O. Pasichnyk, PhD in Ed.
K. Sadykov, Student
Khmelnytskyi National University

NOVEL MIXED-APPROACH LINGUISTIC BENCHMARK FOR THE UKRAINIAN LANGUAGE

Abstract. This theoretical work presents a novel benchmarking system for evaluating large language models (LLMs) on the Ukrainian language. The proposed benchmark, *uastbench*, is designed to quantify an LLM's ability to adhere to Ukrainian grammar and spelling norms, as well as to respond to prompts using the appropriate language and context adherence. *uastbench* utilizes the LLM-as-a-judge approach, as well as proofreading software in evaluating responses. The rating system encompasses a percentage rating for four different categories, the average of which is considered as the uastbench score.

1. Introduction. Large language models (LLMs) have become an integral part of modern natural language processing technologies. However, most existing benchmarks for evaluating LLMs are primarily focused on the English language, leaving other languages, including Ukrainian, largely unaddressed. This gap in evaluating LLMs for the Ukrainian language creates a need for specialized benchmark that takes into account the nuances of this language and adheres to its norms.

The proposed benchmark, termed "*uastbench*", aims to fill this gap by comprehensively evaluating an LLM's ability to adhere to modern Ukrainian grammar and spelling norms, as well as to respond to prompts using the appropriate language and context adherence by using a mix of the LLM-as-a-judge technique [1], as inspired by AlpacaEval [2], and a more traditional proofreading software, such as LanguageTool. These abilities are critically important for ensuring effective human-LLM interaction in various scenarios involving the use of the Ukrainian language.

uastbench differs from existing benchmarks by evaluating the input on its linguistic characteristics, and not on factual accuracy. This allows for a better understanding of the models' Ukrainian language knowledge, as opposed to more traditional linguistic benchmarks, such as the Open-Ko-LLM leaderboard [3], which measures performance on datasets analogous to the common English benchmarks, such as Ko-ARC, Ko-TruthfulQA, Ko-MMLU, etc., representative of ARC [4], TruthfulQA [5], and MMLU [6] respectively. A different approach was taken in response to the observed deficiencies in a subset of less advanced contemporary language models, which frequently exhibit inaccuracies in Ukrainian language processing, including misspellings and conflation with Russian, as well as various additional issues, making them unfeasible for many potential production environments. *uastbench* uses four rating categories, each with well-defined criteria to improve objectivity and repeatability. The said categories are as follows:

- Spelling Accuracy, or *spell* (proofreading software)
- Grammatical Correctness, or *grmmr* (proofreading software)
- Prompt Adherence, or *prmpt* (LLM judge)
- Fluency, or *fluent* (LLM judge)

The proposed benchmark aims not only to provide an objective comparison of existing LLMs for the Ukrainian language but also to foster further development and

improvement of these models. The availability of a specialized evaluation tool will enable developers to identify weaknesses in their models and focus on addressing them, ultimately leading to enhanced quality of LLMs for the Ukrainian language.

The prompts for the benchmarked models are published in the project's **GitHub repo** (See at: https://github.com/int3rrobang/uastbench). A website with the benchmark results is also located at the **repository page** (See at: https://int3rrobang.github.io/uastbench/).

2. Methodology

2.1. LLM-as-a-judge. The methodology of this subset of *uastbench* is based on the principle of employing a language model as a judge to evaluate the quality of responses from another language model. The LLM-as-a-judge evaluation process proceeds as follows: the test language model receives an input prompt in Ukrainian and generates a response. This response is then provided to the judge model along with the input prompt. The judge model, utilizing its embedded knowledge of the Ukrainian language and the context of the prompt, evaluates the response across the two categories using this technique by assigning one of the following ratings: "Unsatisfactory", "Fair", "Average", "Good", and "Excellent". These categories encompass the aforementioned criteria. Each category has a separate set of definitions for the aforementioned ratings. Ratings are converted to percentage values as follows:

Unsatisfactory	0%
Fair	25%
Average	50%
Good	75%
Excellent	100%

Table 1. Values converted into numerical values

We assign a numerical grade to each output and average them out to get a final score for each respective category.

The judge used is Anthropic's claude-3-haiku-20240307 [7]. All the prompts are written in the Ukrainian language. The judge prompts have detailed criteria for each ranking and, crucially, utilize few-shot prompting (Tom B. Brown et al.) [10]. The full prompts are available in the GitHub repository. A crucial aspect of the methodology is the formation of a representative set of prompts for the evaluated model that covers a diverse range of topics and contexts, allowing for an objective assessment of the language model's ability to handle the Ukrainian language across various scenarios. The prompts (n=100) have been either written manually or generated by gpt-4-turbo [7], cherrypicked and modified as necessary.

2.2. Proofreading software. We use proofreading software to rate the following two categories: Spelling Accuracy and Grammatical Correctness. This approach was taken due to cost, as well as the highly insufficient performance on this task even by SOTA LLMs, such as OpenAI's gpt-4-turbo-1106 [7] and Anthropic's claude-3-opus-20240229 [9] as compared to more traditional tools such as LanguageTool [8], which

offers support for the Ukrainian language. For both categories, the baseline score is 100 percent, with each mistake subtracting 15 percentage points, capped by the lower bound of 0%. As with the former category of benchmarks, each text output is evaluated on both *spell* and *grmmr*, and the score of the model is determined by averaging the score of each result. The spell category includes the "misspelling" and "typographical" type errors and the latter encompasses the rest.

2.3. Final result. The final result of the benchmark, or the *uastbench* score, is determined by taking the average of the values obtained from all four benchmarks: spell, grmmr, prmpt, and fluent. Each of these four categories is given equal weight in the final *uastbench* score, reflecting the importance of both linguistic accuracy (as measured by *spell* and *grmmr*) and contextual appropriateness (as measured by *prmpt* and *fluent*) in evaluating an LLM's proficiency in the Ukrainian language. As such, the results between various models are directly comparable so long as the same judge model and spellcheck software is used.

3. Limitations of the approach **and further discussion.**

Cosine similarity between the output of the model being benchmarked and a SOTA model might be used as a cheaper yet potentially more biased alternative to the prompt adherence category.

Attempting to adapt the approach of this study to other, less widespread languages can be challenging due to the necessity of traditional proofreading software for two of the five categories.

It must also be noted that the capabilities of the LLM-as-a-judge approach are inherently limited by the judge model in question. This study uses Anthropic's claude-3-haiku-20240307 [9] as the judge model; however, other models might produce more accurate results.

This benchmark also doesn't measure cultural knowledge. Moreover, a model performing well in a benchmark designed for evaluating performance in Ukrainian may or may not be indicative of a diverse training set and, as such, good general multilingual performance - a potential relationship out of scope of this paper.

REFERENCES

1. Lianmin Zheng, Wei-Lin Chiang, Ying Sheng et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv preprint arXiv:2306.05685.

2. Xuechen Li, Tianyi Zhang, Yann Dubois et al. (2023). AlpacaEval: An Automatic Evaluator of Instruction-following Models. GitHub repository.

3. Open Ko-
LLM Leaderboard. Retrieved from: https://huggingface.co/spaces/upstage/open-ko-llm-leaderboard

4. Peter Clark, Isaac Cowhey, Oren Etzioni et al. (2018). Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. arXiv preprint arXiv:1803.05457.

5. Stephanie Lin, Jacob Hilton, Owain Evans (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods arXiv preprint arXiv:2109.07958.

6. Dan Hendrycks, Collin Burns, Steven Basart et al. (2021). Measuring Massive Multitask Language Understanding. arXiv preprint arXiv:2009.03300.

7. OpenAI et al. (2024). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.

8. LanguageTool – GitHub repository. Retrieved from: https://github.com/languagetool-org/languagetool

9. Anthropic. Claude 3 – Introduction. Retrieved from: https://www.anthropic.com/news/claude-3-family

10. Tom B. Brown et al. (2020) Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165.