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

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

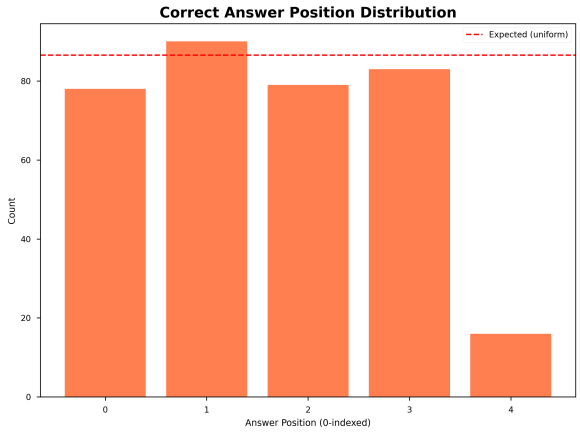

Figure 1: Distribution of correct answer positions across the five answer choices (A corresponds to 0, while the Ukrainian letter Д (equivalent to D) corresponds to 4). The red dashed line indicates the expected count under a uniform distribution. Most positions are close to the expected frequency, while position 4 appears somewhat underrepresented.

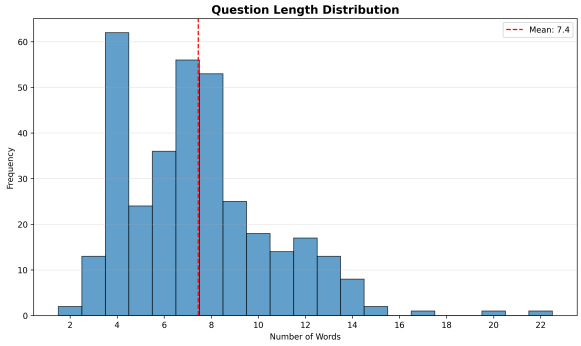

Figure 2: Distribution of question lengths in the benchmark, measured by number of words per question. The red dashed line marks the mean question length.

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.