("supreme") strongly disambiguates what follows. At inference time, the method checks the tail of the tokens generated so far against the lookup table. If a matching prefix is found, the stored continuation is proposed as a draft. The target model then verifies the draft using the standard speculative decoding protocol (Leviathan et al., 2023; Chen et al., 2023). If the proposed tokens are accepted, multiple decoding

[1] https://github.com/osyvokon/dictspec

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 169–183

steps are resolved in a single verification pass. If any proposed token is rejected, decoding falls back to standard autoregressive generation from that point. Because we use the standard speculative decoding verification, the output distribution is provably identical to that of unmodified autoregressive decoding.

Building the lookup table requires only an unlabeled monolingual text corpus, takes a few minutes on a consumer-grade laptop without a GPU, and produces a compact data structure requiring under 5 MB of CPU memory. Because a dictionary is tied to a tokenizer and language/script rather than to model weights, a single table serves every model in a family and can be reused without modification across inputs and sessions.

We evaluate the method across seven checkpoints from five tokenizer families (Gemma 3, Mistral, Mistral NeMo, Qwen 3.5, and Llama 3) in a controlled emulation study over Ukrainian and Crimean Tatar in both Cyrillic and Latin scripts, sweeping dictionary construction parameters across 1,350 total configurations, and validate with wall-clock benchmarks in vLLM.

Our main contributions are:

- We propose DictSpec, a simple and lightweight draft mechanism that accelerates LLM inference for non-Latin-script languages while preserving the target model's output distribution exactly.
- We find that speedup correlates substantially with tokenizer fertility (Pearson $r = 0.72$). The method helps most where tokenization is least efficient.
- We integrate the method into vLLM and evaluate pure DictSpec, prompt n-gram speculation, and a hybrid of the two. Pure DictSpec yields modest positive wall-clock gains on coherent outputs, while the hybrid reaches up to 1.76×.

2 Related Work

2.1 Tokenization Inefficiency in Non-Latin Scripts

Subword tokenization methods such as Byte-Pair Encoding (Sennrich et al., 2016) and SentencePiece (Kudo and Richardson, 2018) build vocabularies that reflect the statistical distribution of the training corpus. When the corpus is dominated by English and other high-resource Latin-script languages, the resulting tokenizer allocates most vocabulary entries to those languages and offers poor coverage of non-Latin scripts. Rust et al. (Rust et al., 2021) showed that tokenizer quality varies dramatically across languages, and that poor tokenizers inflate sequence length and degrade model performance for morphologically rich or non-Latin-script languages. This over-segmentation increases both memory consumption and latency, since each additional token requires a separate autoregressive forward pass. Recent work has documented similar findings for Ukrainian specifically (Maksymenko and Turuta, 2025; Kiulian et al., 2024; Kiulian et al., 2025), and broader studies have shown that tokenizer fragmentation introduces systematic cost and latency disparities across languages (Petrov et al., 2023; Ahia et al., 2023; Hong et al., 2024).

2.2 Extending or Retraining Tokenizers

One line of work addresses tokenization inefficiency by training a new tokenizer or extending the existing vocabulary to better cover the target language (Minixhofer et al., 2022; Cui et al., 2023; Kiulian et al., 2025). Lapa LLM reports replacing a large portion of the base Gemma vocabulary with Ukrainian-specific tokens in order to reduce tokenization cost and improve efficiency for Ukrainian generation (Paniv et al., 2025). Such methods can substantially reduce tokenizer fertility and improve generation quality for underrepresented languages, but they require re-embedding the new vocabulary tokens and continued pretraining on the target language. Moreover, this adaptation must be repeated for every model architecture and size, which makes it a costly process that does not transfer across models. They may also degrade performance on non-target languages by redistributing vocabulary capacity. Our method requires no modification of the model or its tokenizer, and the output distribution remains exactly that of the original model.

2.3 Speculative Decoding

Neural draft models. The idea of verifying multiple predicted tokens in parallel was proposed by (Stern et al., 2018). Speculative decoding was later formalized with distribution-preserving guarantees by (Leviathan et al., 2023) and (Chen et al., 2023). The original formulation uses a smaller language model as the draft model; the target model verifies the proposed block in a single forward pass, accepting or rejecting tokens according

	Step	Example
1.	Extracted n-grams	`персональний комп'ютер (count=12,400)` `персональний комунікатор (count=180)`
2.	Tokenized forms	`_пер сон аль ний _ком п' ют ер;` `_пер сон аль ний _ком уні ка тор`
3.	Context prefix	`_пер сон аль ний _ком`
4.	Candidate continuations	`п' ют ер (prob=98.6%; selected)` `уні ка тор (prob=1.4%; discarded)`
5.	Lookup entry (text)	`_пер сон аль ний _ком → п' ют ер`
6.	Lookup entry (ids)	`[12, 47, 91, 103, 255] → [301, 404, 509]`

Table 1: Schematic illustration of the dictionary construction procedure for the Ukrainian n-gram `персональний комп'ютер` ("personal computer"). Spaces mark token boundaries; _ denotes a token with a leading space.

to a rule that preserves its marginal distribution. A practical limitation is that a suitable draft model must be obtained for each target model family and language: a good English draft model may be a poor draft model for Ukrainian, and training one for a new language requires additional training data and GPU resources.

Architectural methods. A second line of work embeds draft generation into the target model itself. Medusa (Cai et al., 2024) adds multiple decoding heads, each predicting a different future token; EAGLE (Li et al., 2024) builds a lightweight draft network from the target model's hidden states. Both achieve strong speedups but require architecture modifications and additional training for the target model. Most relevant to our motivation, Hong et al. (Hong et al., 2024) target the speed penalty caused by excessive tokenization of non-Latin scripts with a language-specific decoding head (reporting 1.7× speedup), but their method still requires training the new head for each target language and model.

Non-neural methods. The approach closest to ours replaces the neural draft model with n-grams drawn from the prompt or previously generated text (Saxena, 2023). Prompt n-grams work well when output closely resembles input (e.g., document editing or summarization) but offer lower coverage in open-ended generation. REST (He et al., 2024) broadens coverage by retrieving continuations from a datastore. More recent work builds n-gram tries from the input context for in-context learning (Chen et al., 2025), and Stewart et al. (Stewart et al., 2024) show that learning-free n-gram statistics alone can yield competitive speedups. Our method differs in that the lookup table is built offline from a large corpus, giving high coverage of common continuations in the target language even when the prompt provides little relevant context. The two strategies are complementary: prompt n-grams exploit repetition within the current context, whereas DictSpec exploits corpus-level regularities of the target language. We evaluate both methods separately and in combination in our vLLM experiments.

3 Method

3.1 Dictionary Construction

Our method builds a lightweight lookup table that maps token prefixes to their most probable continuations. Building the table requires only an unlabeled text corpus in the target language, is performed entirely offline, and needs no GPU or model access. Table 1 illustrates the full procedure on a concrete example.

We first extract frequent words and short n-grams with their counts (Step 1) and tokenize each using the target model's tokenizer (Step 2). From the resulting token sequences we build prefix-to-continuation statistics: for each shared token prefix (Step 3) we record which continuation is most likely (Step 4) and store only that single most probable suffix (Steps 5–6). Unlike a standard count-based n-gram language model that estimates the full next-token distribution, our table keeps only the top continuation. This approach, combined with an efficient trie (Yata, 2023) implementation, allows us to fit the whole lookup table in just 2 to 5 MB, with median lookup latency below 7 µs.

To keep the table compact and improve acceptance probability, we apply two filters. First, we

retain only the most frequent n-grams. Second, we apply a minimum probability threshold so that we do not store ambiguous continuations, which are likely to be rejected during verification. The resulting dictionary is tied to the tokenizer and language/script, so it can be reused across models in a family. In our setting, building it takes a few minutes on a consumer-grade laptop.

3.2 Speculative Decoding

Speculative decoding accelerates autoregressive generation by having a cheap draft mechanism propose a short continuation of up to γ tokens (the *speculative budget*) that the target model then verifies in a single forward pass. The target model checks the proposal left to right: accepted tokens are appended and generation advances by several positions at the cost of one verification step; when a mismatch occurs, verification stops and the target model supplies the correct next token. Poor proposals reduce efficiency but never compromise correctness. In the worst case the method degrades to standard autoregressive decoding, and the standard verification procedure preserves the exact output distribution of the target model (Leviathan et al., 2023).

3.3 Dictionary-Based Speculative Decoding

The target model must still decide which word or phrase to begin, but once the first part of a word has been generated, the rest is often predictable. Our method uses the dictionary to propose that predictable continuation. The target model then verifies it.

At each decoding step we search the dictionary for the longest suffix of the current token history that appears as a key, backing off to shorter contexts only if necessary. This longest-match rule yields the most specific completion. For instance, referring to Table 1, after the model has generated tokens `_пер сон аль ний _ком`, the dictionary matches the full five-token prefix and proposes the continuation `п' ют ер` (Step 5). The stored continuation is verified by the target model. If the model agrees, all three tokens are accepted in one step; if it instead prefers "комунікатор", verification stops at the first mismatch and the model supplies the correct token. The example also shows why longer context matters: `_ком` alone is ambiguous, whereas the full prefix is specific. When no match is found, DictSpec emits no draft, and decoding proceeds with a standard model step.

Because the dictionary operates over token IDs, it can match partial sequences inside words, not only at word boundaries, so speculation can begin as soon as the recent suffix becomes distinctive.

The dictionary and prompt-local n-gram speculation provide complementary draft sources. DictSpec can propose common language-level continuations that have not appeared in the current prompt, while prompt n-grams can exploit repetition in the input or generated context. In the vLLM experiments we therefore evaluate not only pure DictSpec, but also a hybrid mode: the proposer first queries the corpus dictionary and, when no dictionary draft is available, falls back to vLLM's prompt n-gram proposer. Both sources are still verified by the target model under the standard speculative decoding protocol, so the output distribution is unchanged.

4 Experiments

We evaluate in two stages. First, we run a controlled emulation (Section 4.5) that isolates the speculative decoding mechanism from system-level variables by replaying pre-tokenized reference text. Second, a wall-clock validation with vLLM (Section 4.6) confirms that the emulated gains translate to real throughput improvements in a production serving engine.

4.1 Models

We evaluate seven recent checkpoints from five tokenizer families: Gemma 3 (Gemma Team, 2025), Mistral (Jiang et al., 2023), Mistral NeMo, Qwen 3.5 (Team, 2026), and Llama 3 (Grattafiori et al., 2024). Because each dictionary is tied to the tokenizer, models within a family produce identical emulation results; we therefore report one row per family in the emulation study. For wall-clock validation we select models at two scales: 3–4B on a consumer GPU and 27–70B on a server GPU (Section 4.6). We include the older Mistral 7B v0.1 to assess whether tokenizer quality has improved over three years.

4.2 Languages and Datasets

We focus on settings where tokenization inefficiency is especially relevant: Ukrainian as a mid-resource language using Cyrillic script, and Crimean Tatar in Cyrillic and Latin scripts as a low-resource language. The Latin-script Crimean Tatar setting serves as a contrastive low-resource condition.

Language		Train words	Valid words
Ukrainian		251,618,586	124,703
Crimean (Cyrillic)	Tatar	1,333,103	535,156
Crimean (Latin)	Tatar	118,749	16,020

Table 2: Dataset statistics for dictionary construction (train) and evaluation (valid).

Model	en	uk	cr-cyr	cr-lat
Gemma 3	1.28	2.38	3.30	2.98
Mistral	1.42	3.18	4.20	3.92
Mistral NeMo	1.26	2.62	3.51	3.07
Qwen 3.5	1.30	2.59	3.51	3.01
Llama 3.3	1.23	2.38	3.89	2.94

Table 3: Tokenizer fertility (average tokens per word) for each model–language pair. Higher values indicate more fragmented tokenization.

For Ukrainian we build the dictionary from a sample of CulturaX (Nguyen et al., 2024). For Crimean Tatar we use the QIRIM Crimean Tatar monocorpus (QIRI'M Young, 2024), split by script. In all cases the training-side corpus is used only for offline dictionary construction and never for target-model adaptation. Table 2 summarizes the dataset sizes.

4.3 Tokenizer Fertility

We quantify tokenization inefficiency using *tokenizer fertility*: the average number of tokens per whitespace-delimited word. We compute fertility on the parallel sentences of FLORES+ (NLLB Team et al., 2024) for English, Ukrainian, and Crimean Tatar Latin. For Crimean Tatar Cyrillic, which is absent from FLORES+, we use the QIRIM validation split.

As Table 3 shows, English fertility is uniformly low (1.2–1.4 tokens per word), while Crimean Tatar Cyrillic is the most fragmented (3.3–4.2), followed by Crimean Tatar Latin (2.9–3.9) and Ukrainian (2.4–3.2). Newer tokenizers reduce fragmentation but still exceed 3 tokens per word on Crimean Tatar Cyrillic.

4.4 Dictionary Construction

The offline dictionary is built from frequent words and short n-grams extracted from the training corpus, tokenized with the target model's tokenizer. For each prefix we retain only the most probable continuation above a minimum probability threshold, truncating both prefixes and completions to at most 8 tokens.

We sweep three dictionary-construction parameters:

- N-gram order: unigrams (`n1`), up to bigrams (`n12`), up to trigrams (`n123`).
- Dictionary size: 10k, 50k, 100k, 200k, 500k, 1M entries.
- Minimum continuation probability: 0.2, 0.5, 0.8, 0.9, 1.0.

4.5 Controlled Emulation Study

We first evaluate with an exact speculative-decoding emulator that replays pre-tokenized reference text, simulating the draft-then-verify cycle under identical conditions across all configurations. This isolates the speculative mechanism from factors such as batching, memory bandwidth, and GPU kernel implementation, and is cheap to run and repeat. Because the reference text is drawn from a static corpus rather than generated by a model, acceptance rates may differ from those observed during real inference; we address this gap with wall-clock experiments on live models in Section 4.6.

At each step the emulator queries the dictionary for the longest matching suffix (up to 8 tokens of context) and proposes a continuation of up to $\gamma = 8$ tokens. The sweep covers 5 tokenizer families, 3 language/script settings, 3 n-gram orders, 6 dictionary sizes, and 5 probability thresholds, for a total of 1,350 runs.

4.5.1 Metrics

Emulated speedup. The ratio of total tokens generated to the number of target-model verification steps, i.e. $\frac{N_{\text{tokens}}}{N_{\text{steps}}}$. Under standard autoregressive decoding every token requires one step, so this ratio is 1; values above 1 indicate that speculative drafting has reduced the number of required forward passes. This metric provides a hardware-independent measure of speculative efficiency.

Draft coverage. The fraction of decoding steps where the dictionary proposes a non-empty draft.

Lang	Speedup	Coverage	MAL	Accept
Gemma 3				
uk	1.26	0.59	0.44	0.17
cr-cyr	1.45	0.71	0.63	0.18
cr-lat	1.19	0.57	0.33	0.10
Mistral 7B				
uk	1.43	0.70	0.61	0.17
cr-cyr	1.65	0.77	0.84	0.18
cr-lat	1.35	0.69	0.50	0.08
Mistral NeMo				
uk	1.34	0.64	0.53	0.18
cr-cyr	1.51	0.73	0.70	0.18
cr-lat	1.21	0.61	0.34	0.07
Qwen 3.5				
uk	1.34	0.63	0.54	0.18
cr-cyr	1.51	0.72	0.71	0.19
cr-lat	1.21	0.58	0.36	0.11
Llama 3.3				
uk	1.28	0.61	0.46	0.18
cr-cyr	1.61	0.77	0.79	0.18
cr-lat	1.22	0.62	0.36	0.07

Table 4: Best emulated speculative speedup for each model–language pair (best configuration across all sweep dimensions). Coverage = draft coverage, MAL = mean accepted draft length, Accept = token-level draft acceptance rate.

Mean accepted draft length (MAL). The average number of drafted tokens accepted per draft-producing step.

Acceptance rate. The fraction of all individually proposed draft tokens that are accepted by the target model.

4.5.2 Main Results

We first ask whether the method reduces decoding work at all. Table 4 reports the best achieved speedup for each model–language pair in the sweep. Speedups are highest for Crimean Tatar Cyrillic (up to 1.65×), which also has the highest tokenizer fertility, and lowest for Crimean Tatar Latin (1.19–1.35×).

The per-token acceptance rate is modest (0.07–0.19), but this does not prevent meaningful speedups. Because the speculative budget allows drafts of up to 8 tokens, even a draft where only the first 1–2 tokens are accepted still advances generation by 2–3 tokens per verification step (number of accepted tokens plus one "bonus" token).

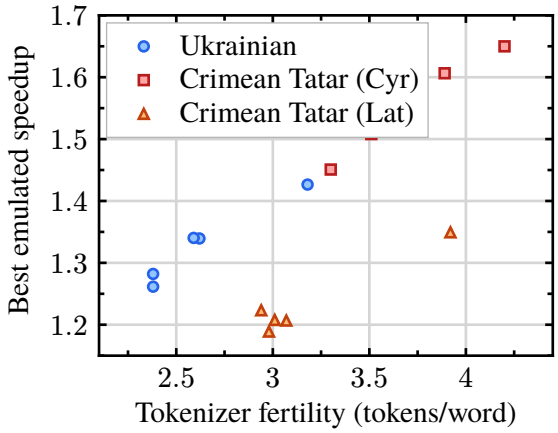


Figure 1: Tokenizer fertility versus best emulated speculative speedup. Higher fertility correlates with larger gains from dictionary-based speculation.

4.5.3 Language and Tokenizer Fertility

Figure 1 plots tokenizer fertility against best achieved speedup for every model–language pair. The correlation is substantial and statistically significant (Pearson $r = 0.72$, $p = 0.002$): Crimean Tatar Cyrillic, the most fragmented setting, consistently achieves the highest speedups. The one outlier is Crimean Tatar Latin, which has higher fertility than Ukrainian yet slightly lower speedups. This is likely because its much smaller training corpus (119k vs. 252M words; Table 2) limits dictionary coverage. This suggests that fertility is the primary driver of gains, provided the dictionary has adequate coverage.

4.5.4 Dictionary Size

Figure 2 shows mean speedup (averaged across models) versus dictionary size for each minimum probability threshold, with the n-gram configuration fixed at `n123`. Speedup increases quickly with dictionary size up to approximately 100k–200k entries, then flattens, suggesting diminishing returns from adding more entries beyond Lower probability thresholds (especially $p_{\min} = 0.2$) give the highest emulated speedups because they preserve more candidate entries, while stricter thresholds trade coverage for higher-confidence drafts. In terms of memory, even the largest dictionaries (1M entries, trigrams) compress to under 5 MB when stored as a trie. For wall-clock validation we therefore use a compact deployment configuration (200k entries, `n123`, $p_{\min} = 0.8$), which occupies

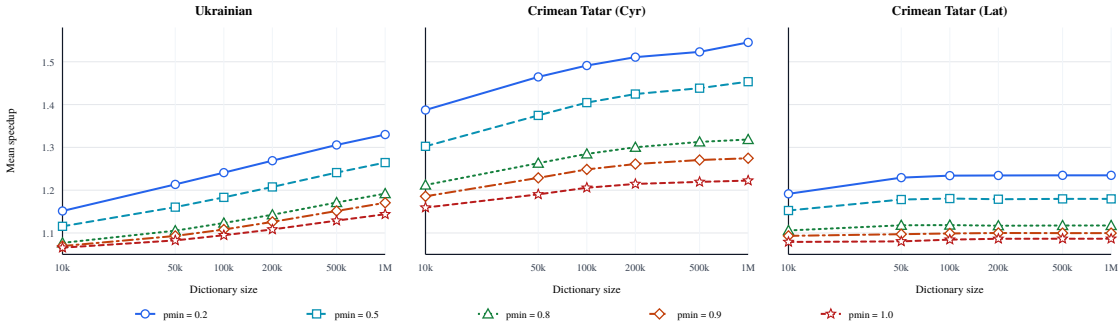


Figure 2: Mean emulated speedup (averaged across models) versus dictionary size for different minimum probability thresholds. N-gram configuration fixed at `n123`.

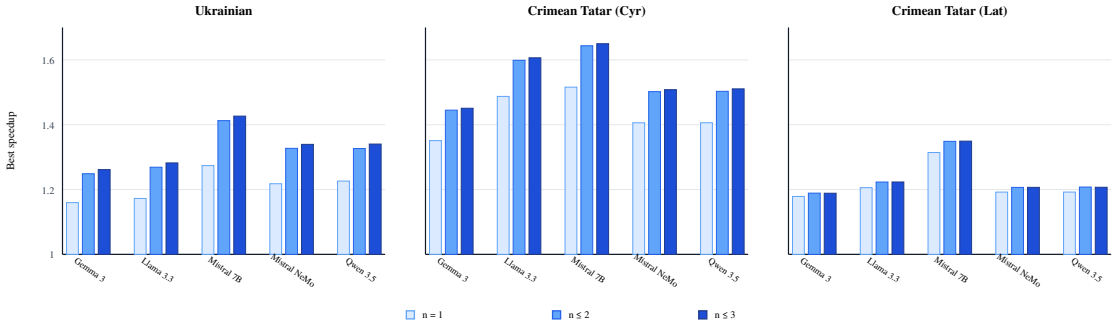


Figure 3: Best emulated speedup by n-gram configuration for each model–language pair.

2–3 MB and keeps the method practical alongside the target model with negligible overhead.

4.5.5 Effect of Longer N-grams

Figure 3 compares `n1`, `n12`, and `n123` for each model–language pair. Adding bigrams (`n12`) consistently improves over unigrams (`n1`), and adding trigrams (`n123`) provides a further but smaller improvement.

4.6 Wall-Clock Validation with vLLM

The emulation study leaves open the question of whether the emulated gains translate to real wall-clock speedups. To answer this, we integrate DictSpec into vLLM (Kwon et al., 2023), an open-source high-throughput serving system, and measure end-to-end generation throughput.

We benchmark five checkpoints listed above on two hardware tiers: Gemma 3 4B, Llama 3.2 3B, and Qwen 3.5 4B on a 16 GB NVIDIA RTX 4090 Laptop GPU; Gemma 3 27B on one 80 GB NVIDIA A100; and Llama 3.3 70B on two 80 GB NVIDIA A100s with tensor parallelism. Models within each family share a tokenizer, so the same trie files from the emulation study apply directly. We use the compact deployment configuration discussed in Section 4.5: `n123`, 200k entries, $p_{\min} = 0.8$. The emulation study uses a speculative budget of $\gamma = 8$ to explore the selected parameter space, but the results show that acceptance drops sharply after the first few draft tokens; we therefore halve the budget to $\gamma = 4$ for the wall-clock experiments.

For each model–language pair we run four configurations: **baseline** (standard autoregressive decoding), **prompt n-gram** speculation using vLLM's built-in prompt-lookup proposer, **DictSpec** (the corpus dictionary alone), and **hybrid** (DictSpec with prompt n-gram fallback). Each configuration serves 20 prompts with a maximum batch size of 5, with a maximum of 2,048 generated tokens per prompt. We report generation throughput (tokens per second) and the token-level acceptance rate for each speculative configuration.

4.6.1 Main Results

Table 5 summarizes the wall-clock benchmarks on coherent outputs. The live results are more

Model	Base tok/s	DictSpec tok/s (×base)	Acc. %	Prompt n-gram tok/s (×base)	Acc. %	Hybrid tok/s (×base)	Acc. %
Ukrainian							
Gemma 3 4B	66.2	66.6 (1.01×)	41.2	75.8 (1.15×)	14.0	**82.8 (1.25×)**	18.0
Llama 3.2 3B	80.7	83.8 (1.04×)	43.6	115.4 (1.43×)	26.6	**122.2 (1.51×)**	28.7
Qwen 3.5 4B	55.6	58.9 (1.06×)	35.2	75.4 (1.36×)	18.4	**75.8 (1.36×)**	18.1
Gemma 3 27B	27.2	27.8 (1.02×)	50.0	29.2 (1.07×)	9.9	**32.1 (1.18×)**	15.9
Llama 3.3 70B	21.1	22.1 (1.05×)	47.8	24.0 (1.14×)	12.9	**28.3 (1.35×)**	23.1
Crimean Tatar (Cyrillic)							
Gemma 3 27B	27.3	29.4 (1.08×)	37.0	**39.5 (1.45×)**	24.0	37.3 (1.37×)	24.7
Llama 3.3 70B	21.2	26.5 (1.25×)	49.0	**45.6 (2.15×)**	46.5	37.4 (1.76×)	36.9
Crimean Tatar (Latin)							
Gemma 3 27B	26.1	26.3 (1.01×)	11.6	**33.9 (1.30×)**	18.3	31.7 (1.22×)	18.5
Llama 3.3 70B	21.2	21.6 (1.02×)	19.4	**29.0 (1.37×)**	24.7	27.4 (1.29×)	18.5

Table 5: Wall-clock generation throughput measured in vLLM on coherent outputs. Throughput columns are reported in tok/s; speculative columns also show speedup over baseline in parentheses. Acc. (%) is the token-level draft acceptance rate. Bold marks the fastest speculative configuration in each row. Models up to 4B run on a 16 GB RTX 4090; Gemma 27B runs on one 80 GB A100; Llama 70B runs on two 80 GB A100s. Dictionary configuration: `n123`, 200k entries, $p_{\min} = 0.8$.

nuanced than the controlled emulation study. Pure DictSpec improves throughput on all coherent rows, but the gains are modest: 1.01–1.25× over baseline. This reflects a practical deployment tradeoff: high acceptance rate is not sufficient by itself; the proposer must also fire often and save enough model steps to overcome lookup and scheduling overhead.

Prompt n-gram speculation is a strong baseline, especially on Crimean Tatar, where the model outputs contain substantial local repetition. The hybrid setting is nevertheless useful: it improves over pure DictSpec on every coherent row and is the fastest configuration for all Ukrainian models. On the larger models, the hybrid reaches 1.18–1.76× speedup, while pure DictSpec reaches 1.01–1.25×.

Manual inspection of the generated outputs reveals that all three smaller models (3–4B parameters) produce severely degenerate text on Crimean Tatar configurations: outputs consist largely of repetitive token loops or wrong-language text (e.g. Kazakh). We therefore omit those rows from Table 5 and use the 3–4B models only for Ukrainian, where their outputs are coherent. The 27B and 70B model outputs are coherent for all three language/script settings.

The wall-clock evaluation refines the interpretation of the method. Pure DictSpec is a high-precision, low-overhead draft source that helps most when tokenizer fertility is high, but as a stand-alone vLLM proposer its live gains are smaller than the emulated reduction in model steps. Prompt n-grams provide strong adaptive coverage when the generated text repeats local context. The best practical deployment is therefore a hybrid view: a small corpus dictionary is a useful complementary draft source that can be combined with prompt-local speculation without changing the target model or its output distribution.

5 Conclusion

We presented DictSpec, a simple method that reduces the inference cost caused by suboptimal tokenization. A lightweight n-gram lookup table, built offline from an unlabeled text corpus, proposes draft token continuations that the target model verifies under the standard speculative decoding protocol, preserving the output distribution exactly. Across seven checkpoints from five tokenizer families and three language/script configurations, emulated DictSpec speedups correlate substantially with tokenizer fertility. In live vLLM serving, pure DictSpec provides modest positive gains on coherent outputs, while a hybrid with prompt-local n-gram speculation reaches up to 1.76×.

The method requires no neural network training, no GPU resources for dictionary construction, and adds less than 5 MB of memory overhead. A single dictionary serves any model that shares the same tokenizer and language/script setting. Thus, DictSpec is most useful as a low-overhead corpus-level draft source for languages whose tokenizers

produce many tokens per word, especially when combined with prompt-local speculation.

Several directions remain for future work. First, combining DictSpec with prompt-based or retrieval-based draft methods (Saxena, 2023; He et al., 2024) could improve coverage and acceptance rates. Second, when the output domain is approximately known (e.g., a medical support system), a domain-specific dictionary could substantially improve acceptance rates. Finally, moving the lookup table to GPU memory, possibly with a custom kernel, could further reduce drafting latency by avoiding CPU-GPU memory communication overhead.

Limitations

DictSpec is most effective when tokenizer fertility is high. For languages with low fertility (e.g., English), the dictionary will match infrequently and gains will be minimal. The method does not improve the quality of the target model's outputs; if the model generates poor-quality text in the target language, our method will not correct this. Similarly, our method does not address the reduced effective context length caused by over-segmentation. Our wall-clock benchmarks cover a limited set of GPU configurations; wall-clock gains may differ under other batched or multi-GPU serving deployments. Finally, this work does not include head-to-head comparisons with approaches that modify the tokenizer itself, such as Lapa LLM (Paniv et al., 2025).

Ethical Considerations

During the preparation of this manuscript, the authors used LLMs extensively to rephrase and polish draft text for improved clarity and readability. The authors reviewed and edited all AI-generated suggestions and take full responsibility for the final content.

Acknowledgments

We thank Dario Stojanovski, Tamara Stankovic and Si-Qing Chen for supporting this work.

References

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R. Mortensen, Noah A. Smith, and Yulia Tsvetkov. 2023. Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D. Lee, Deming Chen, and Tri Dao. 2024. Medusa: Simple LLM Inference Acceleration Framework with Multiple Decoding Heads. In *Proceedings of the 41st International Conference on Machine Learning*, pages 5209–5235.

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. 2023. Accelerating Large Language Model Decoding with Speculative Sampling. *arXiv preprint arXiv:2302.01318*.

Jinglin Chen, Qiwei Li, Zuchao Li, Baoyuan Qi, Guoming Liu, Haojun Ai, Hai Zhao, and Ping Wang. 2025. Faster In-Context Learning for LLMs via N-Gram Trie Speculative Decoding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18040–18051.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and Effective Text Encoding for Chinese LLaMA and Alpaca. *arXiv preprint arXiv:2304.08177*.

Gemma Team. 2025. *Gemma 3 Technical Report*. technical report. Google DeepMind.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, and others. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Zhenyu He, Zexuan Zhong, Tianle Cai, Jason Lee, and Di He. 2024. REST: Retrieval-Based Speculative Decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1582–1595, Mexico City, Mexico.

Jimin Hong, Gibbeum Lee, and Jaewoong Cho. 2024. Accelerating Multilingual Language Model for Excessively Tokenized Languages. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11095–11111.

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

Artur Kiulian, Anton Polishko, Mykola Khandoga, Oryna Chubych, Jack Connor, Raghav Ravishankar, and Adarsh Shirawalmath. 2024. From Bytes to Borsch: Fine-Tuning Gemma and Mistral for the Ukrainian Language Representation. In *Proceedings of the Third Ukrainian Natural Language Processing*

Workshop (UNLP) @ LREC-COLING 2024, pages 83–94, Torino, Italia.

Artur Kiulian, Anton Polishko, Mykola Khandoga, Yevhen Kostiuk, Guillermo Gabrielli, Łukasz Gagała, Fadi Zaraket, Qusai Abu Obaida, Hrishikesh Garud, Wendy Wing Yee Mak, Dmytro Chaplynskyi, Selma Amor, and Grigol Peradze. 2025. From English-Centric to Effective Bilingual: LLMs with Custom Tokenizers for Underrepresented Languages. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 1–13, Vienna, Austria (online).

Taku Kudo and John Richardson. 2018. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast Inference from Transformers via Speculative Decoding. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19274–19286.

Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. 2024. EAGLE: Speculative Sampling Requires Rethinking Feature Uncertainty. In *Proceedings of the 41st International Conference on Machine Learning*, pages 28935–28948.

Daniil Maksymenko and Oleksii Turuta. 2025. Tokenization Efficiency of Current Foundational Large Language Models for the Ukrainian Language. *Frontiers in Artificial Intelligence* 8.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective Initialization of Subword Embeddings for Cross-Lingual Transfer of Monolingual Language Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling Neural Machine Translation to 200 Languages. *Nature* 630(8018):841–846.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. CulturaX: A Cleaned, Enormous, and Multilingual Dataset for Large Language Models in 167 Languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237.

Yurii Paniv, Bohdan Didenko, Mykola Haltiuk, Vladyslav Humennyy, Andrian Kravchenko, Roman Kyslyi, Viktoriia Makovska, Artem Orlovskyi, Bohdan Ruban, Maksym-Yurii Rudko, Anastasiia Senyk, Nazarii Drushchak, Dmytro Chaplynskyi, and Mariana Romanyshyn. 2025. *Lapa LLM v0.1.2 — the most efficient Ukrainian open-source language model*. version 0.1.2.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. Language Model Tokenizers Introduce Unfairness Between Languages. In *Advances in Neural Information Processing Systems*.

QIRI'M Young. 2024. QIRIM Crimean Tatar Monocorpus. Hugging Face Datasets.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135.

Apoorv Saxena. 2023. Prompt Lookup Decoding.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Mitchell Stern, Noam Shazeer, and Jakob Uszkoreit. 2018. Blockwise parallel decoding for deep autoregressive models. *Advances in Neural Information Processing Systems* 31.

Lawrence Stewart, Matthew Trager, Sujan Kumar Gonugondla, and Stefano Soatto. 2024. The N-Grammys: Accelerating Autoregressive Inference with Learning-Free Batched Speculation. *arXiv preprint arXiv:2411.03786*.

Qwen Team. 2026. Qwen3.5: Accelerating Productivity with Native Multimodal Agents.

Susumu Yata. 2023. *MARISA: Matching Algorithm with Recursively Implemented StorAge*.

A Speculative Decoding Visualizations

The following figures visualize speculative decoding outcomes. Each token is color-coded: Accepted draft tokens were proposed by the dictionary and accepted by the target model; Rejected (target corrected) tokens were proposed but replaced by the target model; Bonus (target +1) tokens are additional tokens generated by the target model after accepting a draft; and No draft available tokens had no dictionary match and were decoded autoregressively.

A.1 Ukrainian

Figure 4: Speculative decoding for Ukrainian using Gemma 3 27B. Draft acceptance rate is 42% for this excerpt; speedup is 1.22×

A.2 Crimean Tatar (Cyrillic)

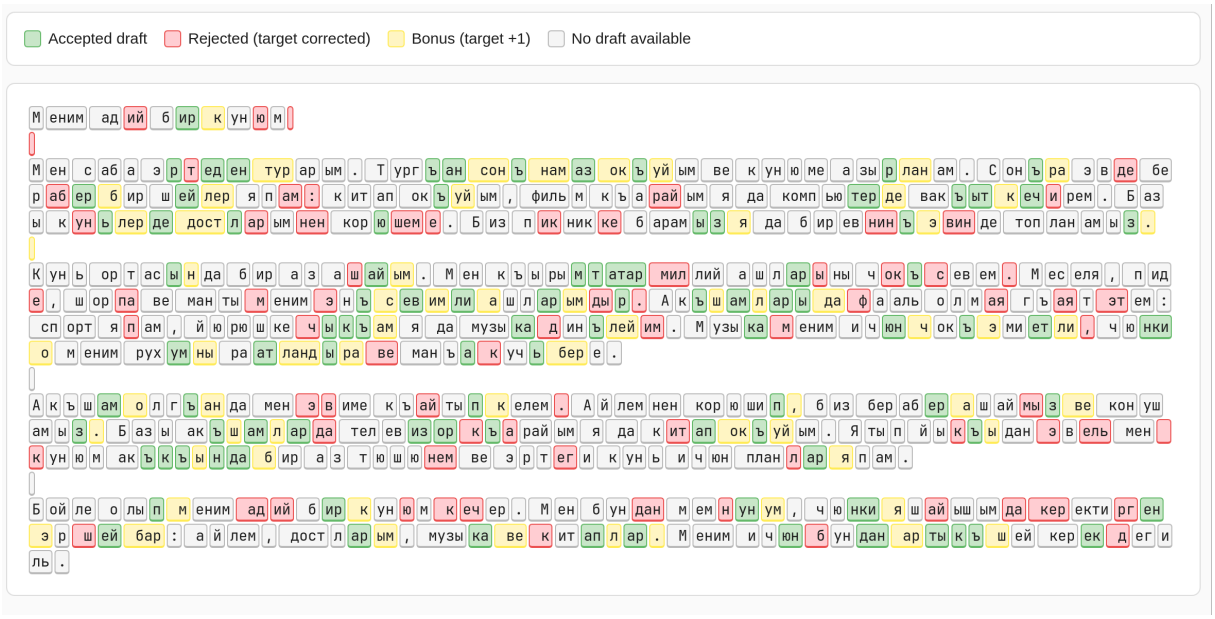


Figure 5: Speculative decoding for Crimean Tatar (Cyrillic script) using Llama 3.3 70B. Draft acceptance rate is 33%; speedup is 1.20×

A.3 Crimean Tatar (Latin)

Figure 6: Speculative decoding for Crimean Tatar (Latin script) using Llama 3.3 70B. Draft acceptance rate is 16%; speedup is 1.09×

B Evaluation Prompts

The following are examples of prompts used in the wall-clock validation experiments (Section 4.6). For each model–language pair, these prompts are sent to the vLLM serving engine as chat-formatted requests, and the model generates up to 2,048 tokens per prompt. The prompts are designed to elicit long-form text in the target language and script.

B.1 Ukrainian

1. Напиши історію України за період 1991–2026 років у десяти параграфах.
2. Напиши короткий текст (3–4 параграфи) про свій звичайний день, використовуючи всі сім відмінків української мови. Після кожного абзацу вкажи, яке слово в якому відмінку вжито.
3. Напиши есе на 1000 слів про значення творчості Лесі Українки для сучасної української ідентичності. Наведи конкретні приклади з її творів.
4. Напиши два варіанти одного й того самого повідомлення: (а) офіційний лист до міського голови Петренко Віктора Івановича з проханням відремонтувати дорогу, і (б) повідомлення другу в месенджері про ту саму проблему. Використовуй відповідний стиль, звертання та лексику для кожного варіанту.

B.2 Crimean Tatar (Cyrillic)

1. Къырымтатар халкъынынъ тарихыны он параграфта язынъыз. Алтын Орда девиринден бугуньге къадар олгъан энъ муим вакъиаларны анълатынъыз.
2. Исмаил Гаспринскийнинъ къырымтатар миллетининъ инкишафындаки ролю акъкъында бинъ сёзлюк бир эссе язынъыз. Онынъ «Терджиман» газетасындан ве дигер фаалиетлеринден мисаллер кетиринъиз.
3. Къырымтатар тилинде сизинъ адий бир кунюнъизни тасвирлеген 3–4 параграфлыкъ бир язы язынъыз.
4. Эки вариантта бир хабер язынъыз: (а) мектеп мудирине расмий бир мектюп ве (б) аркъадашынъызгъа месажда айны меселе акъкъында язув. Эр вариант ичюн мунасип услюп ве сёзлер къулланынъыз.

B.3 Crimean Tatar (Latin)

1. Qırımtatar halqınıñ tarihını on paragrafta yazıñız. Altın Orda devirinden bugünge qadar olğan eñ muim vaqialarnı añlatıñız.
2. İsmail Gasprinskiyniñ qırımtatar milletiniñ inkişafındaki rolü aqqında biñ sözlük bir esse yazıñız. Onıñ «Terciman» gazetasından ve diger faaliyetlerinden misaller ketiriñiz.
3. Qırımtatar tilinde siziñ adiy bir künüñizni tasvirlegen 3–4 paragraflıq bir yazı yazıñız.
4. Eki variantta bir haber yazıñız: (a) mektep müdirine resmiy bir mektüp ve (b) arqadaşıñızğa mesajda aynı mesele aqqında yazuv. Er variant içün münasip üslüp ve sözler qullanıñız.

Scaling ASR for Hutsul Dialect: Multi-Speaker Data Collection, Enhanced Transcription and Cross-Speaker Evaluation

Artem Orlovskyi Zakhar Guzii Bohdan Onyshchenko
Roman Kyslyi Pavlo Khomenko
Kyiv School of Economics, Ukraine
(aorlovsky,zguzii,bonyshchenko,rkyslyi,pkhomenko)@kse.org.ua

Abstract

We present a significant expansion of automatic speech recognition (ASR) resources for the Hutsul dialect of Ukrainian, building on prior work that established the first aligned speech corpus from a single literary source. In this work, we scale the dataset from a single speaker to a multi-speaker corpus comprising 40 speakers and 60.63 hours of audio drawn from diverse sources: YouTube channels (with author permissions), field recordings from native speakers, linguist student recordings, and regional radio broadcasts. To obtain reference transcriptions for audio without existing text, we introduce a novel retrieval-augmented generation (RAG) correction pipeline: audio is first transcribed using ElevenLabs, then corrected through a RAG pipeline backed by a dialect-aware language model. We evaluate fine-tuned ASR models across five distinct speaker datasets, demonstrating that while the best model achieves strong performance on in-domain speakers (character error rate, CER, 3.24%), cross-speaker generalisation remains challenging, with CER ranging from 5.33% to 17.24% depending on speaker characteristics. The cross-speaker gap of 5–14 percentage points indicates that single-speaker-dominated training data is insufficient for robust dialect ASR and motivates future work on speaker-balanced corpora and adaptation methods. All data, code, and models are released publicly to support further research on Ukrainian dialect speech technologies.

1 Introduction

Automatic Speech Recognition (ASR) for low-resource dialects has lagged behind ASR for standard languages, despite the cultural and linguistic value such dialects carry. The Hutsul dialect, spoken in the Ukrainian Carpathians, exhibits phonological, morphological, and lexical features that differ markedly from standard Ukrainian. As Ukrainian ASR systems are increasingly deployed in education, broadcasting, and accessibility tools, their failure on Hutsul speech effectively excludes a sizeable community of speakers. A dedicated dataset is therefore not only a methodological convenience but a prerequisite for equitable language technology in Ukraine: Hutsul speech contains forms that no general-purpose Ukrainian system has been trained to recognise, and absent dialect-targeted resources, errors are systematically biased toward standardisation.

Our previous work (Kyslyi et al., 2026) presented the first dedicated ASR resources for the Hutsul dialect of Ukrainian, centred on a single-speaker corpus derived from readings of the novel "Dido Yvanchyk." That work established a data preparation pipeline, benchmarked multiple ASR architectures, and demonstrated that fine-tuning can reduce Character Error Rate (CER) from over 17% to below 3% on single-speaker dialectal speech.

However, real-world dialect ASR faces challenges that a single-speaker corpus cannot address. The Hutsul dialect exhibits substantial inter-speaker variation: pronunciation, lexical choice, and morphological forms differ across villages, generations, and individual speaking styles. Throughout this paper we use *single-speaker setup* to refer to training and evaluation data drawn from a single human speaker (as in our prior corpus), and *multi-speaker setup* to refer to data drawn from many distinct speakers with diverse demographic, geographic, and stylistic profiles. A model trained on one speaker may fail to generalise to others, limiting practical applicability.

In this paper, we address three key limitations of our prior work:

1. **Speaker diversity:** We expand the corpus from a single speaker to 40 speakers across 60.63 hours of audio, incorporating YouTube content creators, field recordings, university

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 184–198

students, and radio broadcasts.

2. **Transcription for untexted audio:** Unlike the novel-based corpus where ground-truth text existed, our new sources lack reference transcriptions. We develop a RAG-enhanced correction pipeline that combines automatic transcription with dialect-aware language model post-processing.

3. **Cross-speaker evaluation:** We systematically evaluate how a Hutsul-tuned ASR model generalises across speakers with different backgrounds, recording conditions, and dialectal characteristics.

The methodology is dialect-agnostic: the same combination of permissioned audio harvesting, RAG-based dialect correction, and cross-speaker evaluation can be applied to other Ukrainian dialect groups (e.g. Boyko, Lemko, Polissian) wherever a small seed corpus of attested dialect text exists. We view this as a practical recipe for scaling dialect ASR beyond Hutsul.

Our results reveal that while fine-tuned models achieve strong in-domain performance, cross-speaker generalisation remains an open challenge for dialect ASR, motivating further work on multi-speaker training and speaker-adaptive methods.

2 Related Work

2.1 Dialect ASR

Research on ASR for low-resource languages and dialects has expanded significantly, with work in Arabic (Ali et al., 2014), Hindi (Javed et al., 2024), and other language families demonstrating that dialectal variation poses persistent challenges even for large pretrained models. For Ukrainian specifically, existing ASR systems target the standard language (Paniv, 2023; arampacha, 2024), and our previous work (Kyslyi et al., 2026) provided the first Hutsul dialect benchmark.

2.2 Post-Processing and Error Correction for ASR

ASR output for low-resource languages and dialects frequently contains systematic errors that can be addressed through post-processing. Language model rescoring (Radford et al., 2023) and n-best re-ranking are established techniques, but they require language models trained on dialect text, which is scarce for Hutsul.

Recent work has explored using large language models (LLMs) for ASR error correction. The VuykoMistral approach (Kyslyi et al., 2025) demonstrated that a Mistral-based model fine-tuned on Ukrainian dialect text can serve as an effective post-processor for ASR output, correcting morphological and lexical errors while preserving dialect-specific forms. Our RAG-corrector pipeline builds on this idea, using retrieval-augmented generation to ground corrections in attested dialect text.

2.3 Multi-Speaker Dialect Corpora

Building multi-speaker corpora for dialects requires addressing speaker recruitment, consent, recording standardization, and transcription challenges. Prior work on Scottish Gaelic (Klejch et al., 2025) and Arabic dialects (Ali et al., 2014) has documented these challenges. Our approach combines opportunistic data collection (YouTube, radio) with structured elicitation (student recordings, field work).

3 Data Collection and Corpus Expansion

3.1 Overview

Our expanded corpus draws from five distinct source categories, each presenting unique characteristics and challenges. Table 1 summarizes the data sources.

3.2 YouTube Channels

We identified and contacted several Hutsul-language YouTube content creators who produce regular videos featuring dialect speech. After obtaining explicit permission from channel owners, we downloaded and processed audio from two narrative/podcast-style channels (12.55 hours combined: 10.72 hours from one channel and 1.83 hours from another). The content ranges from personal vlogs and storytelling to cultural commentary, providing diverse speaking styles and topics.

Audio was extracted, resampled to 16 kHz, and segmented using our existing pipeline. Unlike the Dido Yvanchyk corpus, these recordings lack reference text, necessitating the RAG-corrector pipeline described in Section 4.

3.3 Yaroslav Zelenchuk Recordings

We collected 4 hours of recordings from Yaroslav Zelenchuk and his father, native Hutsul speakers from the Verkhovyna district (Ivano-Frankivsk region). These recordings include both spontaneous

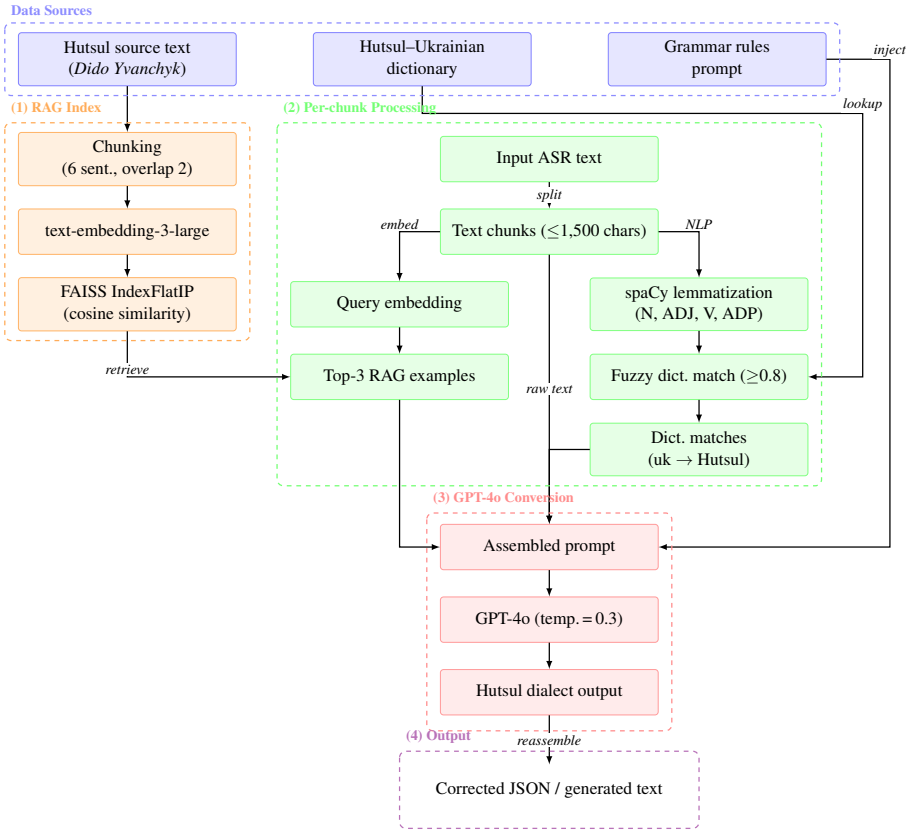


Figure 1: RAG-Corrector pipeline for Hutsul dialect transcription correction. Input ASR transcriptions are split into chunks, simultaneously (a) embedded and matched against a FAISS index of real Hutsul dialect text for retrieval of semantically similar passages, and (b) lemmatised for fuzzy dictionary lookup. Both results and explicit grammar rules are assembled into a GPT-4o prompt (temperature = 0.3) which rewrites the text in authentic Hutsul dialect.

Source	Type	Speakers	Duration	Notes
Dido Yvanchyk	Novel reading	1	15.69h	Updated version of original corpus
Hutsulendia	Broadcast	8	2.89h	Regional Hutsul-dialect programme
Yaroslav recordings	Field recordings	1	4.00h	Native speakers, spontaneous speech
NaUKMA students	Read	28	25.50h	Linguistics students at NaUKMA
YT-channel2	Podcast / narrative	1	10.72h	Hutsul & Bukovynian ethnography
YT-channel1	Podcast / narrative	1	1.83h	Single speaker, used by permission
Total		**40**	**60.63h**	

Table 1: Summary of data sources in the expanded Hutsul dialect corpus.

conversational speech and semi-structured narratives about local history and traditions. Recordings were made with a portable handheld digital recorder using its built-in cardioid microphones at 48 kHz/24-bit, in quiet indoor conditions in the speakers' homes; audio was subsequently downsampled to 16 kHz mono and loudness-normalised before segmentation.

3.4 Linguistics Student Recordings

Linguistics students from the National University of Kyiv-Mohyla Academy (NaUKMA) with Hut-

sul heritage participated in recording sessions. 26 identified speakers, plus an additional 2 distinct but anonymised speaker IDs assigned to segments whose individual identity could not be determined, contributed 25.50 hours of speech for a total of 28 speaker entries (matching Table 1). Recordings include both read passages from Hutsul literary texts and semi-spontaneous speech (e.g., retelling stories, describing their home regions). This source provides younger-generation speakers whose dialect may show more influence from standard Ukrainian, complementing the older speakers in the field-recording and broadcast subsets.

3.5 Hutsulendia Broadcasts

We also obtained 2.89 hours of recordings from the *Hutsulendia* programme, a regional broadcast featuring the Hutsul dialect. The material includes interviews, cultural programmes, and local news segments from 8 distinct speakers. These broadcasts provide professional-quality audio but feature varying degrees of dialect usage, from strong dialect to code-switching with standard Ukrainian. Together with the YouTube and field-recording subsets, this source helps cover speakers from different generations and from a range of villages within the broader Hutsul region.

4 RAG-Enhanced Transcription Pipeline

4.1 Motivation

Our original corpus benefited from an existing written text (the Dido Yvanchyk novel) that could serve as ground truth for alignment. The new data sources - YouTube channels, field recordings, student speech, and radio broadcasts - lack pre-existing transcriptions. Simply transcribing with a standard Ukrainian ASR model produces output riddled with dialect-specific errors: vowel substitutions (y↔i in Ukrainian orthography), morphological normalizations (dialect endings replaced with standard forms), and lexical substitutions (dialect words replaced with standard equivalents).

4.2 Pipeline Architecture

To address this, we developed a two-stage transcription pipeline inspired by the VuykoMistral approach (Kyslyi et al., 2025):

Stage 1: Initial Transcription. Raw audio is transcribed using ElevenLabs STT (ElevenLabs, 2024), which we found in our previous work (Kyslyi et al., 2026) to produce the most reliable word-level timestamps for expressive and dialectal Ukrainian speech. The transcription provides a reasonable first pass but systematically normalizes dialect features toward standard Ukrainian.

Stage 2: RAG-Based Correction. The initial transcription is then processed through a retrieval-augmented generation (RAG) correction pipeline:

1. **Retrieval:** For each transcribed segment, we retrieve the most similar passages from a dialect text knowledge base. This knowledge base includes the full text of the Dido Yvanchyk novel, other Hutsul literary works, and a curated dictionary of Hutsul dialectal forms.

2. **Context construction:** Retrieved passages are combined with the ASR output to form a prompt that provides dialect context.

3. **Generation:** A dialect-aware language model (based on the VuykoMistral architecture) generates a corrected transcription that restores dialect-specific forms while preserving the acoustic content of the original ASR output.

4. **Confidence filtering:** Corrections are accepted only when the model's confidence exceeds a threshold, preventing hallucinated edits. Segments with low correction confidence are flagged for manual review.

4.3 Comparison with Direct ASR

The RAG-corrector pipeline offers several advantages over training a dialect-specific ASR model from scratch for transcription:

- It leverages existing high-quality ASR systems (ElevenLabs) for acoustic modeling while focusing corrections on dialect-specific linguistic features.

- The retrieval component grounds corrections in attested dialect text, reducing hallucination risk.

- The pipeline can be iteratively improved as more dialect text becomes available in the knowledge base.

A detailed evaluation of the RAG-corrector's accuracy on held-out manually transcribed segments is presented in Section 4.4.

4.4 RAG-Corrector Evaluation

A preliminary evaluation by a native Hutsul speaker on a held-out set of 412 manually transcribed segments (drawn from the YouTube and *Hutsulendia* subsets, none of which appear in the RAG knowledge base) shows that the RAG-corrector pipeline improves the proportion of segments rated as fully correct from approximately 40% for raw ElevenLabs output to 71% after dialect-aware correction. The gains are concentrated in dialect-specific lexical items and morphological endings where the baseline STT system defaults to standard Ukrainian forms. A more comprehensive evaluation that disentangles the contribution of retrieval from that of the LLM (i.e. correction with the same dialect-aware LLM but *without* retrieved context) is in progress and is left for future work; this ablation will quantify how much of the observed gain comes specifically from the RAG component versus the underlying LLM's prior knowledge.

5 ASR Models and Training

To benchmark the expanded corpus across architectural families and to make the result directly comparable with our previous single-speaker study (Kyslyi et al., 2026), we evaluate three model families: the encoder–decoder Whisper family (Section 6.1), the CTC-based Wav2Vec 2.0 / XLS-R model (Section 6.3), and the recently released CTC-based Omnilingual ASR models (team et al., 2025) (Section 6.2). Whisper is included because it remains the most widely-deployed open ASR system for Ukrainian and is therefore a natural out-of-the-box baseline for any new dialect; Wav2Vec 2.0/XLS-R is included because it is the strongest pure-CTC baseline previously reported for Ukrainian; and Omnilingual ASR is our flagship system for cross-speaker evaluation, as it achieved the lowest single-speaker CER (2.75%) in our prior work and remains the most accurate model in the multi-speaker setting (Section 6.2).

Building on our previous findings (Kyslyi et al., 2026), our flagship model is OmniASR-CTC-1B. It was fine-tuned on the expanded multi-speaker Hutsul corpus using the same training configuration as our prior work: learning rate 5×10^{-5} with a tri-stage scheduler, per-device batch size of 8 with gradient accumulation of 4, and on-the-fly augmentation including Gaussian noise injection, pitch shifting, speed perturbation (0.8–1.2×), gain modulation, and time/frequency masking (Bartelds et al., 2023).

The fine-tuned model is released as `KSE-RESEARCH-Group` on Hugging Face, together with the expanded `ukr-dialects-audio-dataset` corpus and all training and evaluation scripts (see Section 8).

6 Experimental Results

We evaluate all fine-tuned models across five cross-speaker test sets plus an aggregate set, reporting WER and CER. Full cross-speaker results for individual model families appear in Appendix A; training dynamics are shown in Appendix B.

6.1 Whisper Family

We fine-tuned four Whisper variants: whisper-small, whisper-medium, whisper-large-v3, and arampacha/whisper-large-uk-2 (arampacha, 2024), the latter already adapted to standard Ukrainian via Common Voice 11.0 (Ardila et al., 2020). Fine-tuned checkpoints are publicly available on Hugging Face under the `KSE-RESEARCH-Group` organisation (Section 8). Training was performed in FP16 mixed-precision mode with CER as the primary selection criterion, as character-level evaluation captures dialectal orthographic variation more precisely than WER for the highly inflected morphology of Ukrainian (Thennal D K et al., 2025).

6.1.1 Training Setup

All models were trained using the Hugging Face Seq2SeqTrainer with the AdamW optimizer, a peak learning rate of 1×10^{-5}, 500 linear warm-up steps, and linear decay thereafter. Following the standard Whisper training objective (Radford et al., 2023), models were optimised using cross-entropy loss over decoder token predictions, with padding positions excluded from the loss computation. Checkpoints were saved and evaluated every 500 steps; the checkpoint with the lowest validation CER was retained for final evaluation. Training dynamics for all Whisper variants are shown in the left column of Figure 2 (top: WER, middle: CER, bottom: train/eval loss).

The three generic checkpoints (whisper-small, whisper-medium, whisper-large-v3) were trained for 8,000 steps on the same split of the Ukrainian dialect corpus (27,518 train / 3,347 validation / 3,435 test utterances, maximum token length 448). Per-device batch sizes varied by model due to memory constraints: whisper-large-v3 used batch size 4

with gradient accumulation 2 (effective 8), whisper-medium used batch size 8 with gradient accumulation 4 (effective 32), and whisper-small used batch size 4 with gradient accumulation 4 (effective 16) in Phase 1. The Ukrainian-adapted checkpoint (arampacha/whisper-large-uk-2) was trained for 4,000 steps with batch size 16 and gradient accumulation 2 (effective 32), with gradient checkpointing enabled; its prior Ukrainian fine-tuning yields a stronger initialisation that requires less dialect adaptation.

To reduce hallucination artefacts common in Whisper on expressive speech (Radford et al., 2023), whisper-small was trained in two phases. Phase 1 (steps 1–4,000) used standard training without suppression. Phase 2 (steps 4,001–8,000) introduced no_repeat_ngram_size=3, repetition_penalty=1.3, and condition_on_previous_text=False, while increasing the per-device batch size from 4 to 16 (effective batch size 64). The boundary between phases is marked by the red dashed vertical line at step 4,000 in Figure 2. The two larger generic models (whisper-medium, whisper-large-v3) applied repetition suppression throughout training. The Ukrainian-adapted checkpoint was trained without suppression, as the pre-adapted weights exhibited lower baseline hallucination rates on Ukrainian text. All runs used an NVIDIA GeForce RTX 4090 (48GB VRAM).

The training-step budgets (8,000 for the three generic Whisper checkpoints and 4,000 for the Ukrainian-adapted checkpoint) were chosen empirically from the training dynamics in Figure 2: in each run the validation CER plateaus and starts to drift while training loss continues to decrease, which is the standard sign that further optimisation produces overfitting rather than additional gain. Running beyond 8,000 steps for the larger Whisper variants did not improve validation CER in pilot experiments. The Ukrainian-adapted checkpoint reaches its best validation CER as early as step 3,000 and is therefore trained for only 4,000 steps. We retain the checkpoint with the lowest validation CER for evaluation, which functions as an early-stopping rule.

6.1.2 Results

In-domain test results on the Hutsul corpus are shown in Table 4. whisper-large-v3 achieves the lowest aggregate CER among all Whisper variants (9.32%, WER 25.18%) at step 8,000, followed by whisper-medium (CER 9.99%) and whisper-small (CER 12.54%). The Ukrainian-adapted checkpoint (arampacha/whisper-large-uk-2) attains aggregate CER 10.67% (WER 29.37%), reaching its best checkpoint at step 3,000—the earliest of any variant—which demonstrates the efficiency of dialect adaptation when starting from a language-specific prior, despite not surpassing whisper-large-v3 in final accuracy.

Evaluated across the broader six-dataset Ukrainian dialect benchmark (Table 4), whisper-large-v3 achieves the lowest macro-average WER (25.71%), followed by whisper-medium (28.31%), the Ukrainian-adapted whisper-large-uk-2 (29.82%), and whisper-small (35.79%), with CER following the same order. The Hutsulendia dataset consistently yields the highest WER across all variants (32.01%–45.85%), while near-standard southwestern varieties (YT-channel2) achieve the lowest (16.37%–28.35%).

6.1.3 Analysis of Results

Among the three generic checkpoints, CER decreases monotonically with model scale (Small → Medium → Large-v3), consistent with larger Whisper encoders capturing finer-grained phonetic distinctions (Radford et al., 2023). Notably, arampacha/whisper-large-uk-2 achieves comparable per-dataset CER to whisper-large-v3 on several evaluation sets while using only half the training steps (4,000 vs. 8,000), confirming that prior exposure to standard Ukrainian provides a more efficient starting point for dialect adaptation. Larger models also converge faster: whisper-large-v3 drops below 12% validation CER by step 3,000, whereas whisper-small never reaches this level (Figure 2).

The two-phase training strategy for whisper-small – introducing repetition suppression in Phase 2—yielded consistent improvements across all evaluation datasets, with the largest gain on Dido-Yvanchyk (ΔWER -4.08 pp). The Hutsulendia dataset consistently yields the highest WER across all variants (6–10 pp above the macro-average), driven by spontaneous multi-speaker speech with dense Hutsul-specific vocabulary. Character-level error analysis reveals vowel confusion—"i/и" merger and "o/y" shift—as the dominant failure mode, with morphological suffix errors accounting for the largest share of total character-level errors across all variants.

Dialect-specific vocabulary analysis exposes two

distinct recognition regimes. High-frequency Hutsul lexemes with phonological proximity to standard Ukrainian are handled reliably: “тогди” (*tohdy*, F1=0.90), “файно” (*fayno*, F1=0.86), and “тимунь” (*tymun’*, F1=0.93) are correctly transcribed in the overwhelming majority of occurrences. By contrast, words whose surface form diverges substantially from any standard Ukrainian equivalent suffer systematic misrecognition: “відки” (*vidky*, F1=0.22) is consistently mapped to “звідки” (*zvidky*), “бірше” (*birshe*, F1=0.40) to “більше” (*bil’she*), and “таки” (*taky*, F1=0.83) alternates between “так” (*tak*) and “та” (*ta*). This pattern reveals that the model systematically normalises dialectal forms toward their standard Ukrainian counterparts rather than producing random errors, suggesting that the decoder’s language prior–inherited from multilingual pretraining–actively suppresses out-of-vocabulary dialect forms in favour of the nearest phonologically similar standard word.

6.2 Omnilingual ASR

6.2.1 Training Configuration and Results

We fine-tuned two OmniASR CTC models released by Meta (300M and 1B parameters). Training was performed using the official (team et al., 2025) tutorial on the "ukr-dialects-audio-dataset". Both models were trained with a learning rate of 5×10^{-5} using a tri-stage scheduler. The 300M model was trained on an RTX 5070 Ti (16GB) for 60k steps, while the 1B model was trained on an RTX 4090 (48GB) for 42k steps. We used WER as the primary optimization metric and CTC as loss function. Fine-tuning successfully reduced the initial WER from 76–80% down to around 28.5% and CER from 37–51% to around 11.5%, demonstrating strong dialect adaptation. Detailed cross-speaker evaluation metrics are presented in Table 3 (Appendix A).

6.2.2 Lexical and Morphological Analysis

Evaluating the 1B model on 3,451 utterances revealed a high lexical recall (85.5%) for dictionary-verified Hutsul terms, with only 3.0% of dialectal words dropped entirely. The primary failure mode is orthographic regularization driven by phonological interference (e.g., frequent shifts between "і"(і)↔"и"(у) and "о"(о)↔"у"(u)).

The model tends to overwrite unique Hutsul lexical endings with standard Ukrainian equivalents, resulting in most of the character errors in the suffix zone. Consequently, while standard-adjacent dialect words ("дуже" (duzhe), "файно" (faino)) are robustly transcribed, highly localized lexemes suffer aggressive replacement (e.g., "арідник" (aridnyk) regularized to "нарідник" (naridnyk), and "відтів" (vidtiv) replaced by "вітів" (vitiv)).

6.3 Wav2Vec2

6.3.1 Training Configuration and Results

We fine-tuned the Wav2Vec 2.0 model (XLSR-53 architecture, pre-trained for Ukrainian — `w2v-xls-r-uk`) on the “ukr-dialects-audio-dataset” dataset. Training was performed for 50 epochs on RTX 4090 48 GB.

The model was trained with a learning rate of $1e^{-4}$ using a linear scheduler with 1,000 warmup steps. The per-device batch size was set to 8 with gradient accumulation over 8 steps, yielding an effective total batch size of 64. CER was used as the primary optimization metric.

Training over 21,000 steps achieved stable convergence. Evaluated on the test set of 3,451 utterances, the model reached a final WER of 28.38% and CER of 9.93%.

6.3.2 Lexical and Morphological Analysis

Analysis of the recognition outputs reveals that Wav2Vec2 exhibits error tendency toward orthographic regularization. The model successfully transcribes common vocabulary and dialectal forms that are phonetically close to the literary norm, yet struggles at the morphological level.

The most pronounced failure mode remains the **suffix bottleneck**, whereby distinctive Hutsul endings are systematically replaced with their standard Ukrainian equivalents. The elevated CER (nearly 10%) is driven primarily by vowel substitutions in stems and inflections (e.g., frequent shifts between “і” (і) ↔ “и” (у) and “о” (о) ↔ “у” (u)). The most consistent pattern is the replacement of the dialectal reflexive postfix “-си” with the standard “-ся”, as well as the shift of the particle “си” to “це”/“се”. These substitutions occur in nearly every relevant context, contributing substantially to the overall CER. Representative examples are given below:

1. Postfix “-си” → “-ся”: REF “ни боїмоси” → HYP “не боїмося” (*ny boyimo-**sy** / ne boyimo-**sia***; gloss: “we are not afraid”)

2. Postfix “-си” → “-ся”: REF “вирваласи з хати” → HYP “вирвалася з хати” (*vyrvala-**sy** z khaty / vyrvala-**sia** z khaty*; gloss: “[she] tore [herself] out of the house”)

3. Particle “си” → “се”/“це”: REF “шош си було” → HYP “що ж се було” (*shosh **sy** bulo* / *shcho zh **se** bulo*; gloss: “something **was/this was**”)

4. Hard “-т” → soft “-ть”: REF “чорт и сам любит” → HYP “почорт і сам любить” (*chort y sam liubyt* / *pochort i sam liubyt’*; gloss: “the devil himself loves [it]”)

Dataset	Samples	WER	CER
Dido-Yvanchyk	842	25.80	5.98
YT-channel2	482	20.13	4.65
YT-channel1	103	22.36	5.44
NaUKMA students	1,806	33.88	15.27
Hutsulendia	217	40.31	15.55

Table 2: Cross-speaker evaluation of Wav2Vec2 fine-tuned on Hutsul data

Qualitative analysis of Wav2Vec 2.0 predictions reveals additional error patterns consistent with known Hutsul phonological properties. Dialect-specific vocabulary without standard equivalents (e.g., “лягри”, “кептар”, “сардак”) is frequently dropped or distorted. Word boundary segmentation errors — producing agglutinated forms absent from both standard and dialect — reflect the CTC architecture’s lack of explicit language model priors.

Cross-speaker results (Table 2) reveal a clear performance gradient: in-domain CER of 5.98% degrades to 15.27% on NaUKMA students and 15.55% on Hutsulendia radio broadcasts, consistent with the pattern observed across all model families. Among dictionary-verified dialect-specific lexemes, the model achieves high recall on frequent items but fails systematically on low-frequency, phonologically opaque forms, indicating **lexical regularisation** toward standard Ukrainian for rarer dialect-exclusive vocabulary.

7 Discussion

7.1 The Multi-Speaker Challenge

Our cross-speaker evaluation confirms a well-known but under-documented challenge in dialect ASR: models trained predominantly on one speaker’s data generalize poorly to other speakers of the same dialect. The 5× CER gap between in-domain (3.24%) and the most challenging out-of-domain evaluation (17.24%) underscores the need for diverse training data.

Several factors contribute to this gap:

- **Phonetic variation:** Individual speakers realize dialect phonemes differently, and the model may overfit to one speaker’s particular realizations.
- **Lexical diversity:** Different speakers draw on different subsets of the Hutsul lexicon, and some may code-switch more frequently with standard Ukrainian.
- **Recording conditions:** YouTube, radio, field recordings, and studio recordings differ in noise levels, microphone quality, and acoustic environments.
- **Speaking style:** Read speech (Dido Yvanchyk) differs substantially from spontaneous speech (Yaroslav, students) and broadcast speech (radio).

7.2 The RAG-Corrector Pipeline

The RAG-enhanced transcription pipeline addresses a fundamental bottleneck in scaling dialect ASR: the lack of reference transcriptions for new audio sources. By combining high-quality acoustic transcription (ElevenLabs) with dialect-aware linguistic correction (RAG + VuykoMistral), we can produce training data for speakers and sources that lack existing text.

However, the pipeline is not without limitations. The quality of corrections depends heavily on the coverage of the dialect knowledge base, and the model may struggle with dialectal features not represented in the retrieval corpus. Ongoing manual verification of a subset of corrected transcriptions will be essential for quality assurance.

7.3 Lexical Regularisation as a Structural Limitation

A pattern we observe across all three model families is that errors are not random: dialect-specific lexemes and morphological endings are systematically “corrected” to standard Ukrainian forms (e.g. the dialectal reflexive postfix - replaced by standard -, the particle replaced by /, and rare lexemes such as mapped to phonologically similar standard forms). This indicates that the decoder’s language prior, inherited from large-scale pre-training on standard text, dominates over the acoustic evidence whenever a dialect form has a phonologically close standard counterpart. Fine-tuning on a few

tens of hours of dialect data is not sufficient to overwrite this prior. We read this as a structural limitation rather than a data-quantity problem: a fully Hutsul-aware system likely requires either (i) a foundation acoustic model pre-trained on a large Hutsul-text language model, (ii) a constrained decoding scheme that explicitly biases toward attested dialect forms, or (iii) joint training with a dialect language-model loss. The current RAG-corrector mitigates the symptom at the text level after transcription but does not remove the underlying bias of the acoustic model.

7.4 Generational and Geographic Variation

The Hutsul dialect is not uniform: pronunciation, vocabulary and morphology vary across villages, between mountain and lowland communities, and between older and younger speakers. Our corpus partially reflects this. NaUKMA students contribute younger-generation speech that is closer to standard Ukrainian; field recordings from Verkhovyna contribute older-generation speakers whose dialect is conservative; the *Hutsulendia* broadcast subset and the YouTube channels span both demographics and several villages. We do not yet annotate village-level provenance for every speaker, which is a clear limitation for fine-grained sociolinguistic analysis. The cross-speaker results in Tables 3 and 4 can be read in this light: the largest CER values are concentrated in the subsets that mix several villages and generations (NaUKMA students, Hutsulendia), while subsets dominated by a single speaker or speech style yield substantially lower CER.

7.5 Implications for Low-Resource Dialect ASR

Our findings have broader implications for the dialect ASR community:

- **Single-speaker corpora are necessary but insufficient.** They provide a strong starting point for model development but do not guarantee cross-speaker generalisation.
- **Diverse data collection is essential.** Combining multiple sources (literary readings, field recordings, student speech, broadcasts) captures the full range of dialectal variation.
- **LLM-based post-processing can bootstrap transcription.** For dialects without written tradition, combining ASR with dialect-aware LLMs offers a scalable path to labelled data.

8 Released Resources

To support reproducibility and follow-up work, we publicly release:

- the expanded `ukr-dialects-audio-dataset` multi-speaker Hutsul corpus (40 speakers, 60.63 hours), with train/validation/test splits and per-utterance metadata;
- the fine-tuned models at `KSE-RESEARCH-Group`, the four fine-tuned Whisper variants, and the Wav2Vec2 (XLS-R) checkpoint;
- the RAG-corrector pipeline code (chunking, FAISS index, fuzzy dictionary lookup, prompt assembly), together with the Hutsul–Ukrainian dictionary used for retrieval; and
- the training and evaluation scripts used to reproduce all reported numbers.

All artefacts are hosted on Hugging Face under the `KSE-RESEARCH-Group` organisation and on the project GitHub repository.[1]

9 Conclusion

We present a substantial expansion of Hutsul dialect ASR resources, growing from a single-speaker literary corpus to a multi-speaker dataset comprising 40 speakers and 60.63 hours of audio from diverse sources. We introduce a RAG-enhanced transcription pipeline that enables scalable creation of training data for audio without existing reference text. Our cross-speaker evaluation reveals both the promise and limitations of current approaches: fine-tuned models achieve strong in-domain performance (3.24% CER) but face significant degradation on out-of-domain speakers (up to 17.24% CER).

These results motivate several directions for future work: multi-speaker training with balanced speaker representation, speaker adaptation techniques, improved RAG-corrector pipelines with expanded dialect knowledge bases, and exploration of LM rescoring for CTC-based dialect ASR. All

[1] Hugging Face: https://huggingface.co/KSE-RESEARCH-Group. Code: https://github.com/KSE-RESEARCH-Group.

data, models, and code are publicly released to support continued research on Ukrainian dialect speech technologies.

Limitations

This work has several limitations. First, our cross-speaker evaluation uses a model primarily trained on single-speaker data; results from a model trained on the full multi-speaker corpus may differ substantially. Second, the RAG-corrector pipeline has not yet been comprehensively evaluated with human judgments—we plan to conduct this evaluation as annotation is completed. Third, the data quantities for some speakers (e.g., YT-channel1 with 103 samples) are small, making evaluation estimates noisy. Fourth, we do not yet evaluate the impact of including RAG-corrected transcriptions in training data. Finally, our evaluation is limited to WER and CER; perceptual quality assessments and downstream task evaluations would provide a more complete picture.

Acknowledgments

We thank the YouTube content creators who generously granted permission to use their recordings for research purposes (Larysa Irodenko and Ivanna Stefiuk), the linguistics students at the National University of Kyiv-Mohyla Academy (NaUKMA) who participated in recording sessions, and Yaroslav Zelenchuk for gathering and providing native-speaker recordings from the Verkhovyna region. We also thank the anonymous reviewers of UNLP 2026 for constructive feedback that improved the camera-ready version of this paper.

Use of generative AI. Large language models (specifically OpenAI GPT-4o and Anthropic Claude) were used (i) inside the RAG-corrector pipeline as described in Section 4 (this is part of the methodological contribution), and (ii) to assist with copy-editing of the manuscript: smoothing English prose, harmonising terminology, and checking for typographical errors. All technical content, experimental design, results, and interpretations are the responsibility of the authors, who reviewed and edited every passage produced or revised with LLM assistance.

References

Ahmed Ali, Hamdy Mubarak, and Stephan Vogel. 2014. Advances in dialectal Arabic speech recognition: A study using Twitter to improve Egyptian ASR. In *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*.

arampacha. 2024. arampacha/whisper-large-uk-2: Whisper-Large fine-tuned for Ukrainian speech recognition. Hugging Face model. https://huggingface.co/arampacha/whisper-large-uk-2.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 4218–4222. European Language Resources Association. Cited version is Corpus 11.0 (2022 release).

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL, Long Papers)*.

ElevenLabs. 2024. Elevenlabs speech-to-text API documentation. https://elevenlabs.io/docs/api/speech-to-text.

Tahir Javed, Janki Nawale, Eldho Joshi, Kaushal Bhogale, Sumanth Doddapaneni, Anoop Kunchukuttan, Pushpak Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2024. LAHAJA: A robust multi-accent benchmark for evaluating Hindi ASR systems. *Preprint*, arXiv:2408.11440.

Ondřej Klejch, William Lamb, and Peter Bell. 2025. A practitioner's guide to building ASR models for low-resource languages: A case study on Scottish Gaelic. In *Proc. Interspeech 2025*.

Roman Kyslyi, Yulia Maksymiuk, and Ihor Pysmennyy. 2025. Vuyko Mistral: Adapting LLMs for low-resource dialectal translation. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP)*. Association for Computational Linguistics.

Roman Kyslyi, Artem Orlovskyi, Pavlo Khomenko, Bohdan Onyshchenko, and Zakhar Guzii. 2026. Building ASR resources for the Hutsul dialect of Ukrainian. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Association for Computational Linguistics.

Yurii Paniv. 2023. Ukrainian-TTS: An open text-to-speech system for Ukrainian. GitHub repository.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

Omnilingual ASR team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, and 1 others. 2025. Omnilingual ASR: Open-source multilingual speech recognition for 1600+ languages. *Preprint*, arXiv:2511.09690.

Thennal D K, Jesin James, Deepa P. Gopinath, and Muhammed Ashraf K. 2025. Advocating character error rate for multilingual ASR evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*.

A Detailed Evaluation Tables

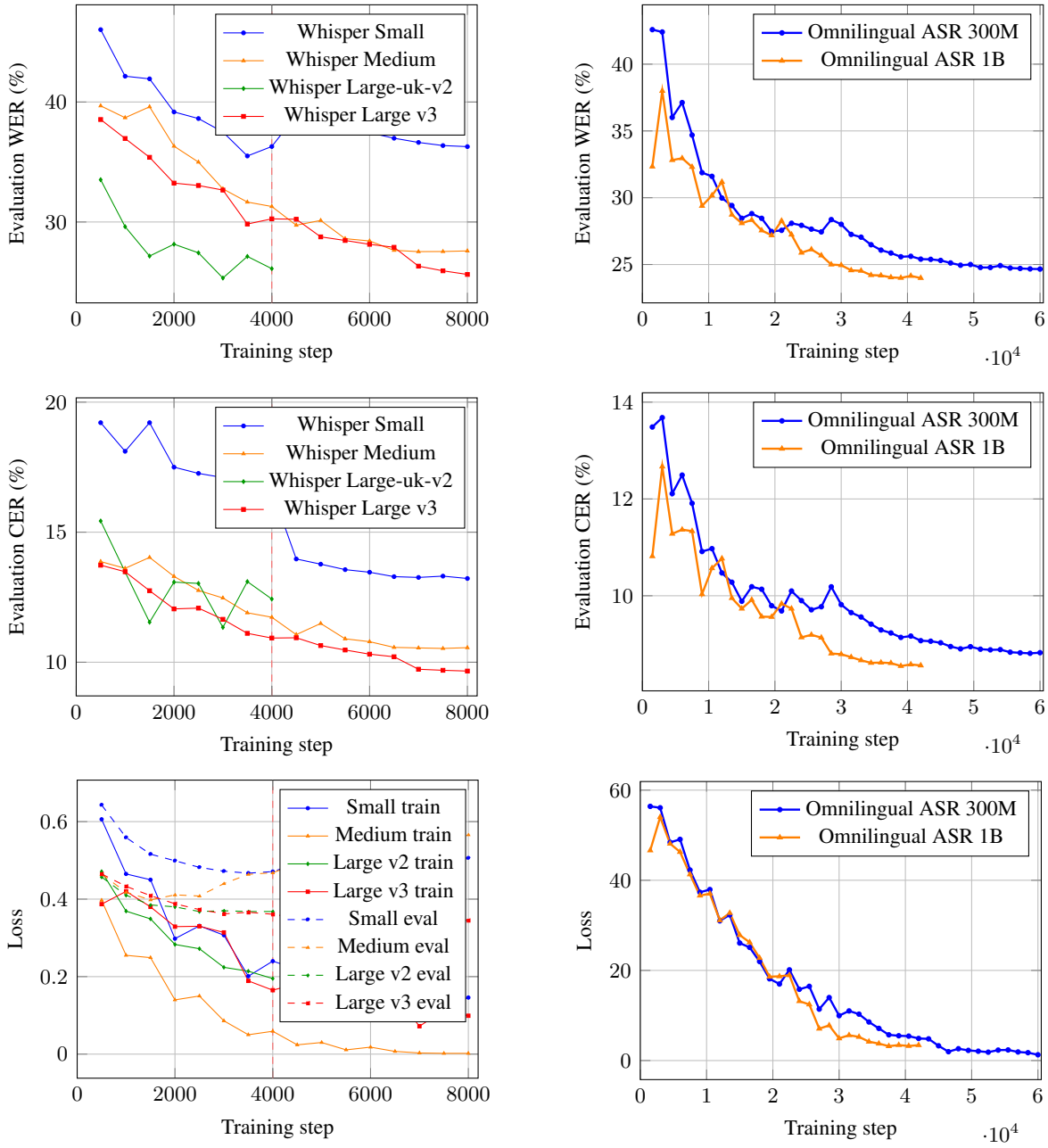
Dataset	Model	Samples	WER	CER
Dido-Yvanchyk	Omnilingual ASR 300M	842	**13.93**	**3.19**
	Omnilingual ASR 1B	842	14.07	3.24
YT-channel2	Omnilingual ASR 300M	483	21.94	5.75
	Omnilingual ASR 1B	483	**20.37**	**5.33**
YT-channel1	Omnilingual ASR 300M	103	**23.41**	6.38
	Omnilingual ASR 1B	103	24.10	**6.19**
NaUKMA students	Omnilingual ASR 300M	1,806	**36.79**	17.21
	Omnilingual ASR 1B	1,806	**36.79**	**16.85**
Hutsulendia (radio)	Omnilingual ASR 300M	217	43.90	**17.13**
	Omnilingual ASR 1B	217	**43.63**	17.24
ukr-dialects-audio-dataset	Omnilingual ASR 300M	3,451	28.99	11.72
	Omnilingual ASR 1B	3,451	**28.76**	**11.49**

Table 3: Cross-speaker evaluation of OmniASR-CTC models fine-tuned on Hutsul data. WER/CER reported as percentages. Bold marks the best result per dataset per metric. Lower is better.

Dataset	Model	Samples	WER	CER
Dido-Yvanchyk	Whisper Small	842	29.88	7.57
	Whisper Medium	842	**19.83**	5.19
	Whisper Large-uk-v2	842	23.50	6.05
	Whisper Large v3	842	19.92	**5.13**
YT-channel2	Whisper Small	483	28.35	7.65
	Whisper Medium	483	19.49	5.07
	Whisper Large-uk-v2	483	19.86	5.22
	Whisper Large v3	483	**16.37**	**4.26**
YT-channel1	Whisper Small	103	34.64	9.56
	Whisper Medium	103	26.82	7.81
	Whisper Large-uk-v2	103	25.54	7.29
	Whisper Large v3	103	**23.18**	**6.23**
NaUKMA students	Whisper Small	1,806	41.86	18.76
	Whisper Medium	1,806	35.00	16.22
	Whisper Large-uk-v2	1,806	37.54	17.08
	Whisper Large v3	1,806	**32.42**	**15.23**
Hutsulendia (radio)	Whisper Small	217	45.85	19.28
	Whisper Medium	217	35.83	14.58
	Whisper Large-uk-v2	217	36.50	15.00
	Whisper Large v3	217	**32.01**	**12.83**
ukr-dialects-audio-dataset	Whisper Small	3,451	35.41	12.54
	Whisper Medium	3,451	26.96	9.99
	Whisper Large-uk-v2	3,451	29.37	10.67
	Whisper Large v3	3,451	**25.18**	**9.32**

Table 4: Cross-speaker evaluation of the Whisper model family fine-tuned on Hutsul data. WER/CER reported as percentages. Bold marks the best result per dataset per metric. Lower is better.

B Training Dynamics



Whisper models: WER, CER, and train/eval loss.

Omnilingual ASR models: WER, CER, and loss.

Figure 2: Training dynamics in a six-plot grid. Left column: Whisper Small/Medium/Large-uk-v2/Large-v3 curves on `ukr-dialects-audio-dataset` (top: WER, middle: CER, bottom: train / eval loss). The red dashed vertical line at step 4000 marks where Whisper Small resumed Phase 2 training. Right column: Omnilingual ASR 300M vs 1B curves (top: WER, middle: CER, bottom: loss). "Large-uk-v2" denotes the Ukrainian-adapted checkpoint `arampacha/whisper-large-uk-2`.

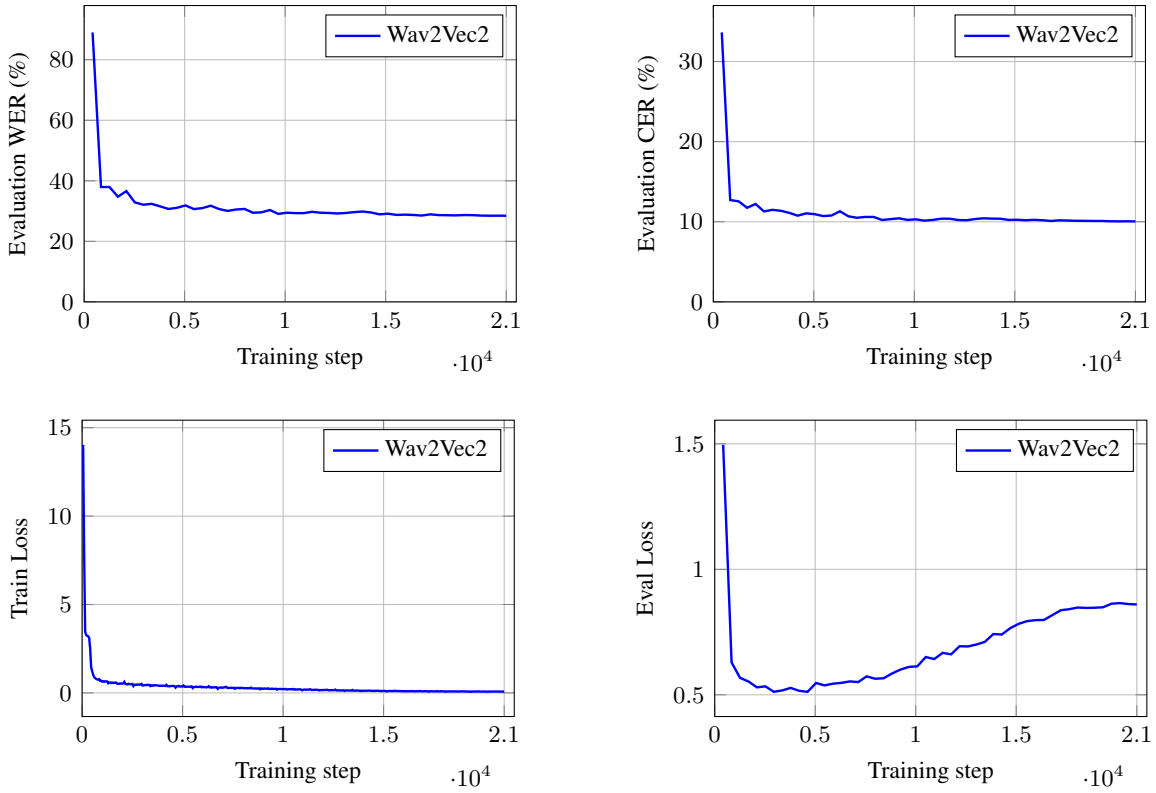


Figure 3: Training dynamics of Wav2Vec2 fine-tuned on the Hutsul dialect dataset over 50 epochs (21,000 steps). Top row: evaluation WER (%) and CER (%). Bottom row: train loss and eval loss. Best checkpoint at step 21,000 (WER = 28.45%, CER = 10.04%). Note: eval loss diverges monotonically from step 2,520 onward while WER/CER continue improving, consistent with the behaviour observed in Whisper models on this task.

C Worst-Case Transcription Examples

Table 5 presents one representative example per model for the two error extremes: the highest-WER utterance (**Worst WER**) and the lowest-CER utterance from the 50-worst list (**Lowest CER**). **REF** = Hutsul dialect reference; **HYP** = model hypothesis.

Model	Type	Dataset	WER	CER	REF	HYP
Whisper Small	Worst WER	Hutsulendia	1.00	1.00	А не і співала, гуляла, ішла, куди мене вела.	А не лізть павалом, ну я гала це і йшла мкуда ми виділи.
	Lowest CER	Dido-Yv.	0.60	0.04	Прото — пустий пусте робит.	Прото, пусти й пусте робит.
Whisper Medium	Worst WER	Hutsulendia	0.80	1.00	А п'ятниця до Пречистої Діви.	А п'єкні це до причищення діли.
	Lowest CER	Dido-Yv.	0.40	0.00	Протерав чоло, так єкби си лише шо з сну прошумав.	Протерав чоло, так єк би си лишешо з сну прошумав.
Whisper Large-uk-v2	Worst WER	Hutsulendia	1.00	0.46	А не і співала, гуляла, ішла, куди мене вела.	А не відставала, бо я і йшла, куда б не виділа.
	Lowest CER	Dido-Yv.	0.43	0.00	Єк флуд пер за дичінов.	Єк флуд перза дичінов.
Whisper Large-v3	Worst WER	Hutsulendia	1.00	0.94	Бути написи, бо я знаю, що- Є ще написи.	- бути, напис робити,- Я знаю, що коли- Ходішь, там є ще знати.
	Lowest CER	Dido-Yv.	0.43	0.00	Зробив навіть и стів новомодний у хаті.	Зробив навіть истівново модний у хаті.
Wav2Vec2	Worst WER	Hutsulendia	1.00	1.00	бо є таке що людина десь погано став	поє таке об одіна рдиєся по панастатаацятттт
	Lowest CER	YT-channel1	0.20	0.00	русалка то та сама лісна котра лиш у воді жила	русалка тота сама лісна котра лиш у воді жила
OmniASR 300M	Worst WER	Hutsulendia	1.00	1.00	рано бо це був вечір	т зра рамон бо це був запобеч вни віів
	Lowest CER	Dido-Yv.	0.80	0.00	то прото були нипрості люде	топрото були ни прості люде
OmniASR 1B	Worst WER	Hutsulendia	1.00	0.94	рано бо це був вечір	ра рамо бо це був побубечи вахоув
	Lowest CER	Dido-Yv.	0.29	0.00	тимунь він цілу ніч мавси на осторозі	тимунь він цілу ніч мавси наосторозі

Table 5: Representative worst-case transcription examples across all models. For each model: Worst WER = utterance with highest WER; Lowest CER = lowest CER from the 50-worst list.

Mining Native Ukrainian Paraphrases: A Multi-Source Comparison

Vladyslav Fesenko, Hanna Dydyk-Meush, Volodymyr Mudryi
Ukrainian Catholic University, Lviv, Ukraine
{fesenko.pn, hanna_dydykmeush, mudryi.pn}@ucu.edu.ua

Abstract

We introduce a Ukrainian paraphrase dataset mined from event-aligned news headlines and compare it with translated and LLM-generated data sources. Candidate pairs are retrieved from native Ukrainian news titles and filtered using semantic and lexical constraints to form a training corpus in a semi-automatic pipeline. Human evaluation indicates that the sources differ in useful ways: LLM-generated paraphrases are generally stronger in meaning preservation, whereas news-mined pairs offer greater lexical variation while remaining fluent and meaning-preserving. We tune mT5-large and mT0-large and evaluate them on several held-out test sets, including a human-validated subset. Relative to Spivavtor-large, the models achieve comparable semantic preservation with lower copying on the combined and human-validated sets. Overall, the findings highlight the value of naturally mined Ukrainian paraphrases as supervision for low-resource paraphrase generation.

1 Introduction

Paraphrase resources serve two distinct roles in NLP: they supervise generation models and they test whether systems can preserve meaning under non-trivial reformulation. The strongest datasets for these purposes remain concentrated in English, including news-derived corpora such as MRPC (Dolan and Brockett, 2005), question-duplicate collections such as QQP and WikiAnswers (DataCanary et al., 2017; Fader et al., 2013), and challenge sets such as PAWS (Zhang et al., 2019). Ukrainian does not yet have a comparable resource designed specifically for paraphrase generation and evaluation.

The missing piece is not only scale but also a systematic comparison of native and synthetic paraphrase sources for Ukrainian. In low-resource settings, paraphrase supervision is often created by translating English corpora or prompting Large Language Models (LLMs) to rewrite text. Both strategies are useful, but they introduce different biases. Translation-based supervision scales cheaply, yet translated text often shows translationese effects and lower linguistic richness than native text (Zhang and Toral, 2019; Vanmassenhove et al., 2021). LLM-based rewriting avoids direct translation transfer, but unconstrained decoding in paraphrase generation can produce trivial copies or near copies, so explicit diversity controls are needed to obtain stronger lexical and structural variation (Thompson and Post, 2020; Holtzman et al., 2019). The result is a familiar trade-off between semantic reliability, naturalness, and lexical change.

We address this problem by mining paraphrases from independent Ukrainian news reports about the same event, which provides semantic anchoring. This idea follows the intuition behind naturally occurring paraphrase resources while adapting it to a low-resource language setting in which native supervision is scarce.

Following recent work on paraphrastic robustness (Srikanth et al., 2024), we treat semantic equivalence and variation in wording and structure as separate properties rather than collapsing them into a single similarity score. This distinction is especially important for Ukrainian, where prior robustness studies show that even small lexical substitutions can substantially affect model behavior and semantic fidelity (Mudryi and Ignatenko, 2025; Mudryi and Laba, 2025). This perspective motivates both our dataset construction choices and our evaluation setup.

Our contributions are:

- We construct a Ukrainian paraphrase corpus by mining news titles about the same event and filtering candidates with semantic and lexical constraints.
- We compare news-mined pairs against translated and LLM-generated data using human

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 199–208

judgments of meaning preservation, fluency, and lexical divergence.

- We evaluate adapter-tuned models trained on a balanced multi-source corpus and show competitive performance against Spivavtor baselines, with lower copying on the combined and human-validated sets.

2 Related Work

Paraphrase datasets differ mainly in how they obtain semantically aligned texts. Some rely on naturally occurring comparable descriptions, such as parallel news reports in MRPC (Dolan and Brockett, 2005), tweets linked to the same URL in Twitter URL (Lan et al., 2017), multiple captions for the same image in MS COCO, or duplicate questions in WikiAnswers and QQP (Lin et al., 2014; Fader et al., 2013; DataCanary et al., 2017). Across these settings, paraphrases are most natural when they arise from independent descriptions of the same underlying event, scene, or intent.

Other resources are constructed synthetically. PAWS, for example, creates high-overlap paraphrase and non-paraphrase pairs through word swapping and back-translation, followed by human validation (Zhang et al., 2019). In low-resource settings, similar synthetic strategies often take the form of translated English corpora or LLM-generated rewrites. These approaches scale well, but translated text may exhibit translationese and reduced linguistic richness (Zhang and Toral, 2019; Vanmassenhove et al., 2021), while unconstrained neural generation often produces conservative rewrites unless diversity is explicitly encouraged (Holtzman et al., 2019; Thompson and Post, 2020).

Low-resource paraphrase generation has largely focused on transfer rather than data construction. LAPA, for instance, adapts pretrained sequence-to-sequence models with adapters and meta-learning under limited supervision (Li et al., 2022). We instead focus on which kind of target-language supervision is most useful when the training budget is fixed.

For Ukrainian, the closest prior resource is Spivavtor (Saini et al., 2024), a text-editing dataset and instruction-tuned model suite that includes paraphrasing alongside other rewriting tasks. It is therefore an important baseline, but not a dedicated paraphrase corpus built from naturally occurring Ukrainian text. We also draw on the evaluation perspective of ParaNlu (Srikanth et al., 2024), which shows that paraphrase quality should be assessed jointly in terms of semantic equivalence and surface variation rather than semantic similarity alone.

3 Dataset Methodology

3.1 Data Source and Candidate Generation

We build our corpus from the *Ukrainian News* dataset (zeusfsx, 2024), a collection of 22.5 million news titles collected from Ukrainian online media. News headlines naturally form paraphrases because independent outlets reporting on the same event write alternative descriptions of the same content.

To identify candidates, we encoded the corpus using a multilingual sentence-transformer (Reimers and Gurevych, 2019) and retrieved nearest neighbors for each title using an approximate nearest-neighbor index (Johnson et al., 2021). Although all retrieved neighbors were considered initially, we restricted the candidate pool to pairs with similarity of at least 0.5. The thresholds were selected conservatively through manual inspection of candidate pairs, prioritizing removal of obvious semantic drift, near-duplicates, large information mismatches, and factual inconsistencies while preserving lexically diverse reformulations. Applying the cutoff therefore reduced the candidate set substantially and prevented unnecessary computational overhead in downstream filtering.

3.2 Filtering Pipeline

Raw candidate pairs contain substantial noise. We therefore applied several filtering steps to remove unsuitable pairs. These include removing near-duplicates with high word-overlap Jaccard scores (> 0.8), pairs with low semantic similarity (cosine score < 0.6), large length mismatches ($\text{len\_ratio} < 0.67$), and pairs containing inconsistent numeric values. Additional heuristics removed mirror duplicates, mixed-language content, non-Cyrillic markup, and temporally distant pairs (publication dates differing by more than three days). Examples of retained pairs and discarded pairs for each rule are provided in Appendix A.1.4 and Appendix A.1.5.

4 Dataset Comparison and Human Evaluation

4.1 Dataset Comparison

We compare our news-mined paraphrase pairs with three alternative Ukrainian sources that differ in construction. The first consists of our *news pairs*, independently written headlines about the same event mined from the *Ukrainian News* corpus. The second includes *LLM-generated pairs*, created through instruction-based rewriting by a large language model. The third contains *translated pairs*, obtained by translating English sentence pairs sampled from ParaNlu into Ukrainian (Srikanth et al., 2024). Finally, we include *Spivavtor-derived pairs*, sampled from the paraphrasing subset of the Ukrainian text-editing dataset (Saini et al., 2024). These sources represent different strategies for constructing paraphrase supervision: native reformulations, LLM rewrites, translated pairs, and instruction-based editing data.

Source	n	Len.	Overlap
News (ours)	213	77.8	0.267
LLM	141	84.9	0.174
Translated	144	48.3	0.180
Spivavtor	152	70.7	0.277
Overall	**650**	**71.2**	**0.230**

Table 1: Validated pairs used in human evaluation, with average paraphrase length (chars) and token overlap.

4.2 Human Evaluation Protocol and Results

For human evaluation, we sampled pairs from all four sources, merged them, shuffled, and removed source labels before annotation. Each pair was evaluated along three dimensions: *meaning preservation*, *lexical divergence*, and *fluency*. Meaning was scored on a three-point scale (0 = different meaning, 1 = similar meaning, 2 = identical meaning), lexical divergence on a three-point scale reflecting word overlap (0 = mostly the same words, 1 = partial overlap, 2 = different wording), and fluency on a binary scale (0 = not natural, 1 = natural). Annotation was performed by a professional linguist. To assess reliability, a subset of the data was independently annotated by a second annotator. Under this scheme, agreement exceeded 80% for meaning and fluency, and was around 75% for lexical divergence. In cases of disagreement, we adopted the labels of the primary annotator due to their linguistic expertise in Ukrainian. After validation and removal of incomplete entries, 650 pairs remained for analysis.

Source	Meaning	Lex. div.	Fluency	Total	n
LLM-gen.	1.94	0.12	0.91	2.98	141
News pairs	1.48	0.36	0.94	2.78	213
Translated	1.39	0.49	0.70	2.58	144
Spivavtor	1.44	0.44	0.65	2.53	152

Table 2: Human evaluation average scores by source family. Total is the sum of Meaning, Lexical divergence, and Fluency.

Table 2 shows a clear ordering: LLM-generated pairs have the highest aggregate score, while Spivavtor-derived pairs have the lowest. This also reveals an important trade-off: LLM outputs are strong on meaning preservation, but lexical divergence is the lowest (0.12), which indicates conservative, safe rewrites.

To check whether this pattern persists among acceptable paraphrases, we apply a stricter subset filter (meaning score ≥ 1 and fluency $= 1$) and recompute lexical divergence.

Source	Lexical div.	n
News pairs	0.259	185
Translated	0.209	91
Spivavtor	0.153	98
LLM-generated	0.078	128

Table 3: Lexical divergence on the strict subset with meaning score ≥ 1 and fluency $= 1$.

Table 3 shows that, under this strict filter, news pairs have the highest lexical divergence and LLM-generated pairs remain the lowest. This confirms that LLM rewrites are often safe edits even when meaning and fluency are acceptable.

To test whether score distributions differ across source families, we ran Kruskal–Wallis tests for meaning, lexical divergence, fluency, and total score. All tests were significant (all $p < 10^{-6}$), which indicates that the differences between data sources are systematic rather than random variation.

Overall, LLM-generated pairs obtain the highest aggregate score, but lexical changes are minimal. In this sense, they are not the strongest paraphrases when the goal is non-trivial reformulation rather than safe editing. At the same time, news-mined data offers the strongest balance between acceptable meaning, fluent language, and non-trivial lexical variation.

5 Fine-Tuning

5.1 Setup and Metrics

For downstream experiments, we train all models on one combined dataset of about 16,000 pairs. We sample the data evenly from four sources (news-mined, LLM-generated, translated, and Spivavtor-derived) to cover a broader paraphrase distribution. This training size is also comparable to the Spivavtor paraphrasing subset scale.

The two adapter-tuned models are mT5-large (Xue et al., 2021) and mT0-large (Muennighoff et al., 2023). We compare them with Spivavtor-large and Spivavtor-XXL. Spivavtor-large follows the same mT5/T5-style encoder–decoder family, while Spivavtor-XXL belongs to a much larger parameter class (13B class), so we treat it as a high-capacity reference.

Evaluation is performed on three held-out sets: the combined test split ($n = 1600$), the Spivavtor test split ($n = 1563$), and a human-validated subset ($n = 502$) containing pairs with meaning score ≥ 1 and fluency $= 1$.

BLEU against a single reference is not sufficient for paraphrase evaluation because many valid rewrites can differ from the gold surface form. We therefore report two reference-based overlap metrics, $BLEU_{ref}$ and $chrF^{++}_{ref}$ (Popović, 2017), together with input-based semantic similarity measured by *BERTScore* (Zhang et al., 2020) using *XLM-RoBERTa-large* (Conneau et al., 2020) and copying measured by $BLEU_{in}$. We do not include METEOR (Banerjee and Lavie, 2005) in the main table because its standard formulation depends on language-specific stemming and synonym resources that are not well-defined for Ukrainian in our setup; a surface-form fallback adds little beyond the other reference-based metrics. We also considered ParaScore (Shen et al., 2022), but do not include it because it is not directly adapted to Ukrainian out of the box; validating paraphrase-specific learned metrics for Ukrainian remains future work.

Following Zou et al. (2025), we compute *iBScore* as a combined measure of semantic preservation and anti-copying behavior:

$$iBScore = BERTScore_{in} - \frac{BLEU_{in}}{100}.$$

Higher *iBScore* indicates a better balance between semantic preservation and non-trivial reformulation.

5.2 Results

Table 4 shows a consistent pattern. Spivavtor-XXL obtains the highest *iBScore* on all three evaluation sets, so we treat it as a high-capacity reference rather than a directly comparable baseline. Among the non-XXL models, mT5-large is best on the combined and human-validated sets, while Spivavtor-large is best on the Spivavtor test split. At the same time, all models keep $BERTScore_{in}$ close to 0.98, which indicates similar semantic preservation.

The reference-based metrics add little extra separation. $BLEU_{ref}$ and $chrF^{++}_{ref}$ (Popović, 2017) show the same broad behavior: Spivavtor-XXL remains strongest on all splits, while the remaining systems stay close on the combined and human-validated sets and differ only modestly on the Spivavtor split. In other words, reference-based overlap metrics do not expose meaningful differences beyond the main copying-aware picture captured by *iBScore*.

The main differences come from input copying. On the combined test split, the adapter-tuned mT5/mT0 models show lower $BLEU_{in}$ than Spivavtor-large (roughly 68–69 vs. 69.46). On the human-validated subset, the gap is larger (roughly 69.7–70.2 vs. 71.99). On the Spivavtor test split, Spivavtor-large has lower copying than the two adapter-tuned models, consistent with an in-domain advantage.

To complement the automatic metrics, we manually evaluated 100 generated outputs using the same annotation dimensions as in the dataset evaluation: meaning preservation, lexical divergence, and fluency. The outputs show moderate meaning preservation: 27% were judged to have different meaning, 39% similar meaning, and 34% identical meaning, giving a mean meaning score of 1.07. Lexical divergence was relatively strong: 19% of outputs used mostly the same words, 45% showed partial wording change, and 36% used different wording, giving a mean lexical divergence score of 1.17. Fluency was the weakest dimension, with 63% of outputs judged natural and 37% judged not natural. Overall, 47% of outputs satisfied the criterion meaning ≥ 1 and fluency $= 1$, while 31% reached the stricter tier of identical meaning and fluent wording. During annotation, we also observed that many outputs with meaning changes or poor fluency came from the translated subset. This suggests that fluency issues in translated source texts can propagate to generated paraphrases. By

Eval set	Model	n	$iBScore \uparrow$	$BERTScore_{in} \uparrow$	$BLEU_{in} \downarrow$	$BLEU_{ref} \uparrow$	$chrF^{++}_{ref} \uparrow$
Combined test	Spivavtor-XXL	1600	0.3797	0.9779	59.82	13.37	36.87
Combined test	mT5-large (tuned)	1600	**0.2984**	0.9807	68.23	12.48	35.49
Combined test	mT0-large (tuned)	1600	0.2947	0.9818	68.71	12.48	35.52
Combined test	Spivavtor-large	1600	0.2867	0.9814	69.46	12.48	35.82
Spivavtor test	Spivavtor-XXL	1563	0.3777	0.9770	59.93	20.90	46.15
Spivavtor test	mT5-large (tuned)	1563	0.2854	0.9812	69.58	18.14	43.31
Spivavtor test	mT0-large (tuned)	1563	0.2701	0.9827	71.25	17.94	43.52
Spivavtor test	Spivavtor-large	1563	**0.3038**	0.9807	67.70	20.19	45.38
Human-validated subset	Spivavtor-XXL	502	0.3525	0.9795	62.70	15.74	41.15
Human-validated subset	mT5-large (tuned)	502	**0.2838**	0.9810	69.73	15.44	40.33
Human-validated subset	mT0-large (tuned)	502	0.2811	0.9827	70.16	15.41	40.42
Human-validated subset	Spivavtor-large	502	0.2625	0.9824	71.99	14.91	40.57

Table 4: Fine-tuning and baseline results on three evaluation sets. Lower $BLEU_{in}$ indicates less copying from the input. Bold marks the best result among non-XXL models; Spivavtor-XXL is included as a high-capacity reference.

contrast, the manually inspected outputs from the news-mined subset were generally stronger, which is consistent with our motivation for using native Ukrainian news text as supervision. These results suggest that the main remaining bottleneck is fluency rather than lexical variation.

Taken together, the automatic and human evaluations show that the adapter-tuned models can reduce copying while preserving meaning in many cases, but generation quality remains limited by fluency and requires further improvement.

6 Conclusion

We construct a Ukrainian paraphrase corpus by mining event-aligned news headlines and filtering candidates with semantic and lexical constraints. We compare this data with translated and LLM-generated pairs using human evaluation of meaning, fluency, and lexical divergence. We then fine-tune two adapter-based generation models and evaluate them against strong baselines on three held-out sets with $BLEU_{ref}$, $chrF^{++}_{ref}$, $BERTScore_{in}$, $BLEU_{in}$, and $iBScore$. Across analyses, news-mined pairs show the strongest lexical divergence under quality constraints, and balanced multi-source training supports competitive meaning-preserving generation with reduced copying on the combined and human-validated sets. The two reference-based metrics show the same broad behavior and do not materially separate the non-XXL models.

Data and Code Availability

For reproducibility, we provide public project repositories for the code, dataset, and model checkpoints. The implementation for data construction, fine-tuning, and evaluation is released at GitHub. The dataset and trained checkpoints are released through Hugging Face at dataset repository and model repository.

Limitations

This study has four main limitations. First, we fine-tuned only two large-model backbones (mT5-large and mT0-large, approximately 1.2B class), so broader architectural coverage remains open.

Second, the dataset is built entirely from news headlines. Headlines are short, compressed, and strongly event-centered, so they do not cover longer sentence structures, paragraph-level context, or other text types such as conversational, instructional, or literary prose. As a result, models trained on this data may not transfer reliably to longer inputs or broader domains. Our mining setup also assumes a one-to-one relation between two headlines. Natural paraphrasing can be more flexible: one sentence may correspond to several sentences, several sentences may be compressed into one, and meaning can be redistributed. Extending the approach beyond headline-level one-to-one pairs remains future work.

Third, due to time and compute limits, we trained on the combined training corpus only. We did not fine-tune separate models for each data source family, so we cannot directly test whether source-specific training reproduces the same ordering observed in human evaluation.

Fourth, we could not fine-tune models in the same scale class as Spivavtor-XXL (10B+), including Aya-101-scale systems. This limits conclusions about the best achievable quality under unconstrained compute.

Ethical Considerations

Our data source consists of publicly available Ukrainian news headlines. Although headlines are short, they can include offensive language, insensitive framing, and descriptions of violence or traumatic events. We therefore treat the corpus as potentially harmful content and assume that downstream models may reproduce such language.

Because headlines are collected from publisher websites, source-specific licensing and reuse terms may apply. We use the corpus for research and evaluation, and we recommend that any dataset release follows the legal and attribution requirements of the original sources.

The corpus can also reflect editorial bias, political framing, and factual inconsistency across outlets. Mining paraphrases from event-aligned headlines improves lexical diversity, but it does not remove these source-level biases. Our filtering pipeline reduces semantic drift and obvious mismatches, yet it cannot guarantee factual correctness of the underlying claims.

Human annotation introduces a second risk: annotators may be exposed to disturbing or offensive text. In this study, annotation was performed by a trained linguist, which improved judgment consistency on meaning and fluency. Future annotation rounds should keep clear skip rules and workload limits to reduce potential harm.

The resulting models can be used for beneficial tasks such as rewriting and educational support, but they can also be misused to rephrase misleading or manipulative content. For this reason, we frame the dataset and models as research resources, report their limitations explicitly, and encourage careful deployment with moderation policies.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

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

DataCanary, hilfialkaff, Lili Jiang, Meg Risdal, Nikhil Dandekar, and tomtung. 2017. Quora question pairs. `https://kaggle.com/competitions/quora-question-pairs`. Kaggle.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013. Paraphrase-driven learning for open question answering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1608–1618, Sofia, Bulgaria. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1224–1234, Copenhagen, Denmark. Association for Computational Linguistics.

Zhigen Li, Yanmeng Wang, Rizhao Fan, Ye Wang, Jianfeng Li, and Shaojun Wang. 2022. Learning to adapt to low-resource paraphrase generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1014–1022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755. Springer International Publishing.

Volodymyr Mudryi and Oleksii Ignatenko. 2025. Precision vs. perturbation: Robustness analysis of synonym attacks in Ukrainian NLP. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 131–146, Vienna, Austria (online). Association for Computational Linguistics.

Volodymyr Mudryi and Yurii Laba. 2025. From benchmark to better embeddings: Leveraging synonym substitution to enhance multimodal models in Ukrainian. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20458–20468, Suzhou, China. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Aman Saini, Artem Chernodub, Vipul Raheja, and Vivek Kulkarni. 2024. Spivavtor: An instruction tuned Ukrainian text editing model. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 95–108, Torino, Italia. ELRA and ICCL.

Lingfeng Shen, Lemao Liu, Haiyun Jiang, and Shuming Shi. 2022. On the evaluation metrics for paraphrase generation. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 3178–3190.

Neha Srikanth, Marine Carpuat, and Rachel Rudinger. 2024. How often are errors in natural language reasoning due to paraphrastic variability? *Transactions of the Association for Computational Linguistics*, 12:1143–1162.

Brian Thompson and Matt Post. 2020. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation*, pages 561–570, Online. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

zeusfsx. 2024. Ukrainian news dataset. https://huggingface.co/datasets/zeusfsx/ukrainian-news.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Wei Zou, Ziyuan Zhuang, Xiang Geng, Shujian Huang, Jia Liu, and Jiajun Chen. 2025. Improved paraphrase generation via controllable latent diffusion. *Frontiers of Computer Science*, 20(1).

A Appendix

A.1 Data Preparation

For candidate generation, we used the sentence encoder checkpoint `paraphrase-multilingual-mpnet-base-v2` (model card: https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2).

Both LLM-based headline paraphrasing and English-to-Ukrainian translation were generated with Gemini 2.5 Flash.

A.1.1 LLM headline paraphrase prompt

For headline rewriting, we used the following system prompt.

System prompt:

> You are an expert news editor rewriting headlines.
> Your task is twofold:
> 1. **Clean**: Remove any artifacts from the input title (e.g., website names, "Read more", "Source: ...", truncated text "...").
> 2. **Paraphrase**: Write a new headline that conveys the EXACT SAME information as the cleaned title but uses different vocabulary and structure.
>
> Guidelines:
> - The paraphrase must be in Ukrainian.
> - Do not change names, numbers, or locations.
> - Avoid trivial changes (like just changing one word). Aim for structural variety.

Structured output schema:

- `cleaned_original`: The original title with artifacts removed.
- `paraphrase`: A semantic paraphrase of the cleaned title with different wording.

A.1.2 English-to-Ukrainian bucket translation prompt

The English source sentences were sampled from ParaNlu (Srikanth et al., 2024) and translated using the prompt below.

System prompt:

> You are an expert translator specializing in preserving semantic nuances and variation.
>
> Task: Translate the following list of English sentences into Ukrainian.
> Context: All these sentences are paraphrases of the SAME meaning.
>
> Constraints:
> 1. **Semantic Equivalence**: The meaning must be preserved.
> 2. **Diversity**: The translations must NOT be identical. You must vary the Ukrainian vocabulary and syntax to match the diversity of the English inputs.
> 3. **Quality & Filtering**: If an English sentence is nonsensical, weird, has artifacts, or would result in a broken/meaningless Ukrainian sentence, **SKIP IT**.
> - Do NOT translate nonsensical inputs.
> - It is strictly VALID (and encouraged) if the number of output Ukrainian sentences is smaller than the input list when some inputs are bad.

Structured output schema:

- `ukrainian_sentences`: list of valid Ukrainian translations; list length may be smaller than input due to filtering.

A.1.3 Ukrainian verbalizer prompts used in adapter training and generation.

During adapter fine-tuning and inference, we prepend one verbalizer sampled from the following fixed set:

- **Ukrainian text:** Перефразуй речення:
  English translation: Paraphrase the sentence:
- **Ukrainian text:** Перепиши речення іншими словами:
  English translation: Rewrite the sentence in other words:
- **Ukrainian text:** Перефразуй цей текст:
  English translation: Paraphrase this text:
- **Ukrainian text:** Перефразуй це речення:
  English translation: Paraphrase this sentence:
- **Ukrainian text:** Перефразуй:
  English translation: Paraphrase:
- **Ukrainian text:** Переформулюй це речення:
  English translation: Rephrase this sentence:
- **Ukrainian text:** Переформулюй цей текст:
  English translation: Rephrase this text:

A.1.4 Examples of retained paraphrase pairs

Table 5 shows examples of pairs retained after filtering. These examples illustrate the type of lexical and structural variation preserved in the final corpus.

Source	Sentence A	Sentence B
News	Рада ухвалила держбюджет-2022	Парламент затвердив державний бюджет на 2022 рік
EN	*The Verkhovna Rada adopted the 2022 state budget.*	*Parliament approved the state budget for 2022.*
News	Гелікоптер і літак, що належить родині Медведчука, передали для потреб ЗСУ.	Гелікоптер і літак Медведчука передали на потреби армії
EN	*A helicopter and an airplane belonging to Medvedchuk's family were transferred for the needs of the Armed Forces of Ukraine.*	*Medvedchuk's helicopter and airplane were transferred for the needs of the army.*
News	Буданов назвав дати оголошення нової мобілізації в РФ.	Буданов назвав дату, коли Росія розпочне нову мобілізацію
EN	*Budanov named the dates for the announcement of a new mobilization in Russia.*	*Budanov named the date when Russia will begin a new mobilization.*
News	Майже кожен другий українець незадоволений своєю роботою	Майже половина українців не задоволена своєю роботою
EN	*Almost every second Ukrainian is dissatisfied with their job.*	*Almost half of Ukrainians are dissatisfied with their job.*

Table 5: Examples of retained news-mined paraphrase pairs. English translations are provided for readability.

A.1.5 Filtering examples

Below we show examples of pairs discarded by the filtering pipeline. For each example, we provide both the original Ukrainian text and an English translation.

Near-duplicate filter (Jaccard > 0.8):

Sentence A (**Ukrainian text**): У Львові відреконструюють старе тролейбусне депо

Sentence B (**Ukrainian text**): У Львові реконструюють старе тролейбусне депо

Sentence A (**English translation**): The old trolleybus depot in Lviv will be reconstructed.

Sentence B (**English translation**): The old trolleybus depot in Lviv will be reconstructed.

Low semantic similarity (cosine < 0.6):

Sentence A (**Ukrainian text**): Кахім Перріс зіграв за збірну Ямайки проти Аргентини

Sentence B (**Ukrainian text**): Два ефектних голи Мессі дозволили Аргентині розгромити Ямайку

Sentence A (**English translation**): Kahim Perris played for the Jamaican national team against Argentina.

Sentence B (**English translation**): Two spectacular goals by Messi allowed Argentina to rout Jamaica.

Length disparity ($\text{len\_ratio} < 0.67$**):**

Sentence A (**Ukrainian text**): У Харкові відкрили новий реабілітаційний центр

Sentence B (**Ukrainian text**): У Харкові відкрили новий реабілітаційний центр для військових, який щодня прийматиме до 200 пацієнтів із різних регіонів України

Sentence A (**English translation**): A new rehabilitation center was opened in Kharkiv.

Sentence B (**English translation**): A new rehabilitation center for military personnel has opened in Kharkiv, which will treat up to 200 patients daily from various regions of Ukraine.

Number inconsistency:

Sentence A (**Ukrainian text**): Низка країн на чолі зі США підписали заяву щодо модернізації ППО України

Sentence B (**Ukrainian text**): США та 12 країн світу підписали заяву щодо нагальної модернізації ППО України

Sentence A (**English translation**): A number of countries, led by the United States, have signed a statement on the modernization of Ukraine's air defense system.

Sentence B (**English translation**): The United States and 12 other countries have signed a statement calling for the urgent modernization of Ukraine's air defense system.

A.2 Fine-Tuning Details and Hyperparameters

Both fine-tuned models (mT5-large and mT0-large) use adapters with bottleneck dimension 128 inserted in transformer attention blocks. We train adapters only, while keeping the base model parameters frozen.

The optimizer is AdamW with learning rate 5×10^{-6} and weight decay 0.01. Batch size is 4. We train up to 15 epochs because the training loss and validation metrics plateaued near this range. We evaluate every three epochs and use early stopping

with patience 3 evaluation rounds.

The learning-rate schedule is linear warmup/decay with warmup equal to 10% of total training steps. We apply gradient clipping with max norm 1.0. For generation during evaluation, we use beam search (num beams $= 5$), repetition penalty 1.5, no-repeat trigram constraint, and adaptive output length with approximately $1.3\times$ source-token budget.

Input lengths are truncated to 128 tokens. For metric computation we report $BLEU_{ref}$, $chrF^{++}_{ref}$, $BLEU_{in}$, $BERTScore_{in}$, and $iBScore$ as defined in Section 5.

Automated CEFR-Level Assessment for Ukrainian Texts

Olha Kanishcheva
Heidelberg University
SET University
kanichshevaolga@gmail.com

Mikhail Kopotev
Stockholm University
University of Helsinki
mihail.kopotev@helsinki.fi

Abstract

The present study evaluates CEFR-based text complexity for Ukrainian using a new dataset compiled from textbooks, designed for language learners. We compare traditional machine learning, transformer-based models, and LLM-based evaluation across A1–B2 language proficiency levels. Results show that explicit linguistic features remain highly effective: a Random Forest classifier achieves the highest macro-F1 (0.576), slightly outperforming fine-tuned XLM-RoBERTa (0.574). While GPT-5.5 shows strong performance (macro-F1 0.564), marking a significant advancement over GPT-4.1, supervised models achieve slightly better scores in this experiment for the proficiency-level assessment. These findings suggest that structured linguistic analysis is a robust alternative to purely neural approaches for Ukrainian CEFR classification.

1 Introduction

The Common European Framework of Reference for Languages[1] (CEFR) is widely used to describe second language proficiency through six ascending levels, from A1 to C2. These levels are based on "can-do" statements that explain what learners are able to perform at each level. However, these descriptors are often too general and subjective to provide an operational definition of the linguistic features that distinguish one level from another. Therefore, more objective, data-driven methods for identifying proficiency levels are required.

Recent developments in NLP offer new possibilities for automated language assessment. Traditional approaches have relied on manually selected features within the Complexity-Accuracy-Fluency framework (Michel, 2017). Although useful, average text-level measures often hide important variation within learner texts. Modern machine learning models, including transformer-based approaches such as BERT and generative transformers, allow more accurate classification by analyzing linguistic patterns in large learner corpora (Schmalz and Brutti, 2021).

Automatic assessment is particularly important for educational purposes. It helps create appropriate learning materials and makes information accessible to different groups of learners. While widely spoken languages have many resources and tools for this task (Misgna et al., 2025; Li and Ng, 2024), Ukrainian still lacks modern systems for text classification. Existing approaches often rely on traditional readability formulas developed for English, which do not account for Ukrainian's rich morphology and flexible syntax.

This paper presents a comparative study of automated text classification for Ukrainian across CEFR levels A1 to B2. Using a curated dataset of approximately 400 educational texts, we evaluate the performance of various classification methods, ranging from classical machine learning with handcrafted linguistic features to modern LLMs. The study examines how lexical diversity, morphology, and syntactic complexity contribute to level prediction, providing the first systematic benchmarks for Ukrainian readability assessment.

2 Related Work

A feature-driven approach to CEFR classification relies on machine learning classifiers, such as support vector machines or random forests, trained on linguistically annotated data with lexical, syntactic, and morphological features (Kurdi, 2020; Sung et al., 2015; Balyan et al., 2018). In the extreme case, Hancke and Meurers utilize 3 821 features, which are grouped according to lexical, morphological, and syntactic complexity. Their study found that the strongest signals are syntactic and lexical, and performance comes from combinations of the

[1] https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 209–222

features. The best model achieved 64.5% accuracy.

In practice, researchers typically select only those linguistic features that are available for automatic processing. They are grouped into four categories.

- **Lexical features**: word frequency, lexical diversity, word length, and lexical density. Word frequency is computed by matching lemmas against a reference corpus; lexical diversity is measured as the proportion of types/tokens in a text; word length is calculated as the mean number of characters per token; and lexical density is calculated as the proportion of content words (nouns, verbs, adjectives, and adverbs) among all tokens. The hypothesis underlying these metrics is that lower-level texts typically use high-frequency, simple vocabulary, whereas higher-level production includes specialized and low-frequency terms (Kurdi, 2020; Kate et al., 2010).

- **Syntactic features**: sentence and clause length, tree depth, subordination ratios, and complex constructions. These measures can be extracted from Universal Dependencies (UD) relations (Fig. 1): sentence length in tokens; clause-based measures that rely on verbal heads; subordination operationalized through dependent-clause relations; and syntactic depth calculated as the mean dependency-path in a sentence. The presupposition here is that the lower-level texts tend to use simpler syntactic structures, whereas advanced texts show greater embedding and more varied syntax (Kate et al., 2010; Dascălu et al., 2012).

- **Morphological features**: inflectional diversity, grammatical paradigms, and derivational complexity (Seiffe et al., 2022). These are computed from UD morphological feature bundles, including case, number, gender, tense, aspect, mood, person, and converbal forms. Such measures are particularly important for morphologically rich languages like Ukrainian, where proficiency is reflected not only in lexical choice but also in control over inflectional paradigms.

- **Semantic and cohesion features**: latent semantic analysis, embedding-based similarity, cohesion metrics, and contextual representations (Liu and Lee, 2023). These features are computed as lexical overlap and/or cosine similarity between adjacent sentences or larger textual units using vector representations of words, sentences, or paragraphs.

The selection of these features is theoretically motivated rather than arbitrary. They represent the main linguistic domains that have repeatedly been associated with proficiency development: lexical sophistication and diversity, syntactic elaboration, morphological control, and discourse-level cohesion. Handcrafted features are especially useful in this study because they can be automatically extracted and vary across proficiency levels. They also allow for the identification of linguistic dimensions that contribute to classification rather than treating proficiency prediction as a black-box task. Thus, the feature set is intended not to be exhaustive, but to provide a linguistically motivated and empirically testable baseline for automatic proficiency assessment.

While these features provide a comprehensive framework for analysis, language proficiency assessment has historically relied mainly on graded word lists and a limited set of grammatical features. This approach is often criticized for its inherent subjectivity, as the selection often depends too heavily on the individual experience and pedagogical intuition of the compilers (Kisselev et al., 2024).

Recent research has shifted toward more holistic methods that integrate a broader range of linguistic indicators. In parallel with these developments, data-driven deep learning methods, including fine-tuned Transformer models such as BERT in its multilingual versions, have emerged as powerful tools that can learn contextualized patterns without the need for explicit feature engineering (Imperial et al., 2025; Lee et al., 2021).

Schmalz and Brutti employed pre-trained BERT-base models to classify written exams from the EFCAMDAT and CLC-FCE corpora. Their approach achieved remarkably high accuracy, reaching nearly 98% when trained on large labeled datasets. Furthermore, they discovered that augmenting learner texts with corrections—whether provided by human evaluators or automated tools like LanguageTool—further improved the system's ability to accurately assign CEFR levels.

Hybrid architectures combine transformer embeddings with handcrafted features, achieving strong performance, especially on small- and medium-sized datasets (Lee et al., 2021). Re-

cently, descriptor-based prompting of instruction-tuned LLMs has emerged as a few-shot approach that uses CEFR level descriptors directly (Imperial et al., 2025).

Cross-lingual and multilingual strategies address the scarcity of CEFR-annotated corpora for many languages. Multilingual models can approximate monolingual performance, and large resources such as the UniversalCEFR corpus facilitate standardized benchmarking across languages (Imperial et al., 2025; Vajjala and Rama, 2018).

Research on CEFR classification for Ukrainian is currently limited, largely because of a shortage of annotated corpora. However, methodologies adapted from other morphologically rich or high-resource languages, such as multilingual transformer fine-tuning and hybrid feature-based approaches, provide a robust framework for developing effective Ukrainian proficiency classifiers.

3 Data Collection

The dataset used for training and evaluation was curated from diverse educational resources that remain copyright-protected and therefore cannot be distributed directly. This study specifically targets A1 to B2 proficiency levels, as they represent the most critical stages for language learners. A list of the textbooks used in this study is provided in Appendix A.

3.1 Initial Dataset

At the outset of this research, we faced a significant challenge: Ukrainian lacked labeled datasets based on CEFR levels A1–B2. Most of the necessary materials were only available in printed format, such as textbooks and teaching aids for Ukrainian as a foreign language. To solve this problem, we digitized these materials and labeled the texts according to the respective CEFR levels, as defined in the sources. Selected statistics for the data collected from these training materials are presented in Table 1.

Examples of texts for the selected CEFR levels are presented in Appendix B.

3.2 Linguistic Feature Extraction and Analysis

To evaluate the complexity of the Ukrainian dataset, we extracted a comprehensive set of linguistic features, categorized into four primary dimensions:

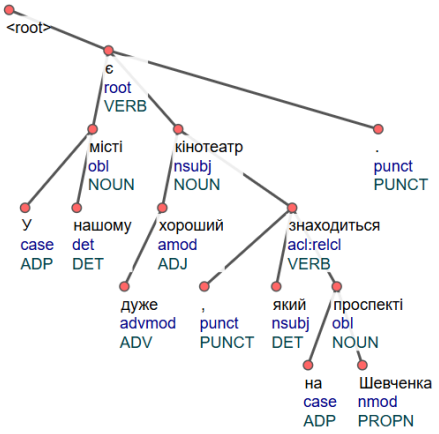


Figure 1: Example of a dependency syntactic tree for a Ukrainian sentence parsed using the UD framework.

- *Descriptive character-based metrics*: average token length (in characters), average syllables per token, and total token count.
- *Lexical diversity* : number of unique tokens and lemmas, number of hapax legomena, and Moving-Average Type-Token Ratio (MATTR) calculated for both tokens and lemmas (Covington and McFall, 2010).
- *Morphological diversity*: part-of-speech (POS) distribution frequencies and the proportion of functional words.
- *Syntactic complexity*: sentence length (mean, median, min, max in both tokens and characters), clause counts per sentence, and dependency tree depth. For each dependency tree, we identify the longest chain of syntactic dependencies from the root node to any dependent node and record its length as the tree's maximum depth. For each text, we then calculate the arithmetic mean of these maximum-depth values in all dependency trees in the text (see the example in Figure 1).

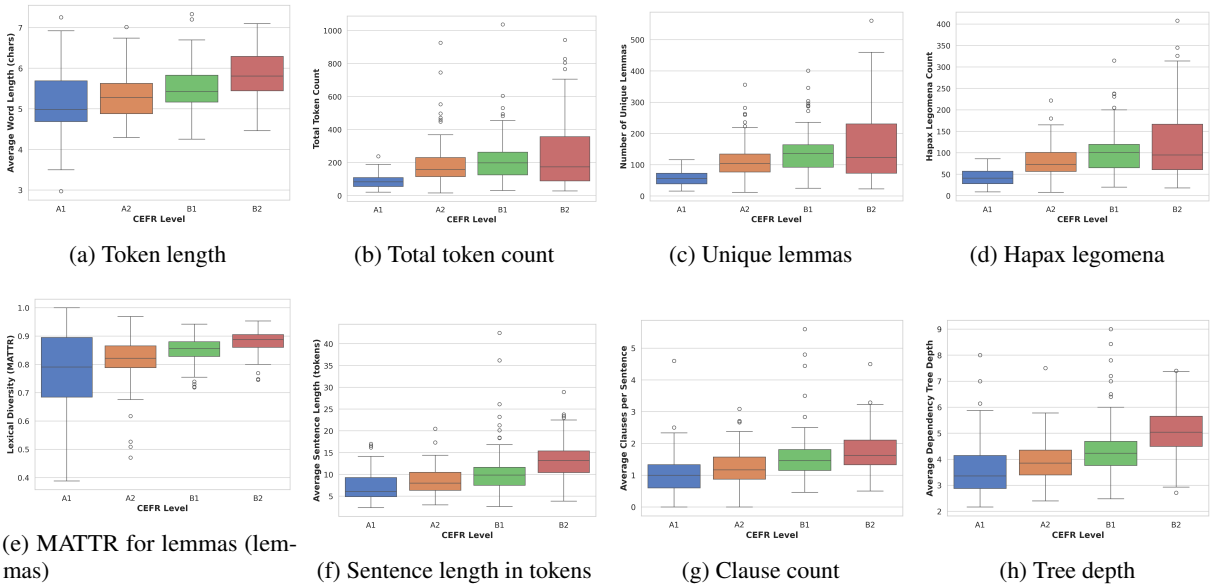
Figure 2 illustrates the distribution of selected metrics. We used boxplots to visualize median values, interquartile ranges, and potential overlaps between the levels. Detailed visualizations for all linguistic features across all CEFR levels, along with the source code for the statistical analysis, are publicly available on the project repository.[2]

The charts in (Figure 2) illustrate the main linguistic characteristics of texts across different lev-

[2] https://github.com/kopotev/fluencymeter/tree/main/UNLP_2026_paper_materials

	A1	A2	B1	B2
Number of files	89	123	110	115
Average text length (tokens)	89	188	215	242
Min/max text length (tokens)	21 / 238	16 / 926	30 / 1037	28 / 943
Total number of tokens	8,270	23,879	24,140	28,224

Table 1: Descriptive statistics for each CEFR level in the initial dataset (Ukrainian educational texts).



(a) Token length (b) Total token count (c) Unique lemmas (d) Hapax legomena

(e) MATTR for lemmas (lemmas) (f) Sentence length in tokens (g) Clause count (h) Tree depth

Figure 2: Distribution of selected linguistic features across CEFR levels A1-B2.

els of Ukrainian language proficiency (A1–B2). Overall, the results show a consistent increase in textual complexity as the proficiency level rises. Lower-level texts (A1–A2) are characterized by shorter lengths, lower average word and sentence lengths, and reduced lexical diversity. At intermediate levels (B1–B2), these indicators gradually increase, reflecting a broader vocabulary and more complex syntactic structures. Thus, the observed trends align with the expected progression of linguistic complexity across CEFR levels.

All feature values for all texts from our dataset are publicly available on GitHub.[3]

3.3 Quantitative Analysis of Text Complexity

To ensure the validity of our corpus and identify the most discriminative linguistic markers for Ukrainian, we performed a comprehensive statistical analysis using one-way analysis of variance (ANOVA).[4] This approach allows us to determine whether the complexity of texts differs significantly between CEFR levels and identify which features most effectively distinguish proficiency stages.

Furthermore, we conducted a linear trend analysis using polynomial contrasts to verify the monotonic progression of complexity from A1 to B2. This dual-method approach confirms that the observed differences are unlikely to be random variation but represent a consistent stepwise increase in linguistic difficulty. The results for the most impactful features are summarized in Table 2.

The statistical analysis revealed that 27 of 30 investigated metrics exhibited significant differences across levels ($p < 0.001$) (Appendix C). The most prominent marker of language proficiency in our study is the *average dependency tree depth* ($F = 41.60, p < 0.001$), which shows a highly significant linear trend. The mean depth increases steadily from 3.68 (A1) to 5.06 (B2), indicating that syntactic structures in Ukrainian become more hierarchical and deeply nested as the level rises.

Syntactic complexity is further evidenced by *sentence length* (mean tokens) and the *average number of clauses*. The average sentence length increases from 7.35 to 13.07 tokens, while the number of clauses nearly doubles.

[3] https://github.com/kopotev/fluencymeter/tree/main/UNLP_2026_paper_materials/figures

[4] https://en.wikipedia.org/wiki/Analysis_of_variance

Linguistic Feature	F-statistics	ANOVA p	Trend p	Mean (A1)	Mean (B2)
Avg. Dependency Tree Depth	41.60	< 0.001	< 0.001	3.68	5.06
Hapax Legomena	39.78	< 0.001	< 0.001	43.11	120.86
Avg. Sentence Length (tok)	36.92	< 0.001	< 0.001	7.35	13.07
Unique Lemmas	34.33	< 0.001	< 0.001	57.85	158.66
MATTR (Lemmas)	31.74	< 0.001	< 0.001	0.77	0.88
Avg. Clauses per Sentence	24.03	< 0.001	< 0.001	1.04	1.75
Noun Frequency	21.66	< 0.001	< 0.001	35.13	99.28
Avg. Token Length (chars)	21.56	< 0.001	< 0.001	5.22	5.85

Table 2: ANOVA results and mean values for selected linguistic features across CEFR levels (A1–B2).

Post hoc comparisons using Tukey's HSD confirmed that these increases are statistically significant at almost every level transition, reflecting a shift from simple sentences to complex multi-clause constructions. Lexical richness metrics also demonstrated robust growth. *Unique Lemmas* ($F = 34.33$) and *hapax legomena* ($F = 39.78$) show substantial increases, particularly during the transition from A1 to A2, where the number of unique words more than doubles.

The *MATTR score for lemmas* score also shows a significant linear trend ($p_{trend} < 0.001$), confirming a more diverse and less repetitive vocabulary at higher proficiency levels. Interestingly, while certain categories such as *gerunds* ($p = 0.094$) and *numerals* ($p = 0.134$) did not show significant differences across the entire range, morphological markers such as *noun frequency* ($F = 21.66$) and *function-word count* ($F = 14.04$) emerged as reliable indicators of complexity. The growth in function words aligns with increased syntactic demands, as they provide the necessary logical and structural cohesion for advanced discourse.

3.4 Thematic Consistency and Data Validation

To ensure that the classification models are driven by linguistic complexity markers rather than specific domain vocabulary, we conducted a thematic distribution analysis using the BERTopic framework.[5] This validation step is crucial for ruling out topic bias, where a model might associate a specific level with a particular subject rather than its grammatical or structural features.

3.4.1 Preprocessing and Model Configuration

Before topic modeling, the text data underwent specialized preprocessing for Ukrainian. This included lemmatization using the `Spacy` (uk_core_news_sm) library to unify inflected word forms and the removal of stop words.

[5] https://huggingface.co/docs/hub/bertopic

The BERTopic pipeline was configured with the following parameters to ensure robust cluster formation:

- **embeddings:** We used the *multilingual-e5-base* model to generate high-dimensional document representations.
- **dimensionality reduction:** UMAP was employed with $n\_neighbors = 30$ and $n\_components = 10$ to preserve local structures.
- **clustering:** HDBSCAN was configured with $min\_cluster\_size = 15$ and $min\_samples = 2$ to minimize noise while identifying distinct thematic groups.
- **vectorization:** A `CountVectorizer` with $ngram\_range = (1, 2)$ was used to capture both individual terms and common phrases.

3.4.2 Analysis of Results

The thematic analysis shows a clear progression in lexical content across CEFR levels. By applying lemmatization, we achieved higher semantic density within clusters, allowing for the identification of specific linguistic markers for different proficiency stages. The key topics identified by the BERTopic model are visualized in Figure 3.

The distribution of these topics across CEFR levels (Figure 4) shows that our data follow a natural path of language learning. Although some subjects appear at all levels, others serve as clear markers of specific proficiency stages.

- **Topic 3**, (related to *classroom/student/room* "аудиторія/ студент/кімната"), is a dominant marker of the **A1-level texts (32.6%)**, representing the basic concrete vocabulary typical of beginners.
- **Topic 6**, (focusing on *culture/technology/profession* "культура/

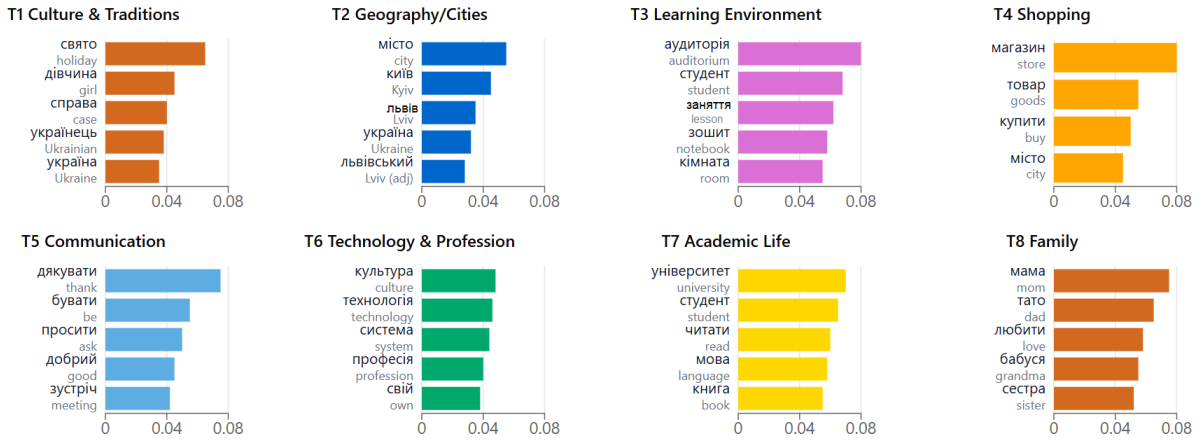


Figure 3: Top words associated with the identified latent topics (Topic Word Scores).

технологія/професія″), becomes significant at the **B2 level (25.2%)**, reflecting the shift toward abstract and professional discourse.

Although the outlier rate for Topic 0 (Unclassified) is relatively high due to the lexical diversity of the corpus, the remaining clusters show clear separation. The consistent presence of Topic 1 and Topic 2 across multiple levels suggests a shared thematic foundation, ensuring that classification performance is driven by structural linguistic complexity and syntactic markers rather than purely topic-specific vocabulary.

4 Model Training and Evaluation

In this section, we describe our experiments to evaluate the effectiveness of various computational methods for the automated CEFR classification of Ukrainian texts. To provide a comprehensive assessment, our experiments are structured around three distinct approaches: (i) feature-driven machine learning, which uses the handcrafted linguistic features analyzed in the previous section to train classical classifiers; (ii) transformer-based deep learning, which leverages state-of-the-art pretrained language models to capture deep contextual representations without manual feature engineering; and (iii) LLM inference, which uses few-shot prompting techniques.

The objective of these experiments is to determine which methodology best captures the morphological and syntactic nuances of Ukrainian while maintaining high classification accuracy across the A1–B2 proficiency levels.

Linguistic features capturing the linguistic properties of the texts were extracted and used as input variables for classification. The feature set included lexical complexity measures (e.g., average token length, average syllables per token, lexical diversity metrics such as MATTR, and hapax legomena counts), vocabulary size indicators (e.g., total tokens, unique tokens, and unique lemmas), part-of-speech frequency distributions (e.g., frequencies of nouns, verbs, adjectives, adverbs, and function words), sentence-level statistics (e.g., mean, median, minimum, and maximum sentence length in tokens and characters), as well as syntactic complexity measures derived from dependency parsing, such as the average number of clauses and the average depth of syntactic trees.

To investigate the contribution of different linguistic aspects, features were organized into several groups: lexical features, part-of-speech distribution features, sentence-level statistics, and syntactic features. In addition to evaluating these groups separately, a combined feature set containing all extracted features was also tested. Before training, numerical features were standardized using z-score normalization, and missing values were handled using median imputation to ensure model robustness.

4.1 Supervised Machine Learning Algorithms

Four supervised machine learning algorithms were evaluated: Random Forest, Gradient Boosting, Support Vector Machines (SVM), and Logistic Regression. All models were implemented using the Scikit-learn library[6] (Pedregosa et al., 2011).

To ensure optimal performance, each model was integrated into a pipeline that included median imputation for missing values and standard scaling of

[6] https://scikit-learn.org/stable/

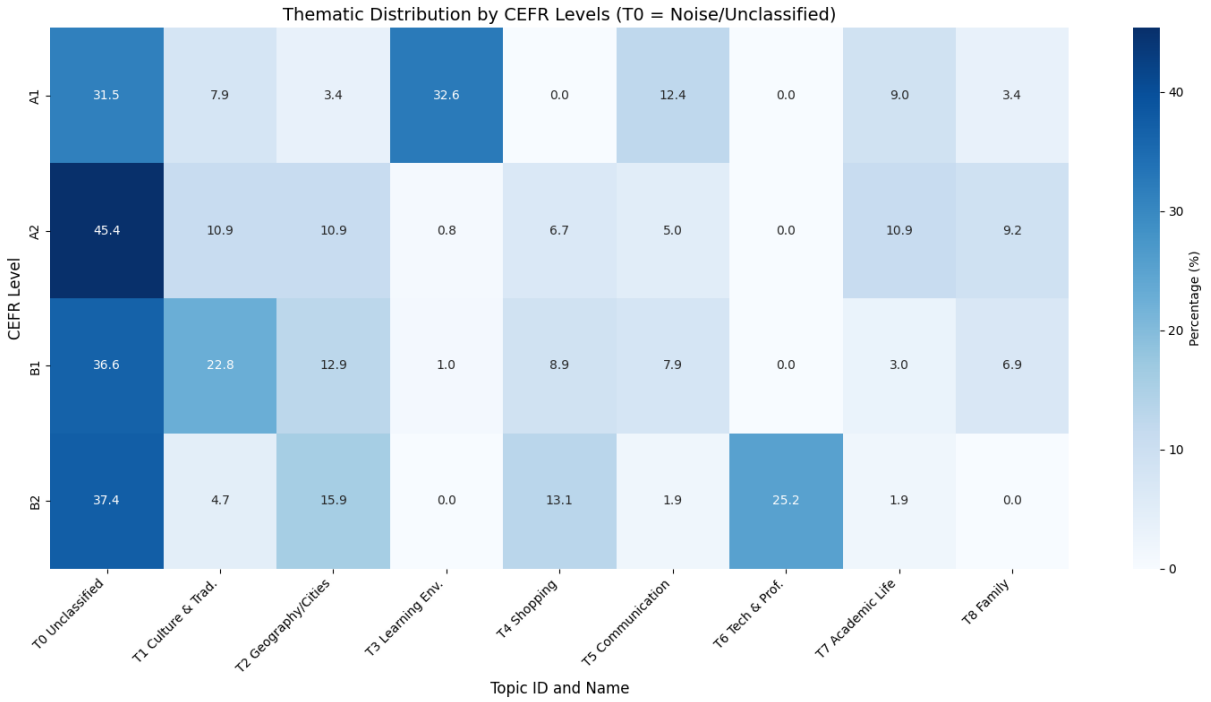


Figure 4: Heatmap of normalized latent topic distribution across CEFR proficiency levels.

linguistic features. Random Forest and Gradient Boosting models were implemented with 300 estimators, providing strong baselines for capturing nonlinear feature interactions. Logistic Regression was trained with a maximum of 2,000 optimization iterations to ensure convergence, while the SVM classifier used a radial basis function (RBF) kernel to address nonlinear decision boundaries.

Model performance was evaluated using stratified five-fold cross-validation, ensuring a balanced distribution of CEFR levels across all folds. For each model and feature configuration, we computed accuracy and macro-averaged precision, recall, and F1-score (Table 3). Macro-averaging ensures that each CEFR level is treated with equal importance, regardless of its frequency in the dataset.

4.2 Transformer-based Deep Learning Methods

For the CEFR classification task, we fine-tuned two state-of-the-art multilingual transformer models: XLM-RoBERTa-base[7] and mBERT (BERT-base-multilingual-cased)[8].

Both architectures were integrated with a sequence classification head and trained on a stratified 80/20 dataset split to ensure consistent class representation across training and testing sets. We used the Hugging Face framework with a learning rate of 2×10^{-5} and weight decay of 0.01. Training was conducted for 6 epochs for XLM-RoBERTa (with a maximum sequence length of 512 tokens) and 4 epochs for mBERT (with a 256-token limit and a batch size of 4). Model performance was evaluated using accuracy and macro-averaged precision, recall, and F1-score to provide a balanced assessment across all CEFR levels (Table 3).

4.3 LLMs and Few-shot Prompting Techniques

In addition to traditional machine learning approaches, we extend our evaluation to modern LLMs to assess their performance on Ukrainian CEFR classification. Specifically, we conducted experiments using the GPT-4.1 and GPT-5 model families via their respective APIs through a few-shot prompting strategy.

The LLM-based classification framework is structured as follows: (i) few-shot prompting is employed by providing the models with a curated set of labeled examples for each CEFR level (A1–B2); these prompts instruct the model to prioritize linguistic complexity – including lexical diversity, syntactic structures, and discourse features – while disregarding text length. This approach ensures

[7] https://huggingface.co/docs/transformers/model_doc/xlm-roberta

[8] https://huggingface.co/docs/transformers/model_doc/bert

consistency, particularly when distinguishing between borderline levels like A2 and B1; (ii) to account for the inherent stochasticity of LLM outputs, we implemented a self-consistency decoding strategy. Each input text is processed through five independent iterations with stochastic sampling. The final proficiency label is then determined by a majority vote across these runs, intended to reduce variability and improve overall prediction reliability.

The configurations for these prompts and the experimental setup are documented in the project repository.[9] Comparative results, showcasing how these LLMs perform alongside the other evaluated models, are presented in Table 3.

4.4 Evaluation Metrics and Results

To ensure a fair comparison, all models, including the feature-based classifiers and the fine-tuned XML-RoBERTa, were evaluated using an identical stratified five-fold cross-validation protocol. This ensures that the performance metrics (macro-F1, accuracy) and the resulting confusion matrices reflect the models' behavior on the same data partitions, eliminating evaluation bias.

The performance of the Random Forest model is visualized through the confusion matrix in Fig. 5. The matrix is presented in row-normalized percentages to facilitate comparison across CEFR levels. Overall, the model demonstrates a robust ability to distinguish between proficiency levels, particularly at the two ends of the scale. The highest class-wise recall rates are observed for A1 and B2 and for the A1 (70.8%) and B2 (68.7%) levels, while the intermediate levels (A2 and B1) show more substantial overlap, which is consistent with the gradual nature of language acquisition.

The transformer-based models demonstrated competitive performance. Fine-tuning XLM-RoBERTa on the CEFR-annotated dataset achieved an overall accuracy of 0.591 and a macro-F1 score of 0.574 (Table 3). While the model performed well on distinct proficiency levels, the intermediate B1 level remained a challenge for neural architectures as well, reflecting the transitional nature of these texts.

Overall, while XLM-RoBERTa achieved the highest accuracy, the feature-based Random Forest model demonstrated nearly identical macro-F1

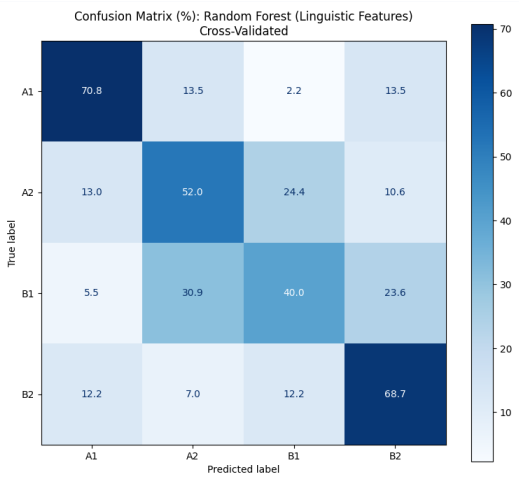


Figure 5: Row-normalized confusion matrix for the Random Forest model, %.

performance, suggesting that explicit linguistic features remain highly effective for Ukrainian CEFR classification.

The evaluation of large language models using prompting shows that few-shot learning provides reasonably strong results but does not always outperform traditional approaches. For GPT-4.1, the standard few-shot setup achieved a macro-F1 score of 0.498 with an accuracy of 0.510. Adding self-consistency led to almost no improvement (macro-F1 = 0.499).

GPT-5.5 performed better overall. Its few-shot configuration reached a macro-F1 score of 0.564 and an accuracy of 0.560, outperforming GPT-4.1 and approaching the performance of the transformer-based models. However, as with GPT-4.1, self-consistency did not bring clear benefits (macro-F1 = 0.561), suggesting that the model already produced stable predictions.

At the same time, supervised transformer models remained slightly stronger. XLM-RoBERTa achieved the highest accuracy (macro-F1 = 0.574, accuracy = 0.591). Among traditional classification methods, Random Forest also performed well (macro-F1 = 0.576).

Overall, the results show that modern LLMs such as GPT-5.5 can handle Ukrainian CEFR classification quite well with few-shot prompting, but supervised models still have a small advantage, especially for more precise distinctions between levels.

[9] https://github.com/kopotev/fluencymeter/tree/main/UNLP_2026_paper_materials/prompts

Model / Approach	Precision	Recall	Accuracy	macro-F1
Baselines (Feature-based ML)				
Random Forest (All linguistic features)	**0.575**	**0.579**	**0.572**	**0.576**
Gradient Boosting (All linguistic features)	0.531	0.533	0.529	0.527
Logistic Regression (All linguistic features)	0.506	0.499	0.501	0.492
SVM (All linguistic features)	0.522	0.507	0.508	0.495
Transformer-based Models				
mBERT (fine-tuned)	0.5299	0.5686	0.550	0.5371
XLM-RoBERTa (fine-tuned)	**0.5648**	**0.5930**	**0.5909**	**0.5742**
GPT-4.1				
Standard Few-Shot	0.620	0.500	0.510	0.4980
With Self-Consistency	0.630	0.500	0.510	0.4994
GPT-5.5				
Standard Few-Shot	**0.620**	**0.550**	**0.560**	**0.5644**
With Self-Consistency	0.620	0.550	0.560	0.5608

Table 3: Performance comparison of different classification methods for Ukrainian CEFR levels.

5 Conclusion and Future Work

This study evaluated computational approaches for the automated classification of Ukrainian texts across four CEFR proficiency levels (A1–B2). Our experimental results demonstrate that while fine-tuned transformer models like XLM-RoBERTa achieve strong performance, with a macro-F1 score of 0.574, traditional classifiers such as Random Forest remain highly competitive (macro-F1 of 0.576) when supplied with robust handcrafted features. In fact, the Random Forest model achieved the highest macro-F1 score, slightly outperforming XLM-RoBERTa in this respect.

The analysis of large language models shows that few-shot prompting achieves moderate results, but LLMs still struggle with fine distinctions between intermediate levels, especially A2 and B1. We also observed that longer input texts may bias the models toward higher-level predictions, likely because of the presence of more complex vocabulary and structures.

While LLMs are flexible and easy to apply, supervised models still offer advantages in this task. Models trained on explicit linguistic features or fine-tuned on annotated data provide more stable and interpretable results for CEFR classification.

Future work will focus on three key areas:

1. **Cross-lingual transfer.** We plan to explore the use of datasets from closely related Slavic languages through translation and adaptation to increase the volume of training data and evaluate cross-lingual feature stability.
2. **Corpus expansion.** We aim to expand our current corpus by collaborating with professional instructors of Ukrainian as a foreign language to ensure high-quality manual annotation and consistency across a wider range of text genres.
3. **Application development.** The long-term goal is to develop a web-based application that provides researchers and educators with a practical interface for the automated analysis and classification of Ukrainian texts according to language proficiency levels.

Limitations

The findings of this study should be considered in light of several limitations:

- Dataset size. The corpus used for this research is relatively small, containing approximately 400 texts. However, this reflects a broader challenge in Ukrainian NLP, where annotated pedagogical datasets are scarce. This study serves as a pilot evaluation to establish baseline benchmarks for larger-scale data collection.
- Level imbalance and coverage. The research focused on levels A1 through B2. While these represent the most active stages of language learning, the lack of C1 and C2 materials in the current dataset limits the model's ability to assess advanced proficiency, where linguistic nuances are more complex.

- Interpretability vs. performance. Although transformer-based models, including modern LLMs, demonstrated strong classification accuracy, they remain black boxes compared to handcrafted linguistic features. This limits their practical pedagogical adoption. The trade-off between the high performance of neural models and the pedagogical interpretability of linguistic features remains a key challenge for automated assessment tools.

Ethical Considerations

We followed ethical guidelines for data use and writing throughout this study. The data came from multiple sources and are protected by copyright, so we cannot share them directly. Instead, we have included a detailed description of all data sources and instructions on how to access them in Appendix A. All data were used under fair-use principles for academic research only. We also used AI tools, including ChatGPT, Gemini, and Grammarly, to edit the manuscript. We carefully reviewed all AI-generated suggestions and take full responsibility for the final content of this paper.

Acknowledgments

We would like to thank the reviewers for their time and effort in reviewing this manuscript. We sincerely appreciate their valuable comments and suggestions, which greatly helped us improve the quality of the work. The authors would like to thank Yurii Prokopenko for valuable assistance in locating relevant educational materials and for helpful consultations during the preparation of this work. This research was partially funded by the Research Council of Finland.

References

Renu Balyan, Kathryn S McCarthy, and Danielle S McNamara. 2018. Comparing machine learning classification approaches for predicting expository text difficulty. In *The Thirty-First International Florida Artificial Intelligence Research Society Conference (FLAIRS-31)*.

Michael A Covington and Joe D McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (MATTR). *Journal of quantitative linguistics*, 17(2):94–100.

Mihai Dascălu, Stefan Trausan-Matu, and Philippe Dessus. 2012. Towards an integrated approach for evaluating textual complexity for learning purposes. In *Advances in Web-Based Learning - ICWL 2012*, pages 268–278, Berlin, Heidelberg. Springer Berlin Heidelberg.

Julia Hancke and Detmar Meurers. 2013. Exploring CEFR classification for German based on rich linguistic modeling. In *Learner Corpus Research*, pages 54–56.

Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Muñoz Sánchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R. Jablonkai, Ekaterina Kochmar, Robert Joshua Reynolds, Eugénio Ribeiro, Horacio Saggion, Elena Volodina, Sowmya Vajjala, Thomas François, Fernando Alva-Manchego, and Harish Tayyar Madabushi. 2025. UniversalCEFR: Enabling open multilingual research on language proficiency assessment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9703–9755, Suzhou, China. Association for Computational Linguistics.

Rohit Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond Mooney, Salim Roukos, and Chris Welty. 2010. Learning to predict readability using diverse linguistic features. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 546–554.

Olesya Kisselev, Mikhail Kopotev, and Anton Vakhranev. 2024. Measuring lexical knowledge in russian as a second language: Exploring the potential of lexical lists for language proficiency assessment. In Gillian Lord and Lara Lomicka, editors, *The Routledge Handbook of Second Language Acquisition and Technology*, chapter 5, pages 65–81. Routledge.

M Zakaria Kurdi. 2020. Text complexity classification based on linguistic information: Application to intelligent tutoring of esl. *Journal of Data Mining & Digital Humanities*.

Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shengjie Li and Vincent Ng. 2024. Automated essay scoring: Recent successes and future directions. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI)*, pages 8114–8122. ijcai.org.

Fengkai Liu and John Lee. 2023. Hybrid models for sentence readability assessment. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 448–454, Toronto, Canada. Association for Computational Linguistics.

Marije Michel. 2017. Complexity, accuracy, and fluency in L2 production. In Shawn Loewen and Masatoshi Sato, editors, *The Routledge Handbook of Instructed Second Language Acquisition*, pages 50–68. Routledge. Available via Lancaster EPrints.

Hiwot Misgna, Byung-Won On, Ingyu Lee, Geunho Choi, and Ha-Young Kim. 2025. A survey on deep learning-based automated essay scoring and feedback generation. *Artificial Intelligence Review*, 58(2):36.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, and David Cournapeau. 2011. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830.

Veronica Juliana Schmalz and Alessio Brutti. 2021. Automatic assessment of English CEFR levels using BERT embeddings. In *Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021)*, pages 295–301.

Laura Seiffe, Fares Kallel, Sebastian Möller, Babak Naderi, and Roland Roller. 2022. Subjective text complexity assessment for German. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 707–714, Marseille, France. European Language Resources Association.

Yao-Ting Sung, Wei-Chun Lin, Scott Benjamin Dyson, Kuo-En Chang, and Yu-Chia Chen. 2015. Leveling L2 texts through readability: Combining multilevel linguistic features with the CEFR. *The Modern Language Journal*, 99(2):371–391.

Sowmya Vajjala and Taraka Rama. 2018. Experiments with universal CEFR classification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 147–153, New Orleans, Louisiana. Association for Computational Linguistics.

A Appendix

The following textbooks were used in the dataset construction:
Палінська О. М. Крок-1 (рівень А1-А2). Українська мова як іноземна: книга для студента / Олеся Палінська, Оксана Туркевич; за ред. Ірини Ключковської. — Львів, 2010. — 102 с.

Українська мова як іноземна. Тексти для читання. Практикум для студентів підготовчого відділення / С. Д. Карпенко, Т. М. Рудакова, О. Д. Будугай та ін.; за ред. С. Д. Карпенко. — Київ: Видавничий дім Дмитра Бураго, 2019. — 256 с.

Українська мова як іноземна. Змістовий модуль «Читання». Рівні складності А2, В1. Методичні вказівки для аудиторної та самостійної роботи студентів підготовчого відділення / уклад. С. Д. Карпенко. — Біла Церква: ВПЦ БНАУ, 2019. — 260 с.

Бакум З. П. Калейдоскоп культур (рівень В1): навчальний посібник / З. П. Бакум, О. О. Пальчикова. — Кривий Ріг, 2014. — 101 с.

Дерба С. М. Українська мова як іноземна: навчальний посібник для студентів-магістрів / С. М. Дерба, Н. С. Ніколаєва. — Київ: Видавництво «Фенікс», 2021. — 136 с.

Тестові завдання до сертифікаційного іспиту з української мови (рівень А1–А2) / укл. Н. М. Малюга, В. А. Городецька. — Кривий Ріг: Видавець Роман Козлов, 2019. — 119 с.

B Appendix

Level	Example (Ukrainian)	Translation (English)
A1	– Привіт! Підеш сьогодні з нами кататися на велосипеді до парку? Я б залюбки, та не можу! Річ у тім, що в мене сестричка захворіла. Зараз до аптеки поспішаю по ліки. А потім буду наглядати за нею, доки батьки з роботи повернуться. – Дуже шкода!	– Hi! Will you go cycling in the park with us today? – I'd love to, but I can't! The thing is, my sister is sick. I'm rushing to the pharmacy for medicine now. And then I'll be looking after her until my parents return from work. – That's a pity!
A2	Мене звати Мухаммед. Я з Камеруну. Моє рідне місто Яунде. Це політична столиця Камеруну. Моя рідна мова французька. Я хочу бути програмістом. Люблю футбол, комп'ютерні ігри, комп'ютерне моделювання. На фото моя родина. ...	My name is Muhammed. I am from Cameroon. My hometown is Yaoundé. It is the political capital of Cameroon. My native language is French. I want to be a programmer. I love football, computer games, and computer modeling. My family is in the photo. ...
B1	Народні звичаї та обряди є давніми формами духовної культури народу. Вони, як і рідна мова, об'єднують людей в один народ. Розрізняють два види обрядів – календарно-обрядові та сімейні. Календарно-обрядові звичаї та обряди пов'язані з календарними циклами (взимку, навесні, влітку, восени). ...	Folk customs and rituals are ancient forms of a nation's spiritual culture. Much like the native language, they unite people into a single nation. There are two main types of rituals: calendar-ritual and family-based. Calendar customs and rituals are associated with seasonal cycles (winter, spring, summer, and autumn). ...
B2	Британська компанія Citymapper запускає нову транспортну послугу Smart Ride. Розробники обіцяють об'єднати в ній переваги трьох видів транспорту: фіксовані зупинки, як у автобуса, можливість замовити поїздку, як в таксі, та єдину транспортну мережу, як в метро. ...	The British company Citymapper is launching a new transport service called Smart Ride. The developers promise to combine the advantages of three modes of transport: fixed stops like a bus, the ability to book a trip like a taxi, and a unified transport network like a subway system. ...

Table 4: Examples of Ukrainian texts by CEFR level.

C Appendix

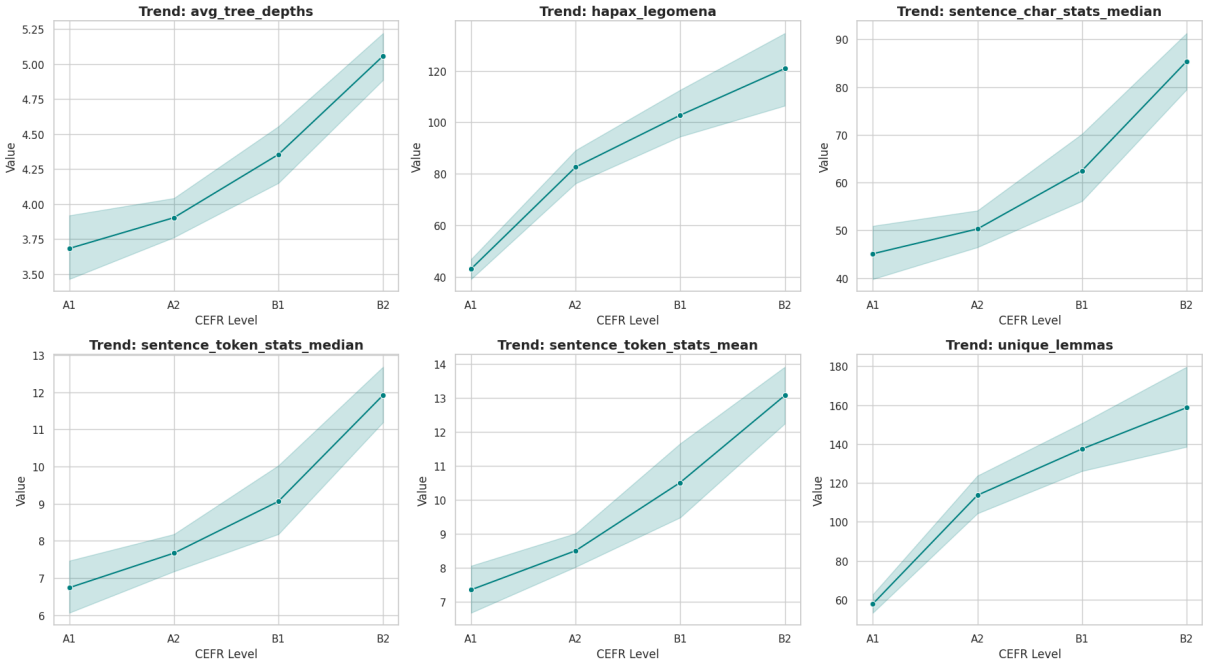


Figure 6: ANOVA trend analysis for key linguistic features across CEFR levels.

An End-to-End Ukrainian RAG for Local Deployment. Optimized Hybrid Search and Lightweight Generation

Mykola Trokhymovych
Pompeu Fabra University
mykola.trokhymovych@upf.edu

Yana Oliinyk
Independent Researcher
oliinykyana@gmail.com

Nazarii Nyzhnyk
Independent Researcher
nazar.nyzhnyk@gmail.com

Abstract

This paper presents a highly efficient Retrieval-Augmented Generation (RAG) system built specifically for Ukrainian document question answering, which achieved 2nd place in the UNLP 2026 Shared Task.[1] Our solution features a custom two-stage search pipeline that retrieves relevant document pages, paired with a specialized Ukrainian language model fine-tuned on synthetic data to generate accurate, grounded answers. Finally, we compress the model for lightweight deployment. Evaluated under strict computational limits, our architecture demonstrates that high-quality, verifiable AI question answering can be achieved locally on resource-constrained hardware without sacrificing accuracy. Code is available at: https://github.com/trokhymovych/unlp-2026-shared-task.

1 Introduction

Large Language Models (LLMs) have emerged as universal tools, demonstrating remarkable capabilities across a wide range of natural language processing tasks (Minaee et al., 2025). While they encode vast amounts of information within their billions of parameters, this internal knowledge is strictly limited by their training data. LLMs become unreliable when tasks require highly specific knowledge - like texts with information that lie beyond models' training corpus or recent facts (Li et al., 2025). In these scenarios, they frequently fall back to so-called hallucinations to fill the gaps in their knowledge.

To bridge this gap, Retrieval-Augmented Generation (RAG) has become the definitive framework (Lewis et al., 2020). By conditioning the generative process on information retrieved from external, domain-specific databases, RAG grounds the LLM in verifiable facts, which aligns with broader efforts in automated fact-checking and knowledge verification (Thorne and Vlachos, 2018; Trokhymovych and Saez-Trumper, 2021). Implementing this, however, adds significant complexity to the data extraction and indexing pipeline. Furthermore, it causes a critical paradigm shift: the precision of the search becomes more important than the raw power of the generative model. Because LLMs rely on the provided context, retrieving incorrect information directly results in an incorrect answer (Niu et al., 2024).

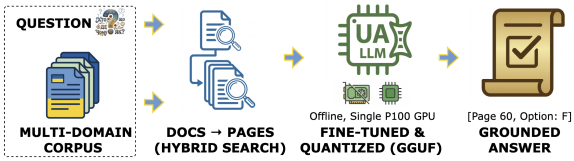


Figure 1: An end-to-end Ukrainian RAG for question answering on local deployment.

Moreover, the RAG paradigm inherently requires processing extensive retrieved contexts. At the same time, most language models are heavily optimized for English (Vargas-Parada, 2025). Applying these models to low- and medium-resource languages like Ukrainian results in a significant decrease in computational efficiency. Default tokenizers fragment Cyrillic text into significantly more subword pieces than Latin text (Maksymenko and Turuta, 2025). This exhausts memory limits, significantly slows inference, and degrades accuracy.

A recent UNLP 2026 Shared Task provided a standardized benchmark for these exact challenges, continuing the community's ongoing efforts to advance Ukrainian NLP and information integrity (Kyslyi et al., 2025; Akhynko et al., 2025). Participants were required to build a system capable of answering multiple-choice questions in Ukrainian, grounding each answer to a specific document and page within a custom, multi-domain corpus. Additionally, the organizers imposed strict

[1] https://github.com/unlp-workshop/unlp-2026-shared-task

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 223–229

hardware constraints: entirely offline inference on a single P100 GPU within a 9-hour limit. In practice, this meant the solution could not rely on third-party LLM providers and had to be both memory-efficient and fast.

This paper presents our solution to the UNLP 2026 Shared Task (see Figure 1), which achieved 2nd place on the final leaderboard. Our main contributions are:

1. A specialized two-stage (document and page-level) hybrid retrieval system.
2. A methodology for generating synthetic Ukrainian QA datasets for model fine-tuning.
3. A customized LLM engineered for high-speed local inference for question answering and grounding.

2 Related Work

2.1 Retrieval-Augmented Generation

Modern RAG architectures have converged on a multi-stage retrieval approach to balance latency with precision (Gao et al., 2024). These pipelines typically begin with a broad search to identify candidate documents, using fast bi-encoders (Reimers and Gurevych, 2019), sparse models such as BM25 (Robertson et al., 1994), or a combination of both. Once a candidate set is retrieved, a more computationally intensive cross-encoder reranker is applied to refine the results and ensure high relevance (Nogueira and Cho, 2019).

Recent studies suggest that hybrid retrieval, which fuses dense semantic embeddings with lexical search, consistently outperforms single methods by capturing both conceptual meaning and exact keyword matches (Akarsu et al., 2026). A common technique for this fusion is Reciprocal Rank Fusion (RRF), an unsupervised method that merges disparate ranked lists by summing the reciprocal of document ranks (Cormack et al., 2009).

Prior to retrieval, processing large documents necessitates chunking (Gao et al., 2024), a step that often causes individual segments to lose their thematic connection to the original source. To address this, contextual embeddings can be utilized to preserve document-wide intent within each fragment (Eslami et al., 2026).

2.2 Language Models for Ukrainian

Modern Large Language Models (LLMs) can effectively handle many languages (Team et al., 2025; Üstün et al., 2024), but they still favor high-resource languages, simply because they dominate the text used to train them. Similar challenges have been previously observed across various NLP tasks, from evaluating text readability to maintaining knowledge integrity and question answering (Trokhymovych et al., 2023, 2024; Zhang et al., 2023). As a result, models struggle with medium-resource languages like Ukrainian. Standard tokenizers break Ukrainian words into too many tokens (Maksymenko and Turuta, 2025). This wastes space, makes processing slower, and hurts the model's overall performance.

To address these gaps, recent work has introduced models specifically optimized for Ukrainian, such as MamayLM (Yukhymenko et al., 2025). Built on the Gemma 3 12B architecture, it underwent continual pre-training on a large, pre-filtered dataset using a combination of data mixing and model merging to gain exceptional Ukrainian cultural and linguistic proficiency. Despite its 12B-parameter size, MamayLM matches or exceeds the performance of significantly larger models, including Llama 3.1 70B, on Ukrainian-specific tasks (Yukhymenko et al., 2025).

2.3 LLM Adaptation and Deployment

LLMs are typically developed as general-purpose models capable of performing a wide range of tasks. However, they can benefit from fine-tuning for specialized applications. Given their massive size, traditional fine-tuning is often computationally impossible under hardware constraints. To address this, previous work introduced LoRA (Low-Rank Adaptation), which only updates small, low-rank matrices that serve as adapters to original model weights modifying key layers such as the attention layers (Hu et al., 2021). This approach enables the efficient adaptation of LLMs for specific tasks such as grounded question answering.

There are also several approaches that enable efficient inference under hardware constraints. In particular, we employ quantization via the GGUF format and the llama.cpp library (Gerganov, 2023). Quantization reduces the numerical precision of model weights, which significantly decreases memory usage and increases throughput with only a moderate impact on accuracy (Rajput and Sharma, 2024). The GGUF format facilitates this process by storing quantized weights alongside essential metadata, enabling efficient loading and execution across diverse hardware configurations.

3 System Architecture

This section describes our final solution for the UNLP 2026 Shared Task. Built on a RAG framework, our system employs a modular pipeline designed to preserve document structure during processing, perform relevant context retrieval across multi-domain corpora, and generate grounded answers to the questions.

3.1 Data Preparation

To convert raw PDF documents into a format suitable for Large Language Models (LLMs), we utilize `pymupdf4llm`[2] to perform layout-aware extraction. This method converts document pages into Markdown, which preserves structural elements such as tables and headers.

Only for lexical search, we implement a custom pre-processing function using `pymorphy3`[3] to perform lemmatization and tokenization. Additionally, we filter out a comprehensive list of Ukrainian stop words[4] and generic artifacts to improve the signal-to-noise ratio during the retrieval. This function is applied to both the query and the corpus for lexical search.

3.2 Hybrid Retrieval Pipeline

Our retrieval strategy is a two-stage process that first identifies the correct document and then finds the most relevant pages within that document for each question.

3.2.1 Document-Level Retrieval.

The initial search narrows the candidate documents to the single most relevant file. We employ a hybrid scoring mechanism that combines:

- Dense Retrieval: We use Perplexity embeddings (`pplx-embed-context-v1-0.6b`[5]), to capture the semantic similarity between the query and the document content. Model choice is motivated by strong performance on benchmarks and compact 0.6B size
- Sparse Retrieval: We utilize the `BM25Okapi` algorithm.

To perform the document search, we represent a query as the question concatenated with its corresponding answer options. To build the vector representation of each document, we embed its first 300 characters and use cosine similarity for ranking. For lexical search, we index each document using its full preprocessed text and use the BM25 score for ranking.

We found that for the majority of questions, the top-ranked item matches for both sparse and dense retrieval and is the correct document. For others, the correct document appears in the top two retrieved documents for at least one approach. Therefore, if the top results match, we return that document. Otherwise, we select the top two documents from each approach and use `jina-reranker-v3`[6] to find the best match. For the reranker, each document is represented by concatenating its first 300 characters with its best BM25 snippet. This algorithm allowed us to achieve near-perfect performance for document retrieval.

3.2.2 Page-Level Retrieval

Once a document is selected, we split it into small parts using syntactic chunking based on Markdown structure. Specifically, we ensure that each chunk is no longer than 500 characters, has a 10% overlap with adjacent chunks, and corresponds to only one page.

Embeddings for each chunk are calculated using the same contextual model as for document-level retrieval. Due to limited resources, we encode chunks in batches of 5 with an overlap of 2 to preserve context. To obtain the final embedding for items in the overlap, we average the results from both batches.

Embeddings are used to identify the chunks ranking via cosine similarity with the question vector. Additionally, we generate a separate ranking based on the BM25 score. We then apply Reciprocal Rank Fusion (RRF) to merge the vector and BM25 scores ranking. Finally, we use a custom cross-encoder based on `BAAI/bge-reranker-v2-m3` to rerank the top-8 candidates, providing a final list of the most relevant chunks for generation. The model was chosen for its strong performance on benchmarks and simple architecture that allows for efficient fine-tuning and inference on a P100 GPU.

[2] https://pymupdf.readthedocs.io/en/latest/pymupdf4llm/

[3] https://pypi.org/project/pymorphy3/

[4] https://github.com/skupriienko/Ukrainian-Stopwords

[5] https://huggingface.co/perplexity-ai/pplx-embed-context-v1-0.6b

[6] https://huggingface.co/jinaai/jina-reranker-v3

3.3 Answer Generation and Grounding

Based on a ranked list of chunks, we build the context out of the full text of the top-3 relevant pages. We pass the context, question, and options to the fine-tuned MamayLM-12B generative model (see the prompt in Appendix A). The model is specifically fine-tuned via Low-Rank Adaptation (LoRA) to perform the dual task of generating an answer with the page number (e.g., "A 2") to ground the response. To optimize the model for deployment in resource-constrained environments (such as Kaggle's P100 GPU), we utilize GGUF 4-bit quantization via the `llama-cpp-python` library.

4 Experimental Setup

4.1 Data

The competition provides a development dataset consisting of 461 questions and 41 documents from two distinct domains. We use this dataset as the primary data for local evaluation of our solution. It should be noted that documents in the corpus vary significantly in length, with some reaching more than 100 pages.

The Shared Task is structured as a code competition, where solution inference is performed on a hidden test set. The leaderboard is split into public (27% of the test data) and private (73% of the test data) sections, with the latter defining the final team ranking. The only information available regarding the test data is that the corpus consists of more than 240 files from a new secret domain and contains a significantly larger number of questions.

To expand the training dataset for custom reranker and foundational LLM fine-tuning, we developed an automated pipeline that synthesizes multiple-choice questions (MCQs) directly from source PDF documents. The pipeline processes documents page by page, first evaluating whether a page contains factual content suitable for question generation. It is explicitly instructed to skip non-factual sections, such as title pages, tables of contents, abbreviation lists, and general introductions.

For each page, the model generates up to ten MCQs. Each question includes exactly six answer options (A through F) with a single correct answer and reference to the document corpus reproducing the structure of the original development data. To ensure each generated question is self-contained and answerable without referencing the source document, the prompt enforces a critical constraint: every question must explicitly state the relevant entity name (e.g., the specific sport or drug).

Generation was performed using OpenAI's `gpt-4o-mini` model, configured with a temperature of 0.7. The complete prompt used for this generation is provided in Appendix A. As a result, this process yielded an additional synthetic dataset of about 7,000 questions.

4.2 Reranker Model Fine-Tuning

The custom reranker is trained using the `BAAI/bge-reranker-v2-m3` model with a hybrid loss objective that combines Listwise Cross-Entropy (weight 1.0) and Pointwise Binary Cross-Entropy (BCE) (weight 0.4) to optimize document chunk ranking. For each query, the model processes a group of candidate chunks and computes a relevance score for each. The model is trained for 5 epochs with a batch size of 2 and gradient accumulation over 4 steps. Final model selection is based on Top-1 Accuracy performance on a hold-out validation subset.

4.3 Generative Model Fine-Tuning

For the generation stage, we fine-tune the `MamayLM-Gemma-3-12B-IT-v1.0` model using Low-Rank Adaptation (LoRA) with a fixed prompt structure (see Appendix A). The training data is processed using a syntactic masking strategy, where the loss is calculated exclusively on the model's generated answer (letter and page number) by masking the prompt tokens. To accommodate long-context documents, the implementation includes a dynamic truncation mechanism with a maximum sequence length of 4000 tokens and utilizes Flash Attention 2 and gradient checkpointing for efficiency. The model is trained using the AdamW 8-bit optimizer and a cosine learning rate scheduler, with a custom evaluation suite that independently tracks accuracy for both the multiple-choice answer and the cited page number. We fine-tune the model for two epochs, first using synthetic data and subsequently on the competition development data, to ensure maximal adaptation of the model to the testing set.

5 Results

5.1 Evaluation Metrics

The evaluation metric is a weighted average that assesses three specific components of the model's output across N questions. Half of the total score

(0.5) is determined by the accuracy of the multiple-choice answers, a_i, where the model receives a point only for an exact match with the ground truth.

The remaining half of the score focuses on the retrieval reference and is split equally (0.25 each) between the document ID, d_i, and the page proximity, p_i. The document score is binary, rewarding the model for identifying the correct file. The page proximity score is more granular; it calculates the distance between the predicted and true page numbers relative to the total number of pages in the document. Crucially, this proximity credit is only awarded if the correct document was identified first ($d_i = 1$). This structure penalizes complete retrieval failures while providing partial credit for "near misses" on specific page locations within the correct document.

5.2 Competition Results

In the final ranking, we achieved second place with a score of 0.942 on the private test set and 0.920 on the public set. Our pipeline successfully processed the hidden test corpus, exactly meeting the strict 9-hour compute limit on a single P100 GPU. Local validation scores on the development set closely matched these final results, demonstrating strong generalization to the unseen domain. Finally, our two-stage hybrid retrieval achieved near-perfect accuracy for document identification and a page-level recall@3 of 0.92 on the development dataset.

6 Conclusion

We presented an efficient, end-to-end RAG system for Ukrainian QA, achieving 2nd place in the UNLP 2026 Shared Task. By combining a two-stage hybrid retrieval pipeline, fine-tuning based on synthetic data, and a quantized `MamayLM-12B model`, we demonstrated that accurate, grounded question answering is viable on local, resource-constrained hardware.

Limitations

Our approach presents several limitations, primarily related to the strict hardware constraints of the shared task. First, to minimize computational overhead, we bypass the processing of images and charts, thereby omitting visual context that could otherwise enhance retrieval and generation accuracy. Second, a marginal fraction of complex PDF pages failed during the layout extraction phase. Although Optical Character Recognition (OCR) could recover this text, the additional processing time outweighed the benefits, given the low incidence rate. Finally, we relied on standard 4-bit GGUF quantization to ensure reliable execution on the provided legacy GPU architecture (P100). Exploring alternative, state-of-the-art quantization techniques could potentially yield further performance improvements if deployed on more modern hardware.

Ethical Considerations

While our system advances Ukrainian NLP, the foundational models and synthetic data pipeline may inherit biases from their pre-training corpora. Furthermore, despite RAG's grounding mechanisms, the risk of hallucination persists. Given the inclusion of sensitive domains like pharmaceutical data, this system requires human oversight for real-world applications. On the other hand, our focus on fully offline, local inference ensures strict data privacy and reduces the environmental footprint of deployment.

We acknowledge the use of AI tools in the preparation of this manuscript. As the authors are non-native English speakers, Google Gemini and Grammarly were used to correct grammar and refine language, improving readability. Additionally, a generative AI model was utilized to create visual elements for the paper's teaser image. Ethically, these tools were not used to generate scientific claims, experimental data, or core ideas, ensuring all intellectual contributions remain solely those of the authors.

Acknowledgments

The work of Mykola Trokhymovych is supported by the Google PhD Fellowship and MCIN/AEI /10.13039/501100011033 under the Maria de Maeztu Units of Excellence Programme (CEX2021-001195-M).

References

Meftun Akarsu, Recep Kaan Karaman, and Christopher Mierbach. 2026. From BM25 to corrective RAG: Benchmarking retrieval strategies for text-and-table documents. *Preprint*, arXiv:2604.01733.

Kateryna Akhynko, Oleksandr Kosovan, and Mykola Trokhymovych. 2025. Hidden persuasion: Detecting manipulative narratives on social media during the 2022 Russian invasion of Ukraine. In *Proceedings of*

the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025), pages 194–202.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 758–759, New York, NY, USA. Association for Computing Machinery.

Sedigheh Eslami, Maksim Gaiduk, Markus Krimmel, Louis Milliken, Bo Wang, and Denis Bykov. 2026. Diffusion-pretrained dense and contextual embeddings. *Preprint*, arXiv:2602.11151.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for Large Language Models: A survey. *arXiv preprint arXiv:2312.10997*.

Georgi Gerganov. 2023. llama.cpp: LLM inference in C/C++. `https://github.com/ggml-org/llama.cpp`. Accessed: 2026-04-06.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of Large Language Models. *CoRR*, abs/2106.09685.

Roman Kyslyi, Nataliia Romanyshyn, and Volodymyr Sydorskyi. 2025. The UNLP 2025 shared task on detecting social media manipulation. In *Proceedings UNLP 2025*, pages 105–111, Vienna, Austria (online). Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of NIPS '20*.

Moxin Li, Yong Zhao, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See-Kiong Ng, Tat-Seng Chua, and Yang Deng. 2025. Knowledge boundary of Large Language Models: A survey. In *Proceedings of ACL'25*, pages 5131–5157.

Daniil Maksymenko and Oleksii Turuta. 2025. Tokenization efficiency of current foundational large language models for the Ukrainian language. *Frontiers in Artificial Intelligence*, 8.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2025. Large Language Models: A survey. *Preprint*, arXiv:2402.06196.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of ACL'24*, pages 10862–10878.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *CoRR*, abs/1901.04085.

Saurabhsingh Rajput and Tushar Sharma. 2024. Benchmarking emerging deep learning quantization methods for energy efficiency. In *2024 IEEE 21st International Conference on Software Architecture Companion (ICSA-C)*, pages 238–242.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Text Retrieval Conference*.

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

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of COLING'18*, pages 3346–3359.

Mykola Trokhymovych, Muniza Aslam, Ai-Jou Chou, Ricardo Baeza-Yates, and Diego Saez-Trumper. 2023. Fair multilingual vandalism detection system for Wikipedia. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 4981–4990, New York, NY, USA. Association for Computing Machinery.

Mykola Trokhymovych and Diego Saez-Trumper. 2021. WikiCheck: An end-to-end open source automatic fact-checking API based on Wikipedia. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 4155–4164, New York, NY, USA. Association for Computing Machinery.

Mykola Trokhymovych, Indira Sen, and Martin Gerlach. 2024. An open multilingual system for scoring readability of Wikipedia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6296–6311, Bangkok, Thailand. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of ACL'24*, pages 15894–15939.

Laura Vargas-Parada. 2025. Large language models are biased — local initiatives are fighting for change. *Nature*.

Hanna Yukhymenko, Anton Alexandrov, and Martin Vechev. 2025. MamayLM: An efficient state-of-the-art Ukrainian LLM. https://huggingface.co/blog/INSAIT-Institute/mamaylm.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of EMNLP'23*, pages 7915–7927.

A Prompts

A.1 Answer Generation Prompt

The following prompt is used for final answer generation with page grounding (Figure 2).

```
Context (excerpts from PDF files - each excerpt
    is
separated by ``` characters and contains a page
number enclosed in []):
```
Page: [<page_number_1>]
<page_text_1>
```
```
Page: [<page_number_2>]
<page_text_2>
```
```
Page: [<page_number_3>]
<page_text_3>
```

Question: <question>
Options:
A: <option_A>
B: <option_B>
C: <option_C>
D: <option_D>
E: <option_E>
F: <option_F>
Instructions:
- Answer the Question using the Context.
- Return the letter of the correct answer (A B C
    D E F)
  and the page number where the information was
    found,
  separated by a space (e.g., A 1).
- Think carefully; first eliminate the obviously
  irrelevant options.
```

Figure 2: Prompt template for answer generation with grounding (translated from Ukrainian).

A.2 Synthetic Question Generation Prompt

The following prompt is used for synthetic MCQ generation (Figure 3).

```
SYSTEM:
You are an expert Ukrainian-language exam
    question writer.

DOMAIN CONTEXT:
{domain_description}

YOUR TASK:
You will receive the text of a single page from
    a Ukrainian PDF document. You must:

1. IDENTIFY the specific subject: the exact
    sport name
   (e.g. "strongman", "sambo") or drug name
   (e.g. "retabolil", "fervex"). This is the
    ENTITY NAME.
2. DECIDE whether this page contains specific
    factual
   content suitable for question generation (
    specific
   rules, dosages, penalties, contraindications,
     etc.).
   SKIP pages that are tables of contents, title
     pages,
   abbreviation lists, or general introductions.
3. If suitable, generate up to 10 MCQs that:
   - Are written entirely in Ukrainian.
   - CRITICAL: Every question MUST explicitly
    name the
     ENTITY NAME. Generic questions are
     FORBIDDEN.
   - Are answerable ONLY from the provided page
    text.
   - Have exactly 6 options (A-F), one correct
    answer,
     and 5 plausible distractors.
   - NEVER use quotation marks or braces.
   - Match the style of: {few_shot_examples}
4. If NOT suitable, return an empty questions
    list.

RESPONSE FORMAT (strict JSON):
{
  "entity_name": "<sport or drug name>",
  "questions": [
    {
      "question": "...",
      "A": "...", "B": "...", "C": "...",
      "D": "...", "E": "...", "F": "...",
      "correct_answer": "A"
    }
  ]
}

Return ONLY valid JSON. "correct_answer" must be
     one of: A, B, C, D, E, F.
If not suitable: {"entity_name": "", "questions":
     []}

USER:
Domain: {domain}
Document page text: {page_text}
```

Figure 3: Prompt template used for synthetic question generation. Note: examples appear in Ukrainian in the actual prompt.

Qwen Goes Brrr: Off-the-Shelf RAG for Ukrainian Multi-Domain Document Understanding

Anton Bazdyrev Ivan Bashtovyi Ivan Havlytskyi
Oleksandr Kharytonov Artur Khodakovskyi
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

Abstract

We participated in the Fifth UNLP shared task on multi-domain document understanding, where systems must answer Ukrainian multiple-choice questions from PDF collections and localize the supporting document and page. We propose a retrieval-augmented pipeline built around three ideas: contextual chunking of PDFs, question-aware dense retrieval and reranking conditioned on both the question and answer options, and constrained answer generation from a small set of reranked passages. Our final system uses Qwen3-Embedding-8B for retrieval, a fine-tuned Qwen3-Reranker-8B for passage ranking, and Qwen3-32B for answer selection. On a held-out split, reranking improves Recall@1 from 0.6957 to 0.7935, while using the top-2 reranked passages raises answer accuracy from 0.9348 to 0.9674. Our best leaderboard run reached 0.9452 on the public leaderboard and 0.9598 on the private leaderboard. Our results suggest that, under strict code-competition constraints, preserving document structure and making relevance estimation aware of the answer space are more effective than adding complex downstream heuristics.

1 Introduction

1.1 Motivation & Context

Real-world document understanding goes beyond extracting an answer span from a passage. A system must navigate long, heterogeneous, domain-specific documents, locate the right evidence among distractors, and tie that evidence to a concrete decision. Generative models alone tend to hallucinate when grounding is required, while retrieval-only pipelines lose critical context once chunks are stripped from their surrounding document structure.

The UNLP 2026 shared task (Sydorskyi et al., 2026) makes these difficulties concrete. Submissions must produce three coupled outputs — the correct multiple-choice answer, the source document, and the page grounding the answer — under scarce task-specific training data, diverse PDF formatting, and strict runtime budgets imposed by the code-competition setting.

1.2 Shared Task Overview

The shared task is organized as a Kaggle code competition. Given a set of PDF documents and a six-option multiple-choice question, the system must predict: (1) the correct answer from options A–F, (2) the supporting document, and (3) the supporting page. The visible training data contain two domains, while the hidden submission set additionally includes an unseen `domain_3`. The visible test set is only a dummy subset; at submission time it is replaced with a substantially larger hidden set of approximately 240 PDFs. All solutions must run within nine hours on Kaggle hardware without internet access.

This setup strongly favors efficient multi-stage pipelines. The system must maintain high retrieval recall under domain shift, but it must also keep the final answerer small enough to fit the competition runtime budget.

1.3 Contributions

We make three main contributions:

1. We present a state-of-the-art retrieval-augmented generation pipeline for Ukrainian multi-domain document understanding that combines structure-aware chunking, question-aware retrieval, and answer-aware reranking within the practical constraints of a code-only shared-task setting.

2. We show that highly competitive performance can be achieved with largely off-the-shelf pretrained models, and that careful pipeline design—particularly preserving document structure and conditioning relevance es-

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 230–239

timation on the full multiple-choice instance—matters more than adding complex downstream heuristics.

3. We construct new Ukrainian resources for retrieval-based multiple-choice question answering, including training data for both dense retrieval and reranking.

2 Related Work

2.1 UA-SQuAD

UA-SQuAD is a key resource for Ukrainian question answering, providing supervised data for reading comprehension in a language with comparatively limited task-specific QA benchmarks (Ivanyuk-Skulskiy et al., 2021). More broadly, extractive QA datasets such as SQuAD and its multilingual extensions have been central for training and evaluating models that align questions with answer-bearing passages, and they often serve as useful transfer supervision even beyond purely extractive settings (Rajpurkar et al., 2016, 2018). In this sense, UA-SQuAD is relevant not only as a benchmark, but also as a practical source of supervision for Ukrainian retrieval and evidence-selection pipelines.

2.2 Document Processing and Contextual Chunking

Prior work on long-document understanding has shown that document processing and segmentation strongly influence retrieval and downstream QA quality, especially for PDFs and other structured sources where headings, layout, and local reading order carry important information (Wang et al., 2025; Lin, 2024). While fixed-size chunking is a common baseline, more structure-aware and context-preserving chunk construction can yield better retrieval units by retaining document titles, section paths, or neighboring context, which makes passages more self-contained for both ranking and answering (Chen et al., 2024; Merola and Singh, 2025). This line of work motivates contextual chunking as a natural design choice in retrieval-augmented document QA (Lu et al., 2025).

2.3 Retrieval in Ukrainian Language

Retrieval for QA is commonly built from dense retrievers, rerankers, and answer modules, with multilingual and cross-lingual transfer playing a major role when native-language retrieval supervision is scarce (Wang et al., 2024; Limkonchotiwat et al., 2024). This is particularly relevant for Ukrainian, where retrieval resources remain more limited than in English, making adaptation from multilingual models and QA-derived supervision especially important (Haltiuk and Smywiński-Pohl, 2025). For evidence-based QA, retrieval is not only a search problem but also a passage selection problem, which is why strong reranking and evidence filtering are often necessary in addition to first-stage retrieval (Limkonchotiwat et al., 2024).

3 Dataset

3.1 Data Source & Task Format

The official competition data was provided in two releases: an initial training set of 40 questions and a subsequently released development set of 461 questions. Both sets are grounded in the same 41 PDF documents spanning two visible domains. Each question has exactly six answer options and is annotated with the correct option, the supporting document, and the supporting page. The hidden submission set includes a third unseen domain, which makes generalization across domains central to system design.

The underlying documents are not short passages but full PDFs with varied formatting, section hierarchies, tables, and long local dependencies. This makes page-level retrieval non-trivial even when document-level identification is relatively easy.

3.2 Auxiliary Training Data

To fine-tune our models, we used a combined dataset that merges competition supervision with auxiliary Ukrainian extractive QA data. We specifically leveraged UA-SQuAD (Ivanyuk-Skulskiy et al., 2021), a Ukrainian adaptation of the widely used SQuAD 2.0 benchmark (Rajpurkar et al., 2018). This dataset provides paragraph-level contexts from Wikipedia paired with localized questions. Incorporating this data allows the model to learn fine-grained context comprehension and exact answer localization before adapting to the longer, noisier PDFs in the competition domain. The final reranker training set contains:

- 461 competition examples;
- 16,658 answerable UA-SQuAD question-context pairs;
- hard negatives mined from impossible UA-SQuAD questions, dense retrieval errors, con-

fusing wrong-answer passages, and same-document wrong-page passages.

3.3 Large-Scale Retrieval Pretraining Corpus

To address the scarcity of high-quality retrieval datasets for the Ukrainian language, we constructed a novel 80,000-example pretraining corpus.[1] Although our final, strongest pipeline did not strictly require these additional pretraining signals, this dataset represents a significant standalone resource. To our knowledge, it is the first large-scale dataset explicitly designed for training bi-encoder embedding models and cross-encoder rerankers in Ukrainian.

We aggregated queries and passages from four established English retrieval datasets to ensure broad domain coverage. The corpus consists of 30,000 rows from Natural Questions (Aarsen, 2024), 20,000 from HotpotQA (Yang et al., 2018), 20,000 from MS MARCO (Nguyen et al., 2016), and 10,000 from GooAQ (Khashabi et al., 2021). This composition captures a wide range of text styles, from formal encyclopedia articles and multi-hop reasoning contexts to everyday web searches and layperson explanations. Overall, the queries range from 5 to 630 characters (with a median of 45), while the positive passages vary significantly from 10 to 9,437 characters (with a median of 372).

We made several deliberate design choices to ensure the dataset is highly effective for contrastive learning:

- **Original Pre-mined Hard Negatives:** Rather than mining new negatives post-translation, we explicitly preserved the original hard negatives provided within the source datasets. Every row includes up to three hard negatives originally pre-mined via dense retrievers or BM25 by the respective dataset creators, forcing the models to learn subtle semantic differences while avoiding translation-induced mining artifacts.
- **Passage Length Diversity:** The extreme variance in document length ensures the resulting models become robust to different chunking strategies, whether they are processing tight 512-token windows or full-page contexts.
- **Cross-Domain Generalization:** Combining diverse domains helps build a generalized retrieval mechanism, reducing the risk of overfitting before task-specific adaptation.

Because the source material is in English, the entire corpus is translated into Ukrainian. To maintain high semantic accuracy during translation, we utilized high-capacity, language-optimized model: our self-hosted TranslateGemma-27B (Finkelstein et al., 2026). The dataset is distributed as a Parquet file natively compatible with standard contrastive learning frameworks.

Table 1 summarizes the resulting training sources, the number of positive and negative examples, and the corresponding negative-mining strategies.

4 Evaluation Metric and Competition Constraints

The competition metric combines answer correctness and reference quality, where reference quality includes both document identification and page localization. For a test set of N questions, the final score is defined as

$$\text{Score} = \frac{1}{2}\cdot\frac{1}{N}\sum_{i=1}^{N} a_i + \frac{1}{4}\cdot\frac{1}{N}\sum_{i=1}^{N} d_i + \frac{1}{4}\cdot\frac{1}{N}\sum_{i=1}^{N} p_i,$$

where

$$a_i = \mathbf{1}(\hat{y}_i = y_i)$$

indicates whether the predicted answer $\hat{y}_i$ matches the gold answer y_i, and

$$d_i = \mathbf{1}\left(\widehat{\text{Doc}}_i = \text{Doc}_i\right)$$

indicates whether the predicted supporting document matches the gold document.

The page-level component is defined as

$$p_i = \left(1 - \frac{\left|\widehat{\text{Page}}_i - \text{Page}_i\right|}{n_i}\right)\mathbf{1}(d_i = 1),$$

where $\widehat{\text{Page}}_i$ is the predicted page, Page_i is the gold page, and n_i is the number of pages in the gold document. Thus, page proximity is credited only when the predicted document is correct; otherwise, the page contribution is zero.

This metric has an important modeling consequence. The system does not need only to retrieve relevant evidence; it must retrieve evidence that is both answer-discriminative and localized tightly enough to preserve page-level precision. Because

[1] https://huggingface.co/datasets/G37A/ua-retrieval-80k-silver

Dataset	Positives	Negatives	Negative Mining	Original Source
Competition train	40	200	Ground truth (wrong options)	41 PDFs (2 visible domains)
Competition dev	461	2,305	Ground truth (wrong options)	Same 41 PDFs (Released later)
UA-SQuAD	16,658	5,766	Impossible Qs & dense errors	Ukrainian Extractive QA
Retrieval Pretraining	80,000	∼240,000	Pre-mined (BM25 & Dense)	NQ, HotpotQA, MS MARCO, GooAQ

Table 1: Dataset composition detailing the volume of positive and negative examples. The mining strategy column highlights how contrastive pairs were sourced, while the original source column indicates the underlying documents or datasets.

the hidden test set is much larger than the visible one and includes an unseen domain, explicit domain probing is forbidden and brittle domain-specific routing is risky.

5 Method

5.1 Technical Details

Retrieval and reranking were run with vLLM-based inference, while the final answerer used constrained single-token generation over the answer alphabet. All major design choices were made under the practical requirement that the complete pipeline fit within the Kaggle code-competition budget on 2×T4 hardware.[2]

Our development process followed a simple progression. We first stabilized chunking, then improved first-stage recall, then focused on reranking, and only after that scaled the final answerer. This ordering turned out to be important: a stronger generator alone could not compensate for structurally weak evidence.

5.2 Contextual Chunking

Our first design decision was to avoid treating a PDF page as plain text. Instead, we chunk documents in a structure-aware way.

PDF processing. We initially used Docling (Team, 2024), a state-of-the-art document understanding library offering layout analysis and table structure recognition. However, Docling proved both slow and unstable in the competition environment, failing on a non-trivial fraction of PDFs. We replaced it with a custom chunker built on `pymupdf4llm`, which converts PDF pages to Markdown via fast native rendering. This replacement yielded a substantial speedup with no drop in leaderboard score.

[2] https://github.com/nuinashco/unlp2026_shared_task

Chunk structure. Each chunk contains three nested levels of context:

1. a short document prefix (global context extracted from the document start);
2. the current heading path (section context);
3. the local chunk body.

In the retained configuration, chunks are built with a maximum length of 512 tokens, 64-token overlap, and a 128-token document prefix. On the 41 competition training documents, this produced 3,362 contextual chunks.

5.3 Dense Retrieval

We use dense bi-encoder retrieval over contextualized chunks. Each query is built from the full multiple-choice instance: the question, answer options, and a short instruction emphasizing document and page grounding. The retriever returns the top-20 chunks for reranking. The exact prompt is provided in Appendix A.

We compared off-the-shelf and fine-tuned embedding models of different sizes; the results are shown in Table 2. The pretrained `Qwen3 8B` embedder (Zhang et al., 2025) gave the best overall retrieval backbone and remained strongest without task-specific adaptation. Fine-tuning it on UA-SQuAD brought no gain, suggesting that larger embedding models already generalize well in this setting and overfit quickly on narrow supervision.

Fine-tuning was more useful for smaller models. For `Qwen3 0.6B`, training on UA-SQuAD and our `80k` corpus substantially reduced the gap to the pretrained `8B` model, while adding UNLP development data degraded performance. We also tested the off-the-shelf diffusion-style embedding model `pplx-embed-v1-4b` (Eslami et al., 2026), but it performed markedly worse than the Qwen3-based alternatives.

Model	Public	Private
Qwen3-Embedding-8B	0.9426	0.9592
Qwen3-Embedding-8B + squad	0.9414	0.9591
Qwen3-Embedding-0.6B + 80k + squad	0.9390	0.9581
Qwen3-Embedding-0.6B + squad	0.9345	0.9534
Qwen3-Embedding-0.6B	0.9370	0.9507
Qwen3-Embedding-0.6B + full*	0.9289	0.9427
Qwen3-Embedding-0.6B + 80k	0.9281	0.9400
pplx-embed-v1-4b	0.8737	0.8759

* 80k + squad + unlp

Table 2: Embedding comparison with fixed off-the-shelf `Qwen3-Reranker-8B` and `Qwen3-32B-AWQ` generator (Yang et al., 2025). `80k` denotes our machine-translated auxiliary retrieval corpus.

Model	Public	Private
Qwen3-Reranker-8B	0.9426	0.9592
Qwen3-Reranker-8B + unlp	0.9286	0.9477
Qwen3-Reranker-8B + squad	0.9452	**0.9598**
Qwen3-Reranker-8B + squad + unlp	0.9330	0.9514
Qwen3-Reranker-8B + 80k + squad	**0.9456**	0.9590
Qwen3-Reranker-4B	0.9429	0.9562
Qwen3-Reranker-4B + squad	0.9253	0.9519
Qwen3-Reranker-0.6B	0.9172	0.9395
Qwen3-Reranker-0.6B + squad	0.9007	0.9257
Qwen3-Reranker-0.6B + squad + unlp	0.9082	0.9360
No Reranker (Baseline)	0.9099	0.9243

Table 3: Reranker comparison with fixed off-the-shelf `Qwen3-Embedding-8B` retriever and `Qwen3-32B-AWQ` generator. The Private leaderboard evaluates on an unseen third domain.

Overall, the results indicate a practical trade-off: if compute allows, a larger pretrained embedder is the strongest choice; under tighter resource limits, fine-tuning a smaller model can recover much of the gap.

5.4 Option-Aware Reranking

The second stage of our pipeline employs Qwen3-Reranker over the top-20 chunks returned by the dense retriever. Crucially, the reranker receives the question concatenated with all six answer options, rather than the question in isolation. This structural change proved decisive: for multiple-choice QA, passage usefulness often depends on which specific alternatives must be distinguished, not just on broad topical similarity. The exact system and user prompts are provided in Appendix B.

To train the model for this discriminative behavior, we relied on the hard negatives detailed in our dataset formulation, specifically emphasizing retrieved chunks that support plausible but incorrect options. After scoring all candidate chunks, we retain only the top-2 passages to pass forward to the final answer generator.

We evaluated reranker models across several sizes and fine-tuning configurations; the results are shown in Table 3. The `Qwen3-Reranker-8B` model fine-tuned exclusively on the auxiliary UA-SQuAD dataset (`+ squad`) provided the strongest overall performance on the Private leaderboard.

As with the embedding stage, the large 8B reranker proved susceptible to overfitting when exposed to competition-specific data. Introducing the UNLP development set into the fine-tuning mix (`+ unlp` and `+ squad + unlp`) noticeably degraded performance, suggesting the model overfit to the characteristics of the two visible domains. Conversely, combining UA-SQuAD with the diverse 80k corpus (`+ squad + 80k`) avoided this degradation, achieving the highest Public score while remaining highly competitive on the unseen domain.

For smaller models (`4B` and `0.6B`), the off-the-shelf versions performed reasonably well but consistently degraded when fine-tuned in any configuration, suggesting that smaller rerankers lack the capacity to learn the complex question-and-all-options formatting without catastrophic forgetting of their general retrieval capabilities.

5.5 Answer Selection and Localization

The final answer is produced by Qwen3-32B AWQ quantized (Lin et al., 2024) from the question, six answer options, and a small set of top reranked passages. We expose passage rank in the prompt so that lower ranks indicate stronger relevance, and constrain decoding to a single token from A–F, which makes inference stable and efficient. The predicted document and page are taken from the highest-ranked selected chunk. The exact system and user prompts are provided in Appendix C.

We also tested a confidence-based variant that adaptively passed one, two, or three passages to the generator based on the top reranker score. Although this was conceptually appealing, it performed slightly worse than the simpler fixed evidence-packing strategy, which we therefore retained.

5.6 Results

Table 4 summarizes key experiments on the competition leaderboard.

The largest single gain comes from document-context prepending: removing it drops the public score from 0.945 to 0.861, confirming that structural signals are critical for passage disambiguation. Replacing the prepended document context with an 80-token summary of the document's first 1024 tokens also hurts: the compressed summary loses the structural signals that make raw prepending effective. Conditioning retrieval and reranking on the full multiple-choice instance, together with reranker fine-tuning, pushes the final system to **0.9598** on the private leaderboard.

To support reproducibility, we release the trained checkpoints and auxiliary resources in a public Hugging Face collection.[3]

Variant	Public	Private
Generator only (no RAG)*	0.3352	0.3392
Early system (BM25, Qwen2.5-32B)	0.9035	0.9114
Ablations from the final pipeline		
w/o document-context prepending	0.8610	0.8891
w/ summary injection (1024→80)	0.9177	0.9408
14B generator, top-10→1	0.9346	0.9483
Rerank Q+A only (retrieval: Q alone)	0.9397	0.9570
Baseline	0.9426	0.9592
Final Pipeline (fine-tuned reranker)	**0.9452**	**0.9598**

* Retrieval score set to 0 (no retrieval component); scores are effectively out of 0.5.

Table 4: Leaderboard scores for key variants. Ablation rows use Qwen3-Embedding-8B, Qwen3-Reranker-8B, top-20→2 selection, full Q+A at both retrieval and reranking, and Qwen3-32B-AWQ, unless otherwise noted.

6 Alternative Approaches

6.1 Gemma3-based models and bidirectionality

We considered a Gemma3-based (Team et al., 2025; Yukhymenko et al., 2025; Paniv et al., 2025) alternative motivated in part by our earlier work (Bazdyrev et al., 2025) on adapting decoder-only models toward bidirectional encoding behavior for downstream understanding tasks. That line of work is relevant here because bidirectionality is appealing for evidence retrieval and page-level localization, where useful signals may depend on relationships spanning both left and right context. More broadly, tasks such as retrieval and named entity recognition often benefit from bidirectional representations when fine-grained contextual interactions are important.

We did not retain this direction in the final system for two practical reasons. First, introducing bidirectional computation substantially increases inference and adaptation cost, which is difficult to accommodate under the strict runtime limits of the shared task. Second, Gemma3 is less convenient in the target hardware setting: Tesla T4 GPUs do not support BF16, while FP16 deployment is known to be unstable for Gemma3 (Han and Han, 2025), and maintaining higher-precision inference is too expensive under the competition budget. We therefore prioritized a Qwen-based pipeline with a better efficiency–quality trade-off for the code-only setting. At the same time, we view bidirectional adaptation of decoder models as a promising direction for tasks where bidirectional dependencies are central and compute constraints are less restrictive.

6.2 Diffusion and Contextual Embeddings

We also explored diffusion-based embedding models inspired by recent work on diffusion-pretrained dense and contextual retrieval embeddings. These models are attractive because they support bidirectional attention, and the contextual variant additionally incorporates document-level context into passage representations.

In our experiments, this direction underperformed the retained Qwen3-based retrieval setup. The off-the-shelf diffusion embedder `pplx-embed-v1-4b` scored well below the best Qwen3 models in the end-to-end pipeline, and the contextual diffusion variant `pplx-embed-context-v1` also performed worse on validation. Even so, we consider this line promising: bidirectional modeling and stronger context handling remain appealing for retrieval, and Ukrainian pretraining with in-domain fine-tuning may yield better results in future work.

6.3 Summary in Context

Rather than prepending the raw first 128 tokens of a document to each chunk — which can be noisy and may truncate before reaching meaningful content — we explored replacing that prefix with a generated summary of the document's leading section (first 1024, 2048, or 4096 tokens). The hypothesis was that a compact, model-generated summary would expose document-level semantics more reliably than a fixed token cut. We used Qwen3-

[3] https://huggingface.co/collections/G37A/qwen-goes-brrr

8B-AWQ to generate these summaries. However, this approach degraded retrieval quality: using the top-1024 token summary, public leaderboard score dropped from 0.9346 to 0.9177.

6.4 Domain Classifier

We added a domain-classifier-first routing stage in which each question was assigned to a predicted domain and retrieval was restricted to the corresponding document partition. Domain labels were obtained by prompting the LLM on each folder's readme, while question domains were classified via a zero-shot prompt. To avoid additional model loads, domain masking was applied directly to the embedding similarity matrix: scores for out-of-domain chunks were set to $-\infty$ before top-k selection, preserving retrieval latency.

On the public leaderboard this configuration achieved our highest score, as the clean separation between the two visible domains allowed the classifier to operate with near-perfect accuracy. However, the private leaderboard introduced an unseen third domain, and the hard routing boundary proved brittle: misclassified questions were irrecoverably directed to the wrong partition, causing a drop from first to fifth place. This confirmed that hard domain routing is a single point of failure whose cost scales with the misclassification rate, and that domain-agnostic retrieval provides the robustness required when hidden-domain generalization is part of the evaluation.

6.5 Hybrid Retrieval

We tested sparse retrieval for two purposes: providing broader document context, and exact keyword matching. For the first, sparse signals were consistently outperformed by chunk prepending. For the second, combining dense and sparse signals via rank fusion yielded no gains over dense retrieval alone — likely because the competition data relies heavily on synonyms and rephrasing, which lexical matching cannot bridge.

6.6 Agentic Inference

Finally, we prototyped an agentic solution in which the model interleaves retrieval, reasoning, and query reformulation over multiple steps. The system used a tool-use loop (search $\rightarrow$ retrieve pages $\rightarrow$ read evidence $\rightarrow$ answer) driven by Qwen3.5-4B (Team, 2026), with hybrid dense+sparse retrieval and an answer repair step. On the competition development set, retrieval navigation was strong (document recall 1.00, page recall 0.96), but final answer accuracy reached only 0.83 — the small reader model consistently mishandled negation and other fine-grained linguistic cues in Ukrainian answer choices.

Beyond this accuracy gap, embedding a multi-round tool-use agent in a Kaggle code-only competition is not straightforward, so we prioritized a single robust retrieval–reranking–generation pipeline instead.

7 Conclusions & Future Work

7.1 Summary of Findings

Our findings can be summarized in four main points. First, we introduced new Ukrainian training resources for retrieval-based multiple-choice document question answering and used them to support both retriever and reranker adaptation. Second, we showed that a strong retrieval-augmented pipeline built largely from modern pretrained components is now sufficient to achieve highly competitive performance for Ukrainian multi-domain document understanding. Third, our embedding experiments revealed a clear efficiency trade-off: larger pretrained embedders already perform very strongly off the shelf, while fine-tuning smaller models can recover much of this gap under tighter compute constraints. Finally, we found that contextualization in chunk construction is one of the most important factors in the entire pipeline, with structure-aware chunking contributing more than many more complex downstream modifications.

Taken together, these results show that strong Ukrainian document QA no longer depends primarily on complex task-specific modeling. Instead, the strongest gains come from combining high-quality pretrained models with careful evidence preparation, efficient retrieval design, and chunk representations that preserve document structure.

7.2 Future Directions

Two directions appear especially promising for future work. First, it would be valuable to explore and benchmark diffusion-based and other bidirectional embedding models in a setting without the compute constraints of the shared task, while also investigating practical inference-speed optimizations for such models. Second, late chunking is a highly relevant extension of our current retrieval setup: our results already show that contextualized chunks matter substantially, and late chunking is ex-

plicitly designed to give each chunk access to full-document context before pooling, which helps preserve cross-chunk dependencies in retrieval (Günther et al., 2025).

Limitations

Although the final pipeline is feasible under the Kaggle code-only constraints, it still relies on comparatively large models. This setting rewards fitting the strongest possible components into a fixed offline runtime budget, whereas real production deployments operate under a broader and more dynamic set of practical constraints.

Part of our auxiliary retrieval pretraining corpus was produced by machine translation from English sources. Although we used a high-capacity translation model, the translated examples were not manually reviewed by human annotators.

Acknowledgements

We thank the organizers of the UNLP 2026 shared task for preparing the benchmark and the code-only competition setup.

References

Tom Aarsen. 2024. natural-questions-hard-negatives. https://huggingface.co/datasets/tomaarsen/natural-questions-hard-negatives. Hugging Face dataset, accessed 2026-04-08.

Anton Bazdyrev, Ivan Bashtovyi, Ivan Havlytskyi, Oleksandr Kharytonov, and Artur Khodakovskyi. 2025. Transforming causal LLM into MLM encoder for detecting social media manipulation in telegram. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 112–119, Vienna, Austria (online). Association for Computational Linguistics.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. 2024. Dense X retrieval: What retrieval granularity should we use? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15159–15177, Miami, Florida, USA. Association for Computational Linguistics.

Sedigheh Eslami, Maksim Gaiduk, Markus Krimmel, Louis Milliken, Bo Wang, and Denis Bykov. 2026. Diffusion-pretrained dense and contextual embeddings. *Preprint*, arXiv:2602.11151.

Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, and 2 others. 2026. TranslateGemma technical report. *Preprint*, arXiv:2601.09012.

Michael Günther, Isabelle Mohr, Daniel James Williams, Bo Wang, and Han Xiao. 2025. Late chunking: Contextual chunk embeddings using long-context embedding models. *Preprint*, arXiv:2409.04701.

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

Daniel Han and Michael Han. 2025. Fine-tune & run gemma 3.

Bogdan Ivanyuk-Skulskiy, Anton Zaliznyi, Oleksand Reshetar, Oleksiy Protsyk, Bohdan Romanchuk, and Vladyslav Shpihanovych. 2021. ua_datasets: a collection of ukrainian language datasets.

Daniel Khashabi, Amos Ng, Tushar Khot, Ashish Sabharwal, Hannaneh Hajishirzi, and Chris Callison-Burch. 2021. Gooaq: Open question answering with diverse answer types. *arXiv preprint*.

Peerat Limkonchotiwat, Wuttikorn Ponwitayarat, Lalita Lowphansirikul, Potsawee Manakul, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, and Sarana Nutanong. 2024. McCrolin: Multi-consistency cross-lingual training for retrieval question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2780–2793, Miami, Florida, USA. Association for Computational Linguistics.

Demiao Lin. 2024. Revolutionizing retrieval-augmented generation with enhanced pdf structure recognition. *Preprint*, arXiv:2401.12599.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. Awq: Activation-aware weight quantization for llm compression and acceleration. *Preprint*, arXiv:2306.00978.

Wensheng Lu, Keyu Chen, Ruizhi Qiao, and Xing Sun. 2025. Hichunk: Evaluating and enhancing retrieval-augmented generation with hierarchical chunking. *Preprint*, arXiv:2509.11552.

Carlo Merola and Jaspinder Singh. 2025. Reconstructing context: Evaluating advanced chunking strategies for retrieval-augmented generation. *Preprint*, arXiv:2504.19754.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268.

Yurii Paniv, Bohdan Didenko, Mykola Haltiuk, Vladyslav Humennyy, Andrian Kravchenko, Roman Kyslyi, Viktoriia Makovska, Artem Orlovskyi, Bohdan Ruban, Maksym-Yurii Rudko, Anastasiia Senyk, Nazarii Drushchak, Dmytro Chaplynskyi, and Mariana Romanyshyn. 2025. Lapa LLM v0.1.2 — the most efficient Ukrainian open-source language model.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

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

Volodymyr Sydorskyi, Nataliia Romanyshyn, Roman Kyslyi, and Olena Nahorna. 2026. The UNLP 2026 shared task on multi-domain document understanding. In *Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026)*, Lviv, Ukraine. Association for Computational Linguistics. To appear.

Deep Search Team. 2024. Docling technical report. Technical report.

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

Qwen Team. 2026. Qwen3.5: Accelerating productivity with native multimodal agents.

Dingmin Wang, Qiuyuan Huang, Matthew Jackson, and Jianfeng Gao. 2024. Retrieve what you need: A mutual learning framework for open-domain question answering. *Transactions of the Association for Computational Linguistics*, 12:247–263.

Zhitong Wang, Cheng Gao, Chaojun Xiao, Yufei Huang, Shuzheng Si, Kangyang Luo, Yuzhuo Bai, Wenhao Li, Tangjian Duan, Chuancheng Lv, Guoshan Lu, Gang Chen, Fanchao Qi, and Maosong Sun. 2025. Document segmentation matters for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8063–8075, Vienna, Austria. Association for Computational Linguistics.

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

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Hanna Yukhymenko, Anton Alexandrov, and Martin Vechev. 2025. Mamaylm v1.0: An efficient state-of-the-art multimodal ukrainian llm.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

A Retrieval Prompt

Instruct: Given a multiple-choice question in Ukrainian, retrieve relevant passages from Ukrainian PDF documents that help identify the correct supporting document and page.
Query: {question}
Options:
{choices}

B Reranking Prompts

System Prompt

Judge whether the Document meets the requirements based on the Query and the Instruct provided. Note that the answer can only be "yes" or "no".

User Prompt Template

Original

<Instruct>: Given a web search query, retrieve relevant passages that answer the query.
<Query>: {question}
Варіанти відповідей:
{choices}

<Document>: {document}

English translation

<Instruct>: Given a web search query, retrieve relevant passages that answer the query.
<Query>: {question}
Answer options:
{choices}

<Document>: {document}

C Answer Generation Prompts

System Prompt

Original

Ти розв'язуєш завдання множинного вибору за наданими уривками з документів. Уважно прочитай запитання, варіанти відповідей і всі уривки. Менший ранг пошуку означає сильніший сигнал релевантності. Якщо уривки суперечать один одному, віддавай перевагу більш прямому та конкретному формулюванню. Якщо інформації недостатньо, все одно обери найімовірніший варіант. Поверни лише одну велику латинську літеру: A, B, C, D, E або F.

English translation

You are solving a multiple-choice task using the provided document excerpts. Carefully read the question, answer options, and all excerpts. A lower retrieval rank indicates a stronger relevance signal. If the excerpts contradict each other, prefer the more direct and specific formulation. If the information is insufficient, still choose the most likely option. Return only one uppercase Latin letter: A, B, C, D, E, or F.

User Prompt Template

Original

Запитання: {question}

Варіанти відповідей:
{choices}

Надані уривки (менший ранг пошуку = сильніший сигнал):
{context_blocks}

Запитання: {question}

Варіанти відповідей:
{choices}

Відповідь (лише одна літера A-F):

English Translation

Question: {question}

Answer options:
{choices}

Provided excerpts (lower retrieval rank = stronger relevance signal):
{context_blocks}

Question: {question}

Answer options:
{choices}

Answer (only one letter A-F):

RAG Pipeline Strategies for Ukrainian Multi-Domain Document Understanding Task

Mykola Nosenko[1] **and Pavlo Kilko**[2]
Taras Shevchenko National University of Kyiv
Kyiv, Ukraine
[1]nikolay.nosenko@knu.ua
[2]pavel.kilko@knu.ua

Abstract

In this work, we present top-performing solution to the UNLP 2026 Shared Task on Ukrainian Multi-Domain Document Understanding. This task focuses on answering multiple-choice questions grounded in domain-specific Ukrainian documents, while also requiring systems to identify the source document and page. We developed a modular retrieval-augmented generation (RAG) pipeline and conducted a series of ablation experiments over its individual components to identify the best-performing strategy at each stage. Based on our evaluation results, we propose two final pipeline configurations that differ in their computational cost and retrieval accuracy: a stronger but more compute-intensive document-level augmentation approach and a lighter summary-based augmentation that is suitable for constrained environments. Our submission achieved 3rd place on the private leaderboard. This demonstrates that isolated curation of RAG components can yield strong performance for Ukrainian document grounded question answering without additional language model adaptations.

1 Introduction

Large language models (LLMs) are increasingly used in applied systems, but factual errors and hallucinations still limit their reliable deployment in domains sensitive to information quality and trustworthiness (Ji et al., 2023).

This is especially evident in tasks that require external context rather than relying only on the model's internal knowledge. Such settings fall within document understanding and document question answering, where system quality depends not only on answer correctness, but also on identifying and using relevant supporting evidence.

One of the most common approaches to address this problem is retrieval-augmented generation (RAG), which combines answer generation with the retrieval of relevant information from external sources. Subsequent work has shown that the effectiveness of RAG depends not only on the base model, but also on the configuration of the entire pipeline, including retrieval, context processing, and answer generation. In many applied scenarios, what matters most is not a new architecture, but the careful selection and coordination of existing components.

These challenges are particularly relevant for mid-resource languages. Research on multilingual RAG shows that transferring English-centric solutions is non-trivial and depends on both language-specific and retrieval-related factors (Wu et al., 2024). For Ukrainian, this is further complicated by limited adapted resources, models, and evaluation practices, despite recent progress in Ukrainian-language LLMs and benchmarks (Paniv, 2025).

The UNLP 2026 Shared Task on Ukrainian Multi-Domain Document Understanding[1] requires systems to answer multiple-choice questions grounded in domain-specific Ukrainian PDF documents while also identifying the source document and page. In this work, we describe the pipeline we developed for this task, focusing not on a new architecture or additional model adaptation, but on retrieval pipeline engineering through the selection and coordination of its components. The resulting system placed third on the private competition leaderboard.

2 Related Work

Retrieval-augmented generation has become a practical framework for building systems (Lewis et al., 2021) in which answers are generated from external context rather than solely from the model's parametric knowledge. In document-grounded question answering, this shifts the focus from generation alone to the design of the full pipeline, including

[1]https://unlp.org.ua/shared-task/

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 240–248

document representation, retrieval, reranking, and context construction.

One line of techreport (Smith and Troynikov, 2024) studies how documents should be represented for indexing and retrieval, including chunking strategies, fragment size, overlap, and ways of preserving the link between local fragments and broader document context. These choices directly affect both retrieval quality and the accuracy of linking answers to specific source segments.

Another important line of work focuses on improving retrieval at the inference stage. This includes dense retrieval, hybrid search, query expansion, hypothetical document generation, and various reranking approaches applied after the initial retrieval step. Reranking models play a particularly important role in modern pipelines, especially cross-encoder approaches and LLM-based schemes, which allow for more precise estimation of the match between a query and a candidate. At the same time, the effectiveness of such methods depends on the specific task setting: techniques that improve performance in open-domain scenarios with noisy user queries do not necessarily yield the same effect in controlled benchmark-oriented settings.

For document understanding tasks, performance depends not only on retrieving relevant text, but also on how context is constructed for the final model (Reuter et al., 2025). Pipeline designs differ in how they combine retrieved fragments, apply contextual enrichment, and preserve the link between retrieval and final answer selection. As a result, system quality is determined by the coordination of the pipeline as a whole.

3 System Architecture

Analyzing existing RAG techniques (Gao et al., 2024), we structure our system as a two-stage pipeline, as shown in Figure 1. Our system covers the two main phases of RAG: the indexing pipeline and the question-answering (QA) pipeline. We adopt a modular architecture in which each pipeline component can be replaced independently. This design directly supports our ablation study (§4): by isolating components, we can measure their individual contributions to overall RAG performance while keeping the rest of the pipeline fixed.

The indexing pipeline consists of four main steps. **Document Loading** converts documents into textual representations, since raw artifacts cannot be embedded directly. **Document Splitting** divides each document into chunks for embedding. **Chunk Augmentation** enriches each chunk with broader contextual information from the source document. Finally, **Embedding and Indexing** encodes chunks as dense vectors and stores them in a vector database using an HNSW index with cosine similarity.

The question-answering pipeline consists of four main steps. **Retrieval** selects relevant chunks from the vector store based on the input question. **Reranking** refines the initial results by reordering the retrieved chunks using a more expressive scoring model. **Context Assembly** combines the highest-ranked chunks into a unified context. Finally, **Answer Generation** prompts the LLM with the assembled context and the input question to produce an answer in the required format.

4 Experiments

4.1 Experimental Setup

Dataset. The development dataset for this task, provided by the organizers, consisted of two parts: a document subset and a question-answer subset. The document subset contained 41 files in PDF format. The question-answer subset included 460 questions from the sports and medicine domains. Each question has six answer choices, one correct answer, a source document ID, and a source page number. Additionally, for embedding model evaluation, we constructed our own UNLP-QA dataset[2]. We aligned the organizers' dataset with the RTEB format (Liu et al., 2025) by manually annotating each question with the corresponding gold context and treating the correct page of the source document as the target retrieved chunk. For local RAG pipeline evaluation, we also prepared a separate subset in TXT format of the source documents by converting the original PDFs with Gemini-2.5-Pro (Comanici et al., 2025).

Language Model. The shared task encouraged participants to use LLMs specifically adapted for the Ukrainian language. Following this recommendation, we chose MamayLM (Yukhymenko et al., 2025), a Ukrainian-adapted fine-tune of Gemma 3 12B (Team et al., 2025). It achieves the highest score on the Ukrainian Language Model Leaderboard (Paniv, 2025) and performs well in our inter-

[2] https://github.com/Dialogus-ex-Machina/unlp-2026-qa-rag-pipeline/blob/main/data/dev_questions_with_context.csv

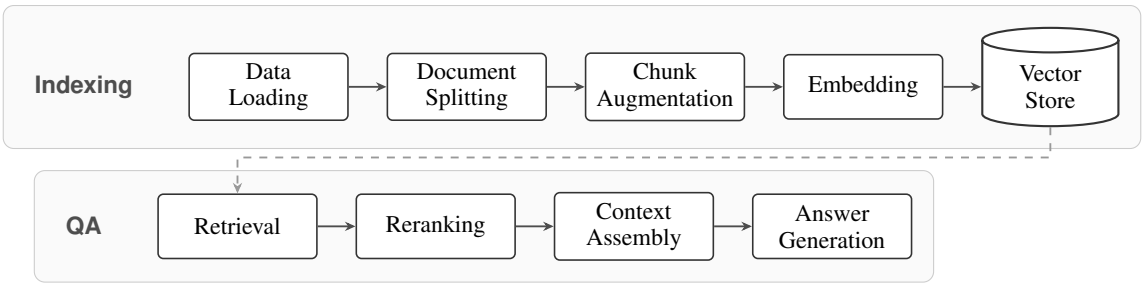


Figure 1: Two-phase RAG pipeline architecture. The dashed arrow indicates the vector store connecting the two phases.

nal evaluations, particularly in question-answering settings when supplied with relevant context. In this work, we do not focus on improving the language model itself through additional fine-tuning or parameter-efficient adaptation methods such as LoRA (Hu et al., 2021). Instead, we keep the LLM fixed and concentrate on improving retrieval and context construction, based on the assumption that modern instruction-tuned models are already sufficiently capable of producing correct answers when provided with well-retrieved and well-structured supporting context.

Metrics. The organizers scoring metric combines answer correctness (50%), document source identification (25%), and page proximity (25%). Since our ablation focuses on retrieval, we isolate the latter two components into separate metrics.

Document source accuracy ($Accuracy_{doc}$) measures correct document identification:

$$Accuracy_{doc} = \frac{1}{N}\sum_{i=1}^{N} d_i \tag{1}$$

where $d_i = \mathbf{1}(Doc_i^{\text{pred}} = Doc_i^{\text{true}})$.

Page accuracy ($Accuracy_{page}$) measures page-level proximity within correctly identified documents:

$$Accuracy_{page} = \frac{1}{N}\sum_{i=1}^{N} p_i \tag{2}$$

$$p_i = \begin{cases} 1 - \dfrac{|Page_i^{\text{pred}} - Page_i^{\text{true}}|}{n\_pages_i} & \text{if } d_i = 1, \\ 0 & \text{if } d_i = 0. \end{cases} \tag{3}$$

In addition, we report **page recall** ($Recall@k$) for correct page retrieval.

Hardware. All experiments run on a single NVIDIA RTX 5060 Ti with 16 GB of VRAM.

4.2 Embedding Model Selection

The embedding model is a critical component of the indexing pipeline because it determines how effectively relevant documents are retrieved during the initial search stage.

With a focus on computationally constrained environments, we selected the top embedding models with fewer than 2B parameters from the Multilingual RTEB leaderboard[3] and reevaluated them for Ukrainian on three retrieval tasks (Table 1) using our UNLP-QA dataset, along with Ukrainian subsets of the multilingual BelebeleRetrieval (Bandarkar et al., 2024) and WebFAQRetrieval datasets.

As a result, Snowflake Arctic Embed L v2.0[4] (568M parameters, 1024 dimensions) achieves the best balance across all three tasks.

4.3 Index Pipeline Optimization

Document Splitting. Drawing on prior work on document splitting strategies (Smith and Troynikov, 2024), we compared four main approaches: RecursiveCharacterTextSplitter (LangChain's default splitter) with varying chunk sizes and overlaps, SentenceSplitter (LlamaIndex's default splitter) at multiple chunk sizes, ClusterSemanticChunker (ChromaDB's proposed splitter) and full-page splitting.

In the results (Table 3 in the Appendix B), full-page splitting achieves the highest $Accuracy_{doc}$ and $Accuracy_{page}$ metrics, outperforming all sub-page strategies.

We attribute these findings to several factors.

First, in the organizer's dataset, each question's answer appears on a single page, so page-level chunks naturally preserve the source metadata needed for the evaluation metric, whereas sub-page

[3] http://mteb-leaderboard.hf.space/?benchmark_name=RTEB(beta)

[4] https://huggingface.co/Snowflake/snowflake-arctic-embed-l-v2.0

Model	BelebeleRetrieval				WebFAQRetrieval				UNLP-QA			
	$R@1$	$R@3$	$R@5$	$R@20$	$R@1$	$R@3$	$R@5$	$R@20$	$R@1$	$R@3$	$R@5$	$R@20$
Qwen3-Embedding-0.6B	82.67	92.11	93.22	97.44	61.19	73.80	77.80	87.21	53.44	72.27	79.96	94.94
Octen-Embedding-0.6B	82.44	92.33	94.00	97.11	62.17	74.34	78.25	87.74	54.25	74.09	80.97	95.75
Jina-Embed-v5-small	88.78	95.56	97.44	99.00	68.62	81.00	85.16	92.64	59.11	78.14	85.22	**97.17**
Arctic-Embed-L-v2.0	87.33	94.89	96.78	98.56	**71.71**	**83.73**	**87.30**	**93.99**	**65.18**	**80.57**	**86.44**	94.74
BGE-M3	88.00	95.67	97.56	99.00	69.26	81.39	84.92	93.01	61.34	78.34	85.43	94.74
Multilingual-E5-Large	**90.44**	**97.44**	**98.11**	**99.44**	71.04	82.67	86.29	93.59	61.34	78.75	85.43	96.15

Table 1: Embedding model comparison on three RTEB Ukrainian tasks. $R@k$ = $Recall@k$ (%).

splitting introduces ambiguity in page assignment when chunks span page boundaries.

Second, modern embedding models may handle longer contexts more effectively, making full-page chunks increasingly practical for such tasks.

Chunk Augmentation. Anthropic, in its engineering blog on contextual retrieval (Anthropic, 2024), showed that enriching chunks with document-level identifying context can improve their relevance during retrieval by making chunks easier to distinguish from similar ones.

In their approach, each chunk is enriched with additional contextual information generated from both the chunk itself and the broader source document. This method can significantly improve retrieval quality, although it is computationally expensive because it requires an LLM call for every chunk.

To make this approach suitable for computational-constrained environments, we adopted document-level summary augmentation, which has previously been used in RAG pipelines for legal-domain applications (Reuter et al., 2025). Instead of producing unique context for each chunk, we force the LLM to produce a short summary once per document, which is then prepended to all chunks derived from that document. This summary-based augmentation offers a favorable trade-off between cost and retrieval quality: it requires only a single LLM call per document, rather than one per chunk, while maintaining a distinction between similar chunks from different documents by grounding each chunk in its document-level context. Details are provided in Appendix A.

4.4 Question Answering Pipeline Optimization

Retrieval Strategies. To optimize the pipeline's retrieval step, various strategies can be incorporated, ranging from different search methods to advanced query transformation techniques. These methods generally aim to address the semantic differences between user queries and encoded document chunks.

In the strategy evaluation, we compared default dense retrieval, hybrid search (dense + BM25 sparse), Hypothetical Document Embeddings (HyDE) (Gao et al., 2023), multi-query expansion (Li et al., 2025), and their combinations.

We define the retrieval limit to the top 20 chunks as a pragmatic heuristic that provides comprehensive coverage of the relevant chunk space, and additional value beyond this threshold tends to degrade overall pipeline performance.

Default dense retrieval achieves the best results (Table 4 in Appendix B). Advanced query transformation and search methods do not improve performance on this dataset. We assume this is related to the nature of the questions: they are carefully hand-crafted with clear intent, unlike ambiguous real-user queries where such techniques typically excel. The evaluated techniques add noise rather than coverage for well-specified questions.

Reranking. For reranking optimization, we tested two main approaches: dedicated reranking models and LLM-based reranking such as listwise scoring (Ma et al., 2023) and pointwise logprob reranking.

The pointwise logprob method is based on adapting the monoT5 approach (Nogueira et al., 2020): for each document, we prompt the language model with the question-document pair, asking whether the document answers the question and requesting a "*Yes*" or "*No*" response. We extract token-level log-probabilities and compute relevance as:

$$s = \sigma\big(\log p(\text{Yes}) - \log p(\text{No})\big) \tag{4}$$

where σ is the sigmoid function. This yields a normalized score in $[0, 1]$ reflecting the model's confidence that the document is relevant.

The results (Table 5 in the Appendix B) did not indicate a consistently better-performing approach.

Without context augmentation, pointwise logprob reranking achieves the best $Recall@k$ results. With context augmentation, BGE Reranker v2-m3 outperforms all other approaches, achieving the highest $Accuracy_{doc}$, $Accuracy_{page}$, and $Recall@k$, though logprob reranking still delivers competitive results.

These findings suggest that a language model fine-tuned for Ukrainian can serve as an effective reranker, approaching the performance of dedicated cross-encoder models. This is particularly relevant in resource-constrained environments where deploying a separate reranking model is not feasible. In such cases, the same LLM can be used for both answer generation and reranking. With targeted fine-tuning on reranking objectives, these models could potentially be used without dedicated rerankers.

Context Assembly. After the reranking step, the top-ranked chunks are concatenated in order to form the final context passed to the language model for answer generation. For the competition submission, we used the top 3 chunks, which in our evaluations provided a balance between relevance page recall, input context compactness, and output generation time.

5 Results

Taking into account the results of all standalone RAG component optimization experiments, we selected two final configurations:

- **Approach A:** document-level contextual augmentation, full-page chunking, Arctic-Embed-L-v2.0 as the embedding model, and BGE Reranker v2-m3 as the reranker.
- **Approach B:** summary-based augmentation, full-page chunking, Arctic-Embed-L-v2.0 as the embedding model, and BGE Reranker v2-m3 as the reranker.

According to the evaluation metrics (Table 2), Approach A achieved the strongest overall performance. However, due to the environmental constraints of the UNLP 2026 Shared Task competition, we were unable to run Approach A in the final submission. Instead, we submitted Approach B, which ranked third on the private leaderboard.

Metric	Approach A	Approach B
$Accuracy_{doc}$	**99.78**	**99.78**
$Accuracy_{page}$	**95.35**	94.70
$Recall@1$	**78.96**	78.52
$Recall@3$	**93.49**	91.54
$Recall@5$	**95.23**	94.14
$Recall@10$	96.53	**96.75**
$Recall@20$	**97.61**	96.96

Table 2: Final system results on the development dataset. Approach A uses document-level contextual augmentation; Approach B uses summary-based augmentation.

6 Conclusion

We presented a modular RAG pipeline for the UNLP 2026 Shared Task on Ukrainian Multi-Domain Document Understanding, achieving 3rd place on the private leaderboard.

Our ablation study highlights several practical findings. For document splitting, full-page chunks consistently outperform sub-page strategies in both document-source accuracy and page-level recall because they naturally preserve the page-level metadata required by the evaluation metric. Chunk augmentation and its variations substantially improve retrieval quality across metrics by enriching chunks with additional document context, making it easier to distinguish them from similar ones. For retrieval, default dense search is most effective on this dataset, since the well-specified competition questions do not benefit from query transformation techniques such as HyDE or multi-query expansion. Among reranking approaches, dedicated cross-encoder models allow us to achieve the strongest overall performance, while LLM-based pointwise logprob reranking is competitive, suggesting that a Ukrainian-adapted language model can serve as an effective reranker in resource-constrained settings and may benefit from further fine-tuning for reranking objectives.

These results show that strong performance on the Ukrainian document-grounded question answering can be achieved through systematic isolated curation of RAG components, without additional language model adaptations. In future work, we plan to explore fine-tuning embedding and reranking models on Ukrainian data and to investigate approaches for visual document understanding.

Limitations

Despite these results, our work has several limitations.

First, we focused on the most impactful parts of the RAG pipeline while leaving aside more elaborate ETL and raw document loading techniques. This choice was partly motivated by the development dataset provided by the organizers, which does not contain corrupted or noisy data that would substantially affect evaluation results.

Second, the QA dataset does not include unanswerable questions. As a result, we cannot fully assess the distinction between cases where the LLM actually knows the answer and cases where it merely predicts one.

Third, the shared task evaluation does not distinguish between answers grounded in retrieved documents and answers generated from the LLM's internal knowledge. As a result, our system always produces an answer, even when the retrieved chunks are not sufficiently connected to the source document. This may lead to confidently incorrect outputs when retrieval fails.

Fourth, although we did not explicitly tune the system for the specific domains represented in the dataset, the final component selection may still have been influenced by the nature of the task and development datasets.

Finally, we do not claim that the proposed system will generalize reliably beyond the shared task environment. In real question-answering scenarios, user queries may be substantially more ambiguous and structurally complex than the benchmark questions used in this work. In real question answering systems, retrieval performance would likely require not only the indexing strategies explored in this work, but also additional query transformation techniques and hierarchical index architectures that align retrieval with domain-level understanding rather than tying it to specific pages or documents.

References

Anthropic. 2024. Introducing contextual retrieval. https://www.anthropic.com/engineering/contextual-retrieval. Anthropic Engineering Blog, accessed 2026-04-08.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Minghan Li, Xinxuan Lv, Junjie Zou, Tongna Chen, Chao Zhang, Suchao An, Ercong Nie, and Guodong Zhou. 2025. Query expansion in the age of pre-trained and large language models: A comprehensive survey. *Preprint*, arXiv:2509.07794.

Frank Liu, Kenneth Enevoldsen, Roman Solomatin, Isaac Chung, Tom Aarsen, and Zoltán Fődi. 2025. Introducing rteb: A new standard for retrieval evaluation. Hugging Face Blog. Accessed: 2026-04-23.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023. Zero-shot listwise document reranking with a large language model. *Preprint*, arXiv:2305.02156.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pre-trained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics:*

EMNLP 2020, pages 708–718, Online. Association for Computational Linguistics.

Yurii Paniv. 2025. Isolating LLM performance gains in pre-training versus instruction-tuning for mid-resource languages: The Ukrainian benchmark study. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 876–883, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Markus Reuter, Tobias Lingenberg, Rūta Liepiņa, Francesca Lagioia, Marco Lippi, Giovanni Sartor, Andrea Passerini, and Burcu Sayin. 2025. Towards reliable retrieval in rag systems for large legal datasets. *Preprint*, arXiv:2510.06999.

Brandon Smith and Anton Troynikov. 2024. Evaluating chunking strategies for retrieval. Technical report, Chroma.

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

Suhang Wu, Jialong Tang, Baosong Yang, Hongcheng Guo, Ruifeng Wen, Zhefeng Wang, Baotian Hu, and Min Zhang. 2024. Not all languages are equal: Insights into multilingual retrieval-augmented generation. *Preprint*, arXiv:2410.21970.

Hanna Yukhymenko, Anton Alexandrov, and Martin Vechev. 2025. Mamaylm v1.0: An efficient state-of-the-art multimodal ukrainian llm.

A Chunk Augmentation Details

This appendix describes the implementation of the two chunk augmentation strategies evaluated in Section 4. Both operate on chunks produced by the splitter and enrich each chunk with additional contextual information before it is passed to the embedding model. During the evaluation, we used Ukrainian versions of the prompts. In this paper, the prompts are presented as direct English translations.

A.1 Summary-based Augmentation

In this strategy, we prompt the language model once per document to produce a short summary of the document, based on two variables: `{char_length}`, a target upper bound on the summary length in characters, and `{document_content}`, a text representation of the document. The resulting summary is prepended to each chunk of that document.

`{document_content}` is built from a fixed window of chunks rather than the full document. Given the window size w, if the document has at most w chunks, all chunks are used. Otherwise, the first $\lfloor w/2 \rfloor$ and last $\lceil w/2 \rceil$ chunks are concatenated, with an explicit placeholder inserted between the head and tail to indicate how many middle chunks were omitted. It bounds the prompt length even for long documents while informing the model that part of the document is missing from the input.

In our final configuration, we use $w = 10$, `{char_length}` = 250, and align each chunk with a single document page.

Summary-based Prompt

You are an expert at document summarization.
Provide a brief summary of the given document text, no longer than `{char_length}` characters. Focus on extracting the most important entities, the main purpose, and key topics. The summary should be concise and optimized to provide context for smaller text fragments. Output only the summary text.
Document:
`{document_content}`

The prompt below illustrates the value of `{document_content}` variable, for the case when a 20-page document is rendered with windows size $w = 10$: the first 5 and last 5 pages are kept, and the omission placeholder is inserted between them.

Example value of `{document_content}` variable for a 20-page document with windows size $w = 10$

[Page 1]
<*text of page 1*>
[Page 2]
<*text of page 2*>
⋮
[Page 5]
<*text of page 5*>
[... PART OF THE DOCUMENT OMITTED ...]
(10 more pages exist between the start and end. When summarizing, take into account that the middle of the document is missing from the input.)
[Page 16]
<*text of page 16*>
⋮
[Page 20]
<*text of page 20*>

A.2 Document-Level Contextual Augmentation

In this strategy we prompt the language model once per chunk to produce a short piece of context that situates the chunk within its document, based on

two variables: `{chunk_content}`, the chunk being contextualized, and `{doc_content}`, the document text surrounding that chunk. The generated context is appended to the chunk's content.

Given the window size w, `{doc_content}` is built by using $\lfloor w/2 \rfloor$ chunks before contextualized chunk and $\lceil w/2 \rceil$ chunks after it. Near document boundaries, the window is shifted rather than shrunk, so that every prompt contains the same number of neighboring chunks. This avoids a size mismatch between the prompts at the start/end of a document and those in the middle.

In our final configuration, we use the window size $w = 4$. Compared to Summary Augmentation, this strategy is substantially more compute-intensive because the number of LLM calls scales with chunks rather than documents.

Document-Level Contextual Augmentation Prompt

We want to situate a chunk in the context of the entire document.
Document:
`{doc_content}`
Chunk:
`{chunk_content}`
Please give a short succinct context to situate this chunk within the entire document for the purposes of improving search retrieval of the chunk. Answer only with the succinct context and nothing else.

B Detailed Experimental Results

Splitting Strategy	$\boldsymbol{Accuracy_{doc}}$	$\boldsymbol{Accuracy_{page}}$	$R@1$	$R@3$	$R@5$	$R@20$
Recursive (600, 0)	89.70	80.47	53.94	69.70	76.77	87.68
Recursive (800, 0)	89.90	80.30	51.92	69.50	76.97	88.28
Recursive (1200, 0)	92.12	81.45	52.32	69.70	74.75	87.48
Recursive (1200, 300)	91.52	81.95	53.94	70.91	77.17	87.88
Recursive (800, 300)	89.70	80.24	53.94	70.91	**77.98**	**88.49**
Full Page	**95.56**	**84.64**	**54.55**	**71.11**	75.76	87.07
Sentence (600, 0)	92.73	83.22	53.33	67.48	74.34	86.87
Sentence (800, 0)	94.14	83.35	49.09	65.66	71.72	83.43
Sentence (1200, 0)	94.34	83.77	48.89	67.07	71.72	84.04
ClusterSemantic (1600, 500)	93.94	82.42	48.49	63.64	68.08	79.80
ClusterSemantic (1200, 300)	93.33	81.46	45.46	63.43	70.10	82.02

Table 3: Document splitting strategy comparison. Numbers in parentheses: (chunk size, overlap) for Recursive and Sentence; (max size, min size) for ClusterSemantic. $R@k$ = $Recall@k$ (%).

Retrieval Strategy	$Accuracy_{doc}$	$Accuracy_{page}$	$R@1$	$R@3$	$R@5$	$R@20$
Without chunk augmentation						
Dense (default)	**95.56**	**84.64**	**54.55**	**71.11**	75.76	87.07
Hybrid (dense + sparse)	86.67	77.22	50.30	71.11	75.56	88.08
HyDE	91.31	81.44	54.34	70.91	**76.97**	**88.28**
HyDE + Hybrid	77.37	68.65	45.45	68.69	76.36	88.08
Multi-Query	94.34	82.64	47.88	67.27	72.73	85.86
Multi-Query + Hybrid	81.41	71.62	43.84	60.81	70.10	88.89
With chunk augmentation (document-level contextual augmentation)						
Dense (default)	**98.18**	88.68	**63.03**	**85.25**	**89.29**	96.16
Hybrid (dense + sparse)	92.53	84.42	60.00	82.83	88.08	**96.36**
HyDE	97.98	**89.15**	62.83	85.05	88.89	96.16
HyDE + Hybrid	90.10	80.26	54.95	79.39	85.05	95.15
Multi-Query	97.98	88.33	58.99	80.20	88.08	95.96
Multi-Query + Hybrid	86.87	77.29	50.51	73.33	82.22	97.17

Table 4: Retrieval strategy comparison with and without chunk augmentation. $R@k$ = $Recall@k$ (%).

Reranking Model	$Accuracy_{doc}$	$Accuracy_{page}$	$R@1$	$R@3$	$R@5$	$R@10$
Without chunk augmentation						
Without reranking	**95.56**	84.64	54.55	71.11	75.76	81.21
Logprob Reranker	88.48	83.38	**67.07**	**79.60**	**83.23**	**86.46**
Qwen3-Reranker-0.6B	88.08	80.68	57.78	74.14	78.18	83.84
Jina Reranker v3	89.49	81.94	60.61	74.75	78.99	83.23
LLM Reranker (listwise)	93.87	84.65	57.55	71.99	78.12	82.28
ContextualAI-Rerank-v2-1B	91.92	83.82	60.81	75.56	78.99	85.05
BGE Reranker v2-m3	93.13	**85.94**	64.24	77.58	80.81	85.05
With chunk augmentation (document-level contextual augmentation)						
Without reranking	98.18	88.68	63.03	85.25	89.29	93.74
Logprob Reranker	98.92	94.92	**80.26**	91.76	94.36	96.31
Qwen3-Reranker-0.6B	95.76	87.89	65.05	86.87	91.11	94.95
Jina Reranker v3	96.97	90.04	68.69	85.45	89.70	94.14
LLM Reranker (listwise)	98.73	90.48	68.53	83.25	88.83	91.88
ContextualAI-Rerank-v2-1B	97.78	92.42	72.93	89.49	91.92	94.95
BGE Reranker v2-m3	**99.78**	**95.35**	78.96	**93.49**	**95.23**	**96.53**

Table 5: Reranking model comparison with and without chunk augmentation. $R@k$ = $Recall@k$ (%).

The UNLP 2026 Shared Task on Multi-Domain Document Understanding

Volodymyr Sydorskyi[1,3], Nataliia Romanyshyn[2], Roman Kyslyi[3], Olena Nahorna[4],
[1]National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
[2]Texty.org.ua
[3]Kyiv School of Economics
[3]Preply
v.sydorskyi@kpi.ua, nataliia.romanyshyn@texty.org.ua, rkyslyi@kse.org.ua lena.kobzar@gmail.com

Abstract

This paper presents the results of the UNLP 2026 Shared Task on Multi-Domain Document Understanding. This Shared Task aims to challenge and assess AI capabilities to find the right information in a stack of domain-specific documents and generalize across domains. Participants were required not only to select the correct answer, but also to localize it by predicting the corresponding document and page. A total of 54 teams registered for the competition, 15 teams submitted systems, and 513 runs were evaluated on a hidden test set via Kaggle in a code-only submission format under constrained computational resources. The Kaggle leaderboard is left open for further submissions. Summarizing the contributions of this work, we establish a Ukrainian multi-domain document understanding benchmark, which consists of: (1) a collected dataset; (2) a proposed evaluation metric; and (3) an analysis of top-performing systems evaluated under a unified framework.

1 Introduction

Working with long and complex documents remains a challenging problem in natural language processing (NLP). While recent large language models (LLMs) have made strong progress on question answering tasks (Ma et al., 2025), they still struggle when required to search through document collections and reliably point to the exact source of the answer (Huang et al., 2025). Moreover, comprehensive benchmarks for evaluating document understanding capabilities remain limited for mid-resource languages such as Ukrainian, where underrepresentation in model pretraining and scarce training resources make robust system development challenging.

To address these challenges, the Fifth Ukrainian NLP Conference (UNLP 2026) organized a Shared Task on Multi-Domain Document Understanding[1]. The goal of the task is to evaluate systems that can both retrieve relevant information from domain-specific documents and use it to answer questions. Unlike standard question answering benchmarks, participants were also required to indicate where the answer comes from by predicting the corresponding document and page.

The shared task is formulated as a multiple-choice question answering problem over collections of domain-specific PDF documents. Given a question and six candidate answers, systems are required to:

1. identify the correct answer from six options
2. specify which document contains the answer
3. pinpoint the exact page number where the answer is located

The dataset spans multiple domains with distinct structures and writing styles, including sporting competition rules, medical product instructions, and military regulations. The military domain was kept hidden from participants to test how well systems generalize to previously unseen data.

The competition was hosted on Kaggle[2] in a code-only submission format under fixed computational constraints, encouraging participants to build efficient and reproducible solutions. A total of 54 teams registered, of which 15 actively participated, producing 513 submissions. Participants explored a range of approaches, from relatively simple hybrid retrieval pipelines to more complex multi-stage systems with reranking and instruction-tuned language models.

The main contributions of this shared task are as follows:

- Introduced a dataset for document understanding in Ukrainian, covering three domains, including one hidden.

[1] https://unlp.org.ua/shared-task/

[2] https://www.kaggle.com/t/3ab59dd1807746c99d0a5c3b72580a8b

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 249–259

- Proposed an evaluation framework for assessing system performance on multi-domain document understanding.
- Provided a comprehensive analysis of top-performing Shared Task systems, evaluated under a unified framework using the proposed dataset and metric.

The remainder of this paper is organized as follows. Section 2 reviews previous work. Section 3 outlines the UNLP 2026 shared task setup, presents the dataset and describes the evaluation metric. Section 4 reports the leaderboard results and summarises the submitted systems. Section 5 concludes the paper, while Section 5 provides an ethics statement, and finally, current limitations are stated.

2 Related Work

Document understanding and question answering (QA) over complex sources have been widely studied in NLP.

Early benchmarks such as SQuAD (Rajpurkar et al., 2016) established reading comprehension as a core NLP task, but focused on short Wikipedia passages. TAT-QA (Zhu et al., 2021) extended this to financial documents combining text and tables, while MultiReQA (Guo et al., 2021) highlighted the difficulty of cross-domain retrieval-based QA. In the visual document understanding space, DocVQA (Mathew et al., 2021a) introduced QA over single-page industry documents, later extended to the multi-page setting by MP-DocVQA (Tito et al., 2023). SlideVQA (Tanaka et al., 2023) expanded multi-page document understanding to slide decks, requiring evidence selection across multiple images alongside answer generation. DUDE (Landeghem et al., 2023) further introduced multi-domain and cross-domain generalization as explicit evaluation goals across visually-rich documents from diverse industries and origins. Our shared task builds on these ideas by combining answer selection with document and page localization but applies them to text-based multiple-choice QA over domain-specific PDF documents in Ukrainian.

Benchmarks for document understanding remain scarce for Ukrainian. Recent efforts have focused on building large language models such as Lapa (Paniv et al., 2025) or MamayLM (Yukhymenko et al., 2025), and Syromiatnikov and Ruvinskaya (2024) explored context-based QA using zero-shot and few-shot LLMs on general-domain texts. The UNLP workshop series has progressively built evaluation infrastructure for Ukrainian: previous shared tasks addressed LLM instruction-tuning (Romanyshyn et al., 2024) and social media manipulation detection (Kyslyi et al., 2025).

The present task extends this line with a more complex scenario: multi-domain document understanding with answer localization over domain-specific PDFs, including a hidden domain to test cross-domain generalization.

Several evaluation frameworks have been proposed for measuring performance in document retrieval and question answering tasks. Mean Reciprocal Rank (MRR) (Voorhees, 1999) and Normalized Discounted Cumulative Gain (NDCG) (Järvelin and Kekäläinen, 2002) are widely used in information retrieval to assess ranking quality, but they do not account for localization within a retrieved document. RAGAS (Es et al., 2024) introduced a suite of reference-free metrics for retrieval-augmented generation, covering faithfulness and answer relevance, yet it does not evaluate page-level precision. In the document understanding space, metrics used in DocVQA (Mathew et al., 2021b) and MP-DocVQA (Tito et al., 2023) focus primarily on answer correctness via Average Normalized Levenshtein Similarity (ANLS), without penalizing incorrect source attribution. Our proposed metric addresses this gap by jointly rewarding answer accuracy, document retrieval correctness, and page-level localization in a single unified score, making it better suited for multi-document, multi-domain settings where source traceability is essential.

3 Task Description

The main goal of the shared task was to build a system capable of retrieving the correct answer to a particular question from multiple-choice options (6 options) based on a set of documents. In addition, the developed system should also provide localization of the corresponding information within the document. Thus, the task can be decomposed into two goals:

- Find the correct answer to a multiple-choice question
- Identify the document and page where the answer was found

Based on the obtained dataset, the system was evaluated across three domains of PDF documents in Ukrainian. Additionally, the third domain was completely hidden from participants during the system development stage, which was intended to ensure robustness of the system to new domains.

3.1 Data

The dataset consists of three types of open-source documents in Ukrainian: sporting competition rules, medical product instructions, and military regulations. In addition to the documents, it includes a set of multiple-choice questions associated with them, along with references to the specific document pages from which the questions were derived (see Appendix B for a data sample example). All documents were provided in PDF format. Each domain was accompanied by a README file describing the content of the documents in both English and Ukrainian (see an example in Appendix A). The documents were sourced from official government resources, such as the Official Portal of the Ministry of Youth and Sports of Ukraine[3], the regulatory documents website of the Ministry of Health of Ukraine[4], and the official web portal of the Parliament of Ukraine[5].

The **sporting rules** are long documents that define competition regulations, including gameplay rules, organizational procedures, participant requirements, officiating, penalties, and integrity measures. Some documents also include ethical provisions and adaptations for people with disabilities. Appendices often contain tables and diagrams.

The **medical instructions** are shorter texts that describe the safe and effective use of medicinal products. They follow a standardized format covering composition, pharmacology, indications, contraindications, dosage, side effects, storage conditions, and regulatory and manufacturer information.

The **military regulations** are extensive documents that define the principles, structure, and functioning of Ukraine's defense sector. They include strategic-level documents outlining defense policy and reform priorities, as well as statutes of the Armed Forces that regulate military service.

3.2 Data Annotation Process

In order to generate questions, we adopted a hybrid approach: a pool of volunteers created questions with multiple-choice answers, and additionally, we generated questions using GPT-4o mini. The latter were validated by volunteers as well as professional Ukrainian linguists to ensure grammatical correctness and eliminate hallucinations. Specifically, we designed an annotation process to assess several aspects: (1) whether the question is accurate, (2) whether the answer options are plausible, and (3) whether the referenced page is correct. For more details, see Appendix C.

Annotators assessed each question along with answers as valid, flagged it for removal, or marked it as uncertain. As 65% of the data was annotated by professional linguists, their judgements dominated the overall data quality profile. Professional linguists applied markedly stricter standards than volunteers, both in terms of rejection and correction. As shown in Table 1, linguists rejected items at nearly twice the rate of volunteers (33.6% vs. 19.4%), a statistically significant difference ($z = 3.24$, $p < 0.01$). Among items approved as valid, linguists further edited at least one field—question phrasing, correct answer, or a page number—in 82.7% of cases, compared to 63.3% for volunteers ($z = 3.77$, $p < 0.001$). These results suggest that professional linguists not only removed a larger share of items but also actively improved the quality of those they retained. The comparison is based on a single domain-matched subset reviewed by both groups ($n = 412$), ensuring that any observed differences in annotation quality are not confounded by domain variation; therefore, the results should be treated as exploratory.

As a result, we obtained a dataset that was partitioned into three subsets: `train` (or `dev`), `public test`, and `private test`. The dataset statistics are presented in Table 2. It is important to note that questions were not written for all documents (as illustrated in the fourth column of Table 2). Additional documents in the training set can be used for system optimization, while additional documents in the test set can be used for more robust system validation.

[3] https://mms.gov.ua
[4] https://mozdocs.kiev.ua/
[5] https://zakon.rada.gov.ua/

Group	Total items	Drop rate	95% CI (drop rate)	Valid items	Correction rate (valid)
Professional linguists	226	33.6%	[27.5%, 39.8%]	150	82.7%
Volunteers	186	19.4%	[13.7%, 25.0%]	147	63.3%

Table 1: Annotation quality metrics by annotator group. Drop rate: proportion of items flagged for removal out of all reviewed items. Correction rate (valid): proportion of approved items with at least one edited field (question, correct answer, or page number).

Domain	Split	#Docs	#Docs w Qs	Avg. len.	#Qs
Sport	train	11	11	78.64	20
	public	19	19	78.21	319
	private	46	46	76.83	704
Medical	train	30	20	8.53	20
	public	50	50	8.26	464
	private	120	106	8.32	942
Military	private	5	5	103	500

Table 2: Dataset statistics.

3.3 Evaluation

$$\begin{aligned}\text{Metric} = 0.5\frac{1}{N}\sum_{i=1}^{N} a_i \\ + 0.25\frac{1}{N}\sum_{i=1}^{N} d_i + 0.25\frac{1}{N}\sum_{i=1}^{N} p_i\end{aligned} \tag{1}$$

$$\begin{aligned}a_i = \mathbb{I}\Big(&\text{Correct\_Answer}_i^{(\text{pred})} \\ &= \text{Correct\_Answer}_i^{(\text{true})}\Big)\end{aligned} \tag{2}$$

$$\begin{aligned}d_i = \mathbb{I}\Big(&\text{DocID}_i^{(\text{pred})} \\ &= \text{DocID}_i^{(\text{true})}\Big)\end{aligned} \tag{3}$$

$$\begin{aligned}p_i = &\left(1 - \frac{\left|\text{Page\_Num}_i^{(\text{pred})} - \text{Page\_Num}_i^{(\text{true})}\right|}{\text{n\_pages}_i}\right) \\ &\cdot \mathbb{I}(d_i = 1)\end{aligned} \tag{4}$$

Final evaluation metric[6] is shown on Equation 1. And it is composed of three terms:

- Answer accuracy - Equation 2
- Document correctness - Equation 3
- Page correctness - Equation 4

The metric is constructed in such a way that its upper bound is equal to one, while the lower bound is theoretically unbounded due to the p_i term. However, by applying a straightforward rule that limits the number of predicted pages for a particular document by the maximum number of pages (n_pages_i), the minimum value is also effectively bounded by zero.

Exploring the first term - Answer accuracy - it is assigned the largest coefficient. This is motivated by the fact that the two subsequent terms reflect the system's localization ability, and we treat them as a single second goal of the task. It is also important to mention that accuracy can be affected by class imbalance; however, since the distribution of correct choices can be controlled, it can always be adjusted to be approximately uniform.

Document correctness is required to evaluate how accurately the system can select the relevant document. For example, this is important in cases where the system needs to retrieve information about a drug that has a very close substitute. It also directly affects the next term of our metric - Page correctness - because selecting any page from an incorrect document is not meaningful.

Finally, Page correctness reflects how close the predicted page is to the correct page, but only if the document has been selected correctly. To measure this closeness, $1 -$ Relative Absolute Error at the page level was used. The motivation for this choice is as follows: using a strict indicator function would be too restrictive, since information relevant to a particular question may be distributed across several pages, or the exact page may not be predicted correctly. In such cases, localization near the correct page is still a more desirable outcome than predictions far from it. Regarding the use of absolute error instead of a quadratic penalty, this choice was made to keep the metric simpler and because it is not obvious how strongly outliers should be penalized.

[6] https://www.kaggle.com/code/vladimirsydor/unlp-2026-document-understanding-metric

Rank	Team	Public Score	Private Score
1	GA	**0.9460**	**0.9598**
2	Bullseye Emoji	0.9211	0.9420
3	Dialogus ex Machina	0.9402	0.9411
4	Golden Retrievers	0.8876	0.9191
5	Daria Belozor	0.8715	0.8802
6	lawandia	0.8528	0.8775
7	PFW	0.8688	0.8722
8	OM	0.8376	0.8314
9	OksanaTkach	0.7720	0.8210
10	catdlia	0.7831	0.8095

Table 3: Leaderboard for UNLP 2026 Shared Task. Final ranking is based on private leaderboard scores; public scores are shown for comparison.

3.4 Data Split and Competition Setup

The Shared Task was conducted on the Kaggle platform[7] using a special type of competition - a Code Competition. This competition format allows all testing data to be hidden from participants, including both target values - correct answer choices, document IDs, and page numbers - and input data - questions, documents, and domain descriptions. This feature is critical for document understanding and information retrieval shared tasks, as it explicitly limits participants from:

- manually finding answers and localizations for test questions,
- overfitting the system to particular documents or questions,
- overfitting the system to all domains available in the test set.

Additionally, all data from the Military domain was placed in the private test set, so participants were not able to explicitly track performance on it via the public leaderboard and ranking. The first two domains were split document-wise in order to avoid data leakage. Detailed data split statistics can be found in Table 2.

4 Results and System Descriptions

In this section, detailed metric results from the top-ranking team, together with a brief description of the top two systems on the public leaderboard, are discussed.

4.1 Overall Results Summary

Metric scores obtained by the top 10 teams on the public and private test sets are presented in Table 3. The overall average absolute deviation between public and private scores is approximately 0.018, with private scores being slightly higher. This deviation is relatively small, indicating consistent performance across the two evaluation settings. Team rankings are also largely preserved, with only minor one-position swaps observed between three pairs of teams.

The results presented in Table 4 reveal several consistent tendencies across domains, metrics, and evaluation settings. First, performance on the private test set is generally higher than on the public test set, indicating that the systems do not exhibit significant overfitting to the public subset and are able to generalize reasonably well to unseen data. This trend is particularly evident for the Sport domain, where improvements are observed across all metrics.

Across domains, the Medical domain demonstrates the highest overall performance, with consistently strong results for all metrics. This suggests that the corresponding documents are either more structured or easier to interpret for retrieval and localization tasks. In contrast, the Sport domain appears more challenging, especially in terms of document and page correctness, indicating potential ambiguity in document selection and localization.

Analyzing the decomposed metrics, answer accuracy is generally high and stable across domains, confirming that systems are effective at selecting the correct answer once relevant information is identified. Document correctness also achieves strong performance, particularly in the Medical and Military domains, suggesting reliable document retrieval. However, page correctness consistently lags behind the other metrics, highlighting the difficulty of precise localization within documents. This gap indicates that, while systems can often identify the correct document, accurately pinpointing the exact page remains a more challenging task.

Finally, the Military domain, available only in the private test set, shows competitive performance levels, which suggests that the systems are robust to domain shifts despite the absence of explicit exposure during development. Overall, the observed trends indicate that the primary bottleneck of current approaches lies in fine-grained localization rather than answer selection or document retrieval.

[7] https://www.kaggle.com

Team	Metric		Answer		Document		Page	
	Pub	Priv	Pub	Priv	Pub	Priv	Pub	Priv
Sport								
GA	0.9195	**0.9481**	0.9279	**0.9574**	**0.9310**	**0.9560**	**0.8914**	**0.9216**
Bullseye Emoji	0.8601	0.9163	0.8840	0.9403	0.8558	0.9091	0.8166	0.8753
Dialogus ex Machina	**0.9198**	0.9220	**0.9310**	0.9176	**0.9310**	0.9503	0.8860	0.9024
Golden Retrievers	0.7675	0.8254	0.8683	0.8991	0.6897	0.7741	0.6435	0.7292
Daria Belozor	0.7981	0.8452	0.8777	0.9048	0.7524	0.8196	0.6845	0.7515
Medical								
GA	0.9629	0.9642	0.9547	**0.9628**	**0.9871**	**0.9841**	0.9551	0.9471
Bullseye Emoji	0.9607	0.9634	0.9569	**0.9628**	0.9763	0.9756	0.9528	0.9523
Dialogus ex Machina	0.9533	0.9528	0.9461	0.9597	0.9828	0.9735	0.9382	0.9186
Golden Retrievers	**0.9703**	**0.9655**	**0.9677**	**0.9628**	**0.9871**	0.9830	**0.9587**	**0.9535**
Daria Belozor	0.9220	0.9075	0.9483	0.9406	0.9440	0.9151	0.8474	0.8337
Military								
GA	–	**0.9684**	–	0.9540	–	**0.9860**	–	**0.9795**
Bullseye Emoji	–	0.9381	–	0.9500	–	0.9300	–	0.9223
Dialogus ex Machina	–	0.9462	–	0.9340	–	0.9700	–	0.9469
Golden Retrievers	–	0.9638	–	**0.9580**	–	0.9760	–	0.9630
Daria Belozor	–	0.8785	–	0.8980	–	0.8820	–	0.8359

Table 4: Performance across 5 top teams, domains, and evaluation metrics. Public (Pub) and private (Priv) scores are reported. Best values per column are highlighted in bold.

4.2 First Place Solution

The winning team created a system that works in three stages: chunks documents, retrieves relevant passages, and then generates an answer (Bazdyrev et al., 2026). First, documents are split into overlapping chunks while keeping important context intact (tables are flattened into text, and each chunk gets a prefix identifying which document it came from). For every multiple-choice option, the system builds an instruction-style query and runs it through dense retrieval using Qwen3-Embedding-8B, pulling the top 20 candidates. A fine-tuned Qwen3-8B reranker then narrows these down to just the top 2 passages. The final answer comes from Qwen3-32B-AWQ, which uses constrained decoding so it can only output a valid option label (A through F). The document and page predictions are simply taken from passage in which reranker scored highest. To train the reranker, the team combined Ukrainian QA data (FIdo-AI/ua-squad) with translated synthetic datasets built from sources like MS MARCO and HotpotQA. Everything was designed to run within Kaggle's hardware limits (two T4 GPUs), relying on vLLM with quantization and trimmed context lengths to fit.

4.3 Second Place Solution

The second place system takes a two-stage approach, pairing hybrid retrieval with a fine-tuned generation model (Nosenko and Kilko, 2026). Documents are first converted to Markdown, then ranked at the document level using both semantic embeddings (computed over the opening text) and BM25. When the two methods disagree, their candidate lists are merged and reranked. Next, the system zooms in to the page level within the chosen document: it chunks pages, scores them with embeddings and BM25, blends the results using Reciprocal Rank Fusion, and applies a custom reranker on top. To boost performance, the team generated roughly 7,000 synthetic QA examples from the training data and used them to fine-tune MamayLM with LoRA. The model learns to predict both the answer option and the page number in one shot (outputting something like "A 2"). For deployment, the model is quantized and packaged in GGUF format, running through llama.cpp. At inference time, it gets the top 3 retrieved pages as context to work with.

5 Conclusion

We believe that the UNLP 2026 Shared Task on Multi-Domain Document Understanding is instrumental in facilitating research on retrieval-augmented question answering for Ukrainian language documents. The task introduced a novel evaluation scenario requiring systems to simultaneously select the correct answer from six options, identify the source document, and pinpoint the exact page, evaluated under a unified metric that jointly rewards answer accuracy and localization quality. The use of a hidden military domain further tested cross-domain generalization under realistic constraints, revealing the ability of top systems to

adapt to previously unseen document types.

A total of 54 teams registered, 15 submitted systems, and 513 runs were evaluated on Kaggle under fixed computational constraints. Teams explored a variety of techniques, from hybrid retrieval combining dense embeddings and BM25 to cross-encoder reranking, synthetic data generation for fine-tuning, and post-answer page verification and demonstrated the creative potential of the NLP research community when working in low-resource settings. Top-performing systems employed models such as Qwen3-32B, Lapa, and MamayLM alongside multilingual retrievers like BGE-M3. All final solutions were required to be released under an open license, promoting reproducibility and accessibility. The Kaggle leaderboard remains open for further submissions.

We hope this shared task will serve as a foundation for future work in Ukrainian NLP, and that the tools, data, and approaches developed through this competition will continue to support progress in document understanding and information retrieval for low-resource languages.

Ethics Statement

To ensure fair competition and support the development of reproducible and transparent solutions, the shared task was conducted under a set of clearly defined rules governing data use, system design, and participant behavior.

By participating in the shared task, all teams agreed to comply with the competition rules. In particular, participants were required to avoid any form of unfair advantage, including leaderboard probing or attempts to infer hidden test data. The hidden domain used for evaluation was explicitly protected, and any effort to retrieve or approximate its content through indirect means was strictly prohibited.

Participants were required to use open-source models in their final solutions, proprietary models were permitted only for data generation purposes. The use of external data was allowed provided the corresponding licenses permitted research use. We encouraged participants to use Ukrainian-specific language models such as Lapa and MamayLM. Final solutions were required to be released under an open license, promoting reproducibility and accessibility for the research community.

The dataset used in the shared task consists of publicly available documents, and no personal or sensitive data was included. Questions were generated and validated through a combination of automated methods and human review, with additional validation to ensure correctness and minimize potential errors.

Limitations

We identify the following limitations of our work:

- The dataset covers only three domains, which consist of relatively well-structured and typical documents. To make the benchmark more challenging and representative of real-world scenarios, it should be extended to a broader range of domains with less structured data, such as newspapers, blogs, and social media posts.
- The proposed evaluation metric was not compared with existing metrics for document retrieval in terms of correlation with subjective or human-centered evaluation criteria. As a result, it lacks sufficient empirical justification. A more thorough comparative analysis would strengthen the validity of the proposed metric.
- The collected benchmarks are primarily aimed at maximizing the proposed metric rather than encouraging a broader diversity of methodological approaches. Consequently, many solutions rely on similar techniques. Expanding the benchmark to promote methodological diversity, rather than optimization of a single metric, would further strengthen the study.

Acknowledgments

We are grateful to the volunteers who assisted in question generation and validation and to Preply that sponsored the validation of the dataset by professional Ukrainian linguists. We also thank the Kaggle platform for hosting the competition infrastructure.

AI-assisted tools (ChatGPT and Claude) were used exclusively to improve the clarity and grammar of this text; they did not contribute to the research design, experiments, or analysis.

References

Anton Bazdyrev, Ivan Bashtovyi, Ivan Havlytskyi, Oleksandr Kharytonov, and Artur Khodakovskyi. 2026. Qwen goes brrr: Off-the-shelf rag for ukrainian multi-domain document understanding. In *Proceedings of*

the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), Lviv, Ukraine. Association for Computational Linguistics. To appear.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th conference of the european chapter of the association for computational linguistics: system demonstrations*, pages 150–158.

Mandy Guo, Yinfei Yang, Daniel Cer, Qinlan Shen, and Noah Constant. 2021. MultiReQA: A cross-domain evaluation forRetrieval question answering models. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 94–104, Kyiv, Ukraine. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

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

Jordy Van Landeghem, Rafał Powalski, Rubèn Tito, Dawid Jurkiewicz, Matthew Blaschko, Łukasz Borchmann, Mickaël Coustaty, Sien Moens, Michał Pietruszka, Bertrand Ackaert, Tomasz Stanisławek, Paweł Józiak, and Ernest Valveny. 2023. Document understanding dataset and evaluation (dude). In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19471–19483.

Chuangtao Ma, Yongrui Chen, Tianxing Wu, Arijit Khan, and Haofen Wang. 2025. Large language models meet knowledge graphs for question answering: Synthesis and opportunities. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24578–24597, Suzhou, China. Association for Computational Linguistics.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021a. Docvqa: A dataset for vqa on document images. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2199–2208.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021b. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Mykola Nosenko and Pavlo Kilko. 2026. Rag pipeline strategies for ukrainian multi-domain document understanding task. In *Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026)*, Lviv, Ukraine. Association for Computational Linguistics. To appear.

Yurii Paniv, Bohdan Didenko, Mykola Haltiuk, Vladyslav Humennyy, Andrian Kravchenko, Roman Kyslyi, Viktoriia Makovska, Artem Orlovskyi, Bohdan Ruban, Maksym-Yurii Rudko, Anastasiia Senyk, Nazarii Drushchak, Dmytro Chaplynskyi, and Mariana Romanyshyn. 2025. Lapa LLM v0.1.2 — the most efficient Ukrainian open-source language model.

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

Mariana Romanyshyn, Oleksiy Syvokon, and Roman Kyslyi. 2024. The UNLP 2024 shared task on fine-tuning large language models for Ukrainian. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 67–74, Torino, Italia. ELRA and ICCL.

Mykyta V. Syromiatnikov and Victoria Ruvinskaya. 2024. Ua-llm: Advancing context-based question answering in ukrainian through large language models. *Radio Electronics, Computer Science, Control*.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: a dataset for document visual question answering on multiple images. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2023. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834.

Ellen M Voorhees. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82.

Hanna Yukhymenko, Anton Alexandrov, and Martin Vechev. 2025. Mamaylm v1.0: An efficient state-of-the-art multimodal ukrainian llm.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287, Online. Association for Computational Linguistics.

A Domain Description Example

This appendix provides an example of a domain description (README file) accompanying the dataset.

This folder contains the rules of sporting competitions for various sports. All documents are written in Ukrainian. The average length of a document is 77 pages.

Each document provides a clear and structured presentation of the rules governing the conduct of competitions, as well as ensuring fairness and the safety of participants.

The main sections of a competition rules document include the following information:

- general points describing the scope of application of the rules;
- terms and abbreviations used in the rules;
- rules of games and competitions, often taking into account different disciplines and formats of the sport concerned;
- competition organization processes and competition regulations;
- requirements for competition venues, equipment, and facilities;
- requirements for competition participants and qualification criteria;
- rules for team formation in team competitions;
- medical supervision of competition participants;
- officiating, scoring, and determination of winners;
- violations of competition rules, penalties, and disqualification;
- participants' protests and appeals against officials' decisions;
- anti-doping rules and measures;
- measures to prevent corruption in competitions.

The rules of some sports also include ethical provisions and adaptations for war veterans, people with disabilities, and individuals with conditions that limit their daily activities.

The documents may include diagrams and tables, most of which are placed in appendices.

The rules of sporting competitions serve as a primary source of information for organizers, officials, and participants, ensuring unified standards and a shared understanding of the rules of play.

B Data Format Example

This appendix illustrates the structure of a single data row from the dataset, clarifying which fields are available to participating systems as input, which must be predicted, and which are technical (additional) ones.

Field Roles

- **Input fields** (visible to the system): `Question_ID`, `Question`, `A`, `B`, `C`, `D`, `E`, `F`
- **Target fields** (output of the system): `Correct_Answer`, `Doc_ID`, `Page_Num`
- **Additional fields**: `Domain`, `n_pages`

C Question Validation Guidelines

This appendix describes the validation procedure for generated questions used in the *UNLP 2026 Shared Task on Multi-Domain Document Understanding*.

C.1 Task Overview

Participants are provided with sets of PDF documents, where each set corresponds to a particular domain. In addition to the PDFs, each set includes two text files, in Ukrainian and English, describing the domain.

For each domain, a set of multiple-choice questions is prepared in a standard quiz format, consisting of one question and six answer options. Participants are required to build a system that can: (i) identify the correct answer, (ii) indicate the exact document and page where the answer was found, and (iii) perform well on these tasks regardless of the domain or document length.

The domains are as follows:

- **domain_1**: rules of sporting competitions across various sports approved by the Ministry of Youth and Sports of Ukraine;
- **domain_2**: medical product instructions;
- **domain_3**: currently undisclosed.

Field	Value (Ukrainian)	Value (English)
Input fields		
Question_ID	0	0
Question	Що означає термін «у грі» на змаганнях з регбі?	What does the term “in play” mean at rugby competitions?
A	гравця позначили маленькою позначкою ’X’	the player was marked with a small ‘X’ marker
B	гравець перебуває в положенні, що дає можливість брати участь у грі	the player is in a position that allows them to participate in the game
C	гравець перетнув лінію вкидання	the player has crossed the throw-in line
D	гравець покараний за положення «поза грою»	the player was penalized for being in an offside position
E	удар ногою, призначений на користь невинної в порушенні команди	a kick awarded in favor of the team not at fault for the infringement
F	гравець опинився ближче 10-ти метрів від суперника	the player ended up within 10 meters of an opponent
Additional fields		
Domain	domain_1	domain_1
n_pages	79	79
Target fields		
Correct_Answer	B	B
Doc_ID	759dfc8486c0f02391d7 cfc1fed753b0608fc601.pdf	759dfc8486c0f02391d7 cfc1fed753b0608fc601.pdf
Page_Num	43	43

Table 5: Example data row split by field role.

C.2 Validation Task

Validators are asked to review each generated question and mark the corresponding value in the Validated? field. The following criteria should be taken into account.

C.2.1 Question Formulation

- The wording of the question should allow one to identify unambiguously the document that contains the answer. For example, it is too broad to ask about clothing in karate in general, since multiple types of karate are represented in the collection. Instead, the question should specify a particular discipline, such as kyokushin karate competitions.

- The question must be answerable using only the content of the given document. No external resources, including internet sources, should be needed.

- Overly trivial questions should be removed. For instance, a question such as “What active substance is contained in caffeine sodium benzoate?” is not informative if the answer simply repeats the name of the product.

- If the name of the medication or sport is omitted from the question, it should be added whenever necessary to avoid ambiguity.

C.2.2 Answer Options

- The answer to each question must be located on a single page of the document. Accordingly, the Page_num field should contain exactly one page number.

- If the relevant information is repeated on multiple pages, such questions should preferably be excluded.

- If the relevant context spans the boundary between two pages, it is better not to formulate a question based on that passage.

- Each question must have exactly one correct answer and five incorrect answers.

- Incorrect answers should be plausible. They may be constructed on the basis of the same document. For example, in a question about active ingredients in a medication, excipients listed in the same instruction may serve as distractors.

- One of the incorrect answers may be intentionally nonsensical in order to test models for hallucinations.
- GPT-generated questions and answers may occasionally contain phrasing errors, lexical mistakes, or other inaccuracies. Such cases should be corrected during validation.

C.2.3 Page Numbering

- In the `Page_num` field, make sure the page number corresponds to the one in the PDF reader rather than the page number printed inside the document itself, if such numbering is present.
- In longer documents, GPT may assign incorrect page numbers. Please correct these errors whenever they are detected.

C.3 Practical Note

For convenience, the generated questions are grouped by document and page. This allows one to work with one document at a time and review all associated questions efficiently.

Author Index