# BelarusianGLUE: Towards a Natural Language Understanding Benchmark for Belarusian

**Maksim Aparovich**[1], **Volha Harytskaya**[2], **Vladislav Poritski**[2], **Oksana Volchek**[2], **Pavel Smrz**[1]
[1] Brno University of Technology, Brno, Czech Republic
[2] Independent researcher, Vilnius, Lithuania
**Correspondence:** belarusianglue@gmail.com

## Abstract

In the epoch of multilingual large language models (LLMs), it is still challenging to evaluate the models' understanding of lower-resourced languages, which motivates further development of expert-crafted natural language understanding benchmarks. We introduce BelarusianGLUE — a natural language understanding benchmark for Belarusian, an East Slavic language, with ≈15K instances in five tasks: sentiment analysis, linguistic acceptability, word in context, Winograd schema challenge, textual entailment. A systematic evaluation of BERT models and LLMs against this novel benchmark reveals that both types of models approach human-level performance on easier tasks, such as sentiment analysis, but there is a significant gap in performance between machine and human on a harder task — Winograd schema challenge. We find the optimal choice of model type to be task-specific: e.g. BERT models underperform on textual entailment task but are competitive for linguistic acceptability. We release the datasets[1] and evaluation code.[2]

## 1 Introduction

Recent advances in NLP, such as large language models (LLMs) based on transformer architectures (Vaswani et al., 2017), have had groundbreaking impact on the field. LLMs are projected to bring significant economic effect in the future (Eloundou et al., 2023), however, in the first place it is anticipated to benefit the largest language communities, such as people speaking English or Chinese (Xie and Avila, 2024). For smaller language communities, especially those of vulnerable languages, state-of-the-art NLP tools may help preserve and promote the linguistic and cultural heritage (Mohanty et al., 2024), maybe even (under favorable circumstances) stimulate language revival, given that enough effort is put into improving the multilingual capabilities of LLMs. The case of Belarusian, an East Slavic language, illustrates this point.

In the UNESCO World Atlas of Languages, Belarusian is characterized as "potentially vulnerable".[3] According to the 2019 population census data, 54% of the residents of Belarus consider Belarusian their native language but only 26% speak it at home.[4] Based on statistical analysis of the census data, it was argued that the true proportion of Belarusian-speaking residents of Belarus is in fact even lower (Sokolov, 2022). In an earlier study, 4% respondents in urban areas of Belarus claimed to be using standard Belarusian, possibly with some Russian words, as their primary language of communication, while 41% reported using substandard, mixed Belarusian–Russian varieties (Kittel et al., 2010). Despite its official status and symbolic importance, Belarusian is a de facto minority language in its home country (Zaprudski, 2007).

To advance computational support of a particular language — such as Belarusian — in the epoch of LLMs, it is crucial to have (1) training data, (2) task-specific fine-tuning data, and (3) evaluation data for this language available in the open. In the case of Belarusian, as shown below, evaluation datasets are the most glaring omission; multilingual benchmarks that include tasks in Slavic languages, such as the recent EU-20 set of benchmarks (Thellmann et al., 2024), can only serve as a proxy indicator of the models' performance in Belarusian. To address this issue, we introduce BelarusianGLUE, the first natural language understanding benchmark for Belarusian, modeled after GLUE-type benchmarks for other languages. It includes five novel expert-crafted datasets.

[1] https://hf.co/datasets/maaxap/BelarusianGLUE
[2] https://github.com/maaxap/BelarusianGLUE
[3] https://en.wal.unesco.org/countries/belarus/languages/belarusian
[4] https://census.belstat.gov.by/saiku/?guest=true&lang=en#query/open//public/F503N_en.saiku



*Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–527
July 27 - August 1, 2025 

The rest of this paper is structured as follows. Section 2 is a review of existing datasets, containing Belarusian data, and natural language understanding benchmarks. Section 3 is a detailed description of BelarusianGLUE, its guiding principles and the datasets included. In section 4, we measure human performance on BelarusianGLUE and compare it with the performance of BERT models and LLMs. Discussion and conclusion follow.

## 2 Related work

### 2.1 Belarusian data in multilingual datasets

Starting from (Buck et al., 2014), Belarusian texts are available in Common Crawl-based massively multilingual corpora: OSCAR (Ortiz Suárez et al., 2019), CC-100 (Conneau et al., 2020; Wenzek et al., 2020), mC4 (Xue et al., 2021), CulturaX (Nguyen et al., 2023), and HPLT (de Gibert et al., 2024) that also includes data from the Internet Archive's crawls. In each of the above, the amount of Belarusian texts ranges from several dozen million to several billion tokens, which is two–three orders of magnitude less than Russian, two orders less than Polish, one order less than Ukrainian, on par with e.g. Kazakh, Armenian, or Icelandic.

Belarusian is not entirely lacking training data in other modalities than text: e.g. Common Voice (Ardila et al., 2020) includes 1873 hours of speech recordings in Belarusian, as of March 2025.

The amount of task-specific data for Belarusian is smaller, and their quality is generally lower. As an example, consider machine translation. Among the parallel corpora available in OPUS (Tiedemann, 2012),[5] the largest ones that include Belarusian data are NLLB (NLLB Team et al., 2022) derived from Common Crawl and ParaCrawl (Bañón et al., 2020), bilingual HPLT and two other Common Crawl-based corpora: CCMatrix (Schwenk et al., 2021) and CCAligned (El-Kishky et al., 2020). We labeled random samples of 100 Belarusian–English aligned sentence pairs from each corpus, following the taxonomy of Kreutzer et al. (2022), and found the ratios of natural, correctly translated sentences to be 17%, 41%, 7%, and 31% respectively in NLLB, HPLT, CCMatrix, and CCAligned.[6]

Instruction tuning is another example. Upadhayay and Behzadan (2024) introduced a version of Alpaca-52K and Dolly-15K instruction tuning datasets machine-translated into 132 languages, including Belarusian. Although the overall quality of the Belarusian translations is high, there is still some noise in translations of English-specific tasks, code snippets, rare words, etc.

Evaluation datasets for Belarusian are scarce: available for some of the more traditional NLP tasks, such as POS tagging and dependency parsing, e.g. in Universal Dependencies (Shishkina and Lyashevskaya, 2022), but not available for most tasks related to natural language understanding, thus providing no guidance for future models supporting Belarusian. The situation has begun to improve only recently: e.g. the question-answering benchmark INCLUDE (Romanou et al., 2024) contains several hundred instances in Belarusian.

### 2.2 GLUE-type benchmarks

Language understanding benchmarks emerged as a way of assessing transformer models' capabilities to understand linguistic structure above the word level and apply this understanding in downstream tasks. The benchmarks are often based on pre-existing datasets and cover a wide range of tasks, from sentiment analysis to question answering and beyond. It is common to frame the tasks as classification, although other types of tasks, e.g. sequence tagging, may also be included in the benchmark.

Most of the GLUE-type benchmarks created to date are monolingual. While the original GLUE and SuperGLUE focused on English (Wang et al., 2019a,b), their influence has since expanded through the development of benchmarks for many genetically and typologically diverse languages, such as Chinese (Xu et al., 2020), Arabic (Elmadany et al., 2022), or Hungarian (Ligeti-Nagy et al., 2024). Russian and Ukrainian, the two East Slavic languages most closely related to Belarusian, are covered by RussianSuperGLUE (Shavrina et al., 2020) and Eval-UA-tion (Hamotskyi et al., 2024) respectively; one more benchmark, MERA (Fenogenova et al., 2024), specifically targeting LLMs, has recently been proposed for Russian. Another neighboring and related language, Polish, has two comprehensive benchmarks: KLEJ (Rybak et al., 2020) and LEPISZCZE (Augustyniak et al., 2022). Language understanding datasets have also been created for smaller Slavic languages, such as Bulgarian (Hardalov et al., 2023) or Slovene (Žagar and Robnik-Šikonja, 2022). Multilingual benchmarks XGLUE (Liang et al., 2020) and XTREME-R (Ruder et al., 2021) include tasks for Slavic languages, however, Belarusian is missing from both.

[5] https://opus.nlpl.eu

[6] The labeled samples can be viewed here.

| Dataset ID | Task | Instance type | No. instances | | |
|---|---|---|---|---|---|
| | | | train | dev | test |
| BeSLS | sentiment analysis | sentence | 1500 | 250 | 250 |
| BelaCoLA | acceptability prediction | sentence | 1992 | 800 | 800 |
| BeWiC | word sense disambiguation | word + sentence pair | 5626 | 400 | 400 |
| BeWSC | coreference resolution | sentence / sentence pair | 570 | 200 | 200 |
| BeRTE-WD | textual entailment | sentence pair | 1080 | 360 | 360 |

Table 1: Datasets summary.

## 3 BelarusianGLUE

BelarusianGLUE is a natural language understanding benchmark that includes five novel expert-crafted datasets (summarized in Table 1), with all tasks formulated as binary classification. Sample instances from each dataset are shown in Table 4 in Appendix A. All Belarusian examples below are given in romanized spelling.

### 3.1 Guiding principles

During the benchmark development we adhered to the following principles:

• **Prefer quality over size.** The material was selected from representative sources for each dataset (see the descriptions below). When necessary to construct or label linguistic data in Belarusian, this was done by at least two fluent speakers of Belarusian with a background in linguistics (M.A. or Ph.D. degree). Each dataset was thoroughly reviewed.

• **When possible, leverage existing resources.** For example, Wikidata properties and labels in Belarusian were leveraged to build a textual entailment dataset, with careful attention given to gender and cultural balance (representation of women, Belarusian objects, etc.). Some of our datasets are modeled after similar resources in other languages: e.g., to construct the train set of a Belarusian Winograd schema challenge (WSC), English instances were translated, incorporating insights from the corresponding Russian dataset and adapting examples where needed. This approach minimized effort and maximized the quality and consistency of new datasets while respecting the unique characteristics of the target language.

• **Take into account the specifics of Belarusian.** While sampling linguistic data from various sources, we tried to reflect the variability of modern Belarusian language. E.g., the distribution of orthographic variants in the sentiment analysis dataset reflects the real-world diversity of written Belarusian: most sentences follow the official modern orthography (*narkamaŭka*), some — less than 10% — use the classical orthography (*taraškievica*), and a tiny minority is written in Latin script (*łacinka*).

• **Embrace open licensing.** While the sources are variously (and not always permissively) licensed, we made sure that none of the passages borrowed or derived from copyrighted work are longer than one sentence. We believe such use of copyrighted material for scholarly purposes to qualify under fair use or similar provisions in most legislations, thus allowing us to publish the novel datasets under an open license.

### 3.2 Tasks

#### 3.2.1 Sentiment analysis

BeSLS is a small dataset of sentiment-labeled Belarusian sentences, partially inspired by a similar English dataset from (Kotzias et al., 2015).

**Data sources:** The sentences were sampled from newspaper articles, reviews posted by the users of online shopping and booking platforms, messages in thematic Telegram channels and other social media, such as Mastodon. Five domains are covered: movie reviews, book reviews, hotel and travel reviews, consumer product reviews, social media posts.

**Methodology of data selection and processing:** In multilingual sources, non-Belarusian sentences were filtered out using Lingua.[7] Following Petrović et al. (2010), we anonymized user mentions in Mastodon posts. The sentences were manually tagged for sentiment polarity using a two-stage approach: initial labeling by one expert followed by comprehensive review and refinement by another. This results in 100% agreement, as either both labelers agree on the label, or the instance is not included in the dataset.

**Dataset structure:** The dataset contains 2000 sentences with positive or negative polarity. The classes are balanced: 50% positive and 50% negative, none of the sentences are neutral. Sentences are equally distributed over domains: 300 per domain in the train set, 50 in the dev and test sets.

[7] https://github.com/pemistahl/lingua-py

#### 3.2.2 Linguistic acceptability

BelaCoLA is a small-scale Belarusian corpus of linguistic acceptability, similar to CoLA (Warstadt et al., 2019) and RuCoLA (Mikhailov et al., 2022), with some inspiration also taken from BLiMP (Warstadt et al., 2020).

**Data sources:** We used five major sources of data to create the corpus:

1) sentences from Russian linguistic publications included in RuCoLA, manually translated into Belarusian and reviewed (we made sure to keep only instances with acceptability judgments transferable from the original Russian sentences to their Belarusian translations, due to deep similarities between Russian and Belarusian grammar);
2) contexts from Belarusian language textbooks and other normative sources;
3) sentences from the Belarusian section of Common Voice project, evaluated as unacceptable by speakers of Belarusian participating in the project;
4) passages produced by lightweight, non-state-of-the-art language models, i.e. hallucinations;
5) outputs of machine translation models.

**Methodology of data selection and processing:** All unacceptable sentences in the corpus were taken from the sources "as is" or with minor simplifications. Unlike the original CoLA, they exemplify not only morphological, syntactic, and semantic violations, but also certain pragmatical anomalies, prescriptive rule violations, and errors produced by language models, such as hallucinations and machine translation errors, which don't always fall neatly into a single category. Their corresponding acceptable sentences were extracted from the sources (if available) or, more typically, constructed by the experts — three of the paper's authors. For example, a hallucinated sentence *Jana navat źlohku zachvalavaŭsia i vyjšaŭ* 'She even got.3SG.M a little excited and left.3SG.M' is transformed to *Jon navat źlohku zachvalavaŭsia i vyjšaŭ* 'He even got.3SG.M a little excited and left.3SG.M' by correcting the gender agreement.

**Dataset structure:** The dataset contains 3592 sentences tagged as acceptable or unacceptable. The class balance is close to 50 : 50. The sentences have been randomly shuffled. Sentences translated from Russian, extracted from Belarusian textbooks and Common Voice data constitute the in-domain set, split into train/dev/test sets. Hallucinations and machine translations constitute the out-of-domain set, split into dev/test sets.

#### 3.2.3 Word in context

BeWiC is a Word-in-Context dataset for Belarusian, similar to the original WiC (Pilehvar and Camacho-Collados, 2019) and RUSSE (Shavrina et al., 2020, section 3.1.2). It can be viewed as a version of word sense disambiguation task.

**Data sources:** The dataset is based on the Explanatory Dictionary of Belarusian (*Tłumačalny słoŭnik biełaruskaj movy*, 1977–1984, vol. 1–5). While a newer dictionary exists (*Tłumačalny słoŭnik biełaruskaj litaraturnaj movy*, 1996, revised 2022), our choice of the older source was deliberate based on several advantages: broader lexical coverage, machine-readable accessibility and, crucially, availability of illustrative contexts.

**Methodology of data selection and processing:** For most words and word senses, the dictionary provides usage examples — phrases or sentences. To make each context one sentence long, we expanded phrases to full sentences by finding suitable contexts on the web or constructing them from scratch. E.g., the phrase *abarvać śpiełyja jahady* 'to pick ripe berries' was expanded to *My abarvali śpiełyja jahady* 'We've picked the ripe berries', and *hruntoŭny adkaz* 'a profound answer' to *Hruntoŭny adkaz vučnia ŭściešyŭ nastaŭnika* 'The student's profound answer made the teacher happy'.

Each instance in the dataset is a pair of contexts $c_1$, $c_2$ containing the target word $w$. The contexts refer either to the same word sense of $w$ or to two different homonyms of $w$. An instance is positive if both $c_1$ and $c_2$ refer to the same word sense of $w$, and negative if $c_1$ and $c_2$ refer to two different homonyms of $w$ (possibly belonging to different parts of speech), which are listed separately in the dictionary with their respective word senses. This is a stronger distinction than in WiC, so that less instances can be constructed from the dictionary data but they are easier to solve for humans and therefore don't require pruning.

**Dataset structure:** The dataset contains 6426 instances. The dev and test sets contain 400 instances each, half of them positive and half negative. None of the sentences repeat across instances in the dev and test sets, and each target word is represented by $\leq 3$ instances. The training set contains all positive and negative instances that can be constructed from the remaining sentences.

#### 3.2.4 Winograd schema challenge

BeWSC is a Belarusian version of the Winograd schema challenge, WSC (Levesque et al., 2012).

The dataset is available in two flavors: WSC proper, formatted as in SuperGLUE, and WNLI, formatted as in GLUE, i.e. converted into an NLI task. The number of instances and their (randomized) ordering are the same in both variants.

**Data sources:** Most of the training instances have been manually translated into Belarusian from the standard English dataset, WSC-285[8]; when adaptation was not possible, we translated those items from the similar Russian dataset RWSD (Shavrina et al., 2020, section 3.1.4) that were created specifically to replace unsuitable English sentences. The dev and test instances are based on or inspired by contexts from fiction books in Belarusian, available on the web.

**Methodology of data selection and processing:** Issues in English → Belarusian translation of the training instances are typically caused by differences in grammar, such as the grammaticalization of gender in Belarusian, or the reflexive pronoun *svoj*, which is equivalent to English possessive pronouns in certain contexts but doesn't have ambiguous reference. In such cases, the sentences were adapted to maintain the overall meaning of the original while altering its grammatical structure and wording.

The dev and test sentences were sampled (with modifications) from a corpus of Belarusian fiction books: we split the texts into sentences using `sentence-splitter`,[9] added morphological tags using `beltagger`,[10] extracted all sentences with a personal pronoun and at least two distinct nouns that precede it and have matching gender/number, then processed the output manually. These instances are intended to be hard to solve by selectional restrictions. Not all of them are Google-proof, as some sentences follow the source contexts rather closely.

**Dataset structure:** The training set has 570 instances, the dev and test sets have 200 instances each. Half of the instances are positive (the antecedent is correct), and half negative.

#### 3.2.5 Textual entailment

BeRTE-WD is a small-scale textual entailment dataset for Belarusian, derived from Wikidata.[11]

**Data sources:** To produce the sentences, we extracted all statements from a June 2024 dump of Wikidata such that: (1) the property relates an entity to a timestamp, a number, or another entity; and (2) all entities in the statement, i.e. one or both, have Belarusian labels available.

**Methodology of data selection and processing:** Each instance in the dataset is a pair of sentences in Belarusian, denoted "text" ($t$) and "hypothesis" ($h$); $t$ is said to entail $h$ if, typically, a human reading $t$ would infer that $h$ is most likely true (Dagan et al., 2006). Unlike many of the standard benchmarks, such as SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018), or XNLI (Conneau et al., 2018), we don't distinguish between contradictory and neutral pairs, so the labels are binary: entailment or non-entailment.

For each of the three value types, as described above, we manually sampled 200 diverse statements. Three fluent speakers of Belarusian then transformed the statements into texts and wrote two hypotheses per text: one entailed, one non-entailed. Additional texts and hypotheses were produced from the same statements grouped into pairs.

The entailed hypotheses have at their core a wide range of phenomena, including but not limited to timestamp or numeric comparison, reasoning about time intervals, conversion of units, domain-specific or world knowledge, logical consequence, paraphrasing, etc. A non-entailed hypothesis is typically produced by modifying the entailed one to make its claim contrary or neutral w.r.t. the text.

**Dataset structure:** The dataset contains 1800 instances. Train/dev/test split was obtained by grouping the statement pairs belonging to each value type into 60 : 20 : 20, so that none of the source statements would overlap between the samples. The dataset is balanced by class (half of the instances entailed, half non-entailed) and by value type (equal counts of timestamps, numbers, and entities).

### 3.3 Evaluation

For BeSLS, BeWiC, BeWSC and BeRTE-WD, accuracy is reported. For BelaCoLA, Matthews correlation and accuracy are reported. Although there have been attempts to evaluate model performance on various datasets with a single metric, such as Krippendorff's alpha (Berdicevskis et al., 2023), or a unified set of metrics, such as accuracy, precision, recall, and F1 (Augustyniak et al., 2022), these choices are still not very popular, so we use the most common evaluation metric(s) for each dataset type, to ensure compatibility of our results with

[8] https://cs.nyu.edu/~davise/papers/WinogradSchemas/WSCollection.html
[9] https://pypi.org/project/sentence-splitter
[10] https://github.com/volchek/beltagger
[11] https://www.wikidata.org

related work for other languages.

## 4 Experiments

### 4.1 Human baseline

We evaluated human performance on each dataset for comparison with the performance of NLP models. The usual procedure, outlined e.g. in (Wang et al., 2019a, §5.2), involves recruiting paid crowdworkers, who are first provided with task-specific instructions, then asked to label a few dozen dev set instances as a pre-screening, and then proceed to annotate a sample of test set instances. Due to the low availability of crowdworkers fluent in Belarusian, we recruited unpaid volunteers from the language community and followed a simplified version of the above procedure. We wrote task-specific instructions including labeled examples from the train set (3 positive + 3 negative) and 5 self-check instances from the dev set. A random sample of 100 instances, balanced by class and certain other parameters (such as in-domain / out-of-domain in BelaCoLA, PoS of the target word in BeWiC, etc.), was then extracted from the test set and split into 5 groups of 20 instances with approximately the same balance of classes in each group. 5 volunteers, all of them competent and fluent speakers of Belarusian, were invited to tag the samples. Each volunteer tagged 20 instances per dataset, i.e. 100 instances in total, without overlaps (a single label obtained per instance). A balanced Latin square (Bradley, 1958) was used to reduce order effects while presenting the samples to volunteers.

Human baseline scores are shown in Table 2. BeSLS gold labels are in near-perfect agreement with speaker judgments, while in all other datasets the agreement is around 90%. The human baseline in BelaCoLA is lower than in other tasks but this difference in scores isn't statistically significant, given the sample size. It might reflect the unique linguistic situation of Belarusian with two codified standards (*narkamaŭka* and *taraškievica*), so that even competent speakers show variation in grammatical form usage.

| Dataset | Metric | Score |
|---|---|---|
| BeSLS | acc | 0.99 |
| BelaCoLA | acc / MCC | 0.87 / 0.754 |
| BeWiC | acc | 0.91 |
| BeWSC | acc | 0.91 |
| BeRTE-WD | acc | 0.89 |

Table 2: Human baseline scores.

### 4.2 BERT baselines

We fine-tuned mBERT (Devlin et al., 2019), XLM-R base (Conneau et al., 2020), mDeBERTa-v3 (He et al., 2023) and a recent monolingual model — Belarusian HPLT BERT (Samuel et al., 2023; de Gibert et al., 2024) on each of the five training sets separately. All tasks were formulated as binary classification: in particular, BeWSC was presented in WNLI format, i.e. as a sentence pair classification task. Evaluation scores for BelaCoLA in-domain and out-of-domain instances were calculated separately. The models were fine-tuned for 5 epochs with learning rate 2e-5, batch size 16 on a single GeForce RTX 4090 GPU. Snapshots were saved once per epoch, and the best snapshot for evaluation was selected by the dev set accuracy.

The results are shown in Table 3. Since mDeBERTa-v3 has the strongest overall performance, we experimented with additional pretraining of this model on Belarusian texts from HPLT 1.2 (deduplicated version[12] with custom filtering applied), adjusting the script published by the model authors.[13] This brought some more performance gains. Further experiments are targeting this enhanced version of mDeBERTa-v3.

With the size of our training datasets ranging between several hundred and a few thousand instances, it may be beneficial to freeze a subset of the model's layers while training the classifier on top of it (Grießhaber et al., 2020). We tried progressively freezing the layers of mDeBERTa-v3 during fine-tuning: only the embeddings or 3, 6, 9, 12 layers in addition to the embeddings. As shown in Table 3, the model's performance on BelaCoLA and BeWiC goes up but is sensitive to the number of layers frozen: in particular, prediction quality drops abruptly when all layers are frozen and only the classification head is trained. Also, freezing layers doesn't help to beat the trivial baseline on BeWSC and BeRTE-WD.

We also tried transfer learning by mixing our training data with instances from larger datasets:

- Sentiment analysis: all positive and negative instances no longer than 500 characters with self-confidence score at least 0.7 were taken from the train folds of MMS (Augustyniak et al., 2023) for all languages in it. The classes were balanced per

[12] https://data.hplt-project.org/one/monotext/deduplicated/be_map.txt

[13] https://github.com/microsoft/DeBERTa/blob/master/experiments/language_model/rtd.sh

| Model | Dataset: Metric: | BeSLS acc | BelaCoLA i.d. acc / MCC | BelaCoLA o.o.d. acc / MCC | BeWiC acc | BeWSC acc | BeRTE-WD acc |
|---|---|---|---|---|---|---|---|
| mBERT | | 0.712 | 0.490 / -0.029 | 0.510 / 0.029 | 0.515 | 0.500 | 0.517 |
| XLM-R | | 0.816 | 0.510 / 0.033 | 0.506 / 0.019 | 0.500 | 0.500 | 0.503 |
| HPLT BERT be | | 0.912 | 0.480 / -0.048 | 0.576 / 0.167 | 0.575 | 0.495 | 0.506 |
| mDeBERTa-v3: | | | | | | | |
| *Original* | | 0.892 | 0.623 / 0.260 | 0.742 / 0.488 | 0.613 | 0.500 | 0.511 |
| *With pre-train* | | | | | | | |
| Frozen: | no layers | 0.916 | 0.620 / 0.274 | 0.784 / 0.589 | 0.678 | 0.500 | 0.522 |
| | only emb. | **0.928** | 0.610 / 0.234 | 0.826 / 0.666 | 0.678 | 0.500 | 0.514 |
| | emb. + 3 | 0.916 | 0.640 / 0.297 | 0.828 / 0.666 | 0.685 | 0.500 | 0.514 |
| | emb. + 6 | 0.924 | 0.633 / 0.322 | 0.772 / 0.570 | 0.690 | 0.500 | 0.517 |
| | emb. + 9 | 0.916 | 0.657 / 0.362 | 0.788 / 0.590 | 0.710 | 0.490 | 0.494 |
| | emb. + 12 | 0.508 | 0.547 / 0.221 | 0.628 / 0.369 | 0.510 | 0.500 | 0.500 |
| *With pre-train & transfer* | | | | | | | |
| Frozen: | no layers | 0.904 | 0.693 / 0.425 | 0.802 / 0.622 | 0.745 | **0.650** | 0.633 |
| | only emb. | 0.896 | 0.717 / 0.461 | 0.822 / 0.661 | 0.760 | 0.500 | 0.661 |
| | emb. + 3 | 0.916 | 0.720 / 0.449 | 0.830 / 0.677 | 0.768 | 0.500 | **0.664** |
| | emb. + 6 | 0.896 | **0.743 / 0.502** | **0.838 / 0.688** | **0.773** | 0.500 | 0.619 |
| | emb. + 9 | 0.920 | 0.707 / 0.454 | 0.826 / 0.671 | 0.748 | 0.600 | 0.600 |
| | emb. + 12 | 0.848 | 0.637 / 0.348 | 0.730 / 0.480 | 0.510 | 0.515 | 0.494 |

Table 3: Results of BERT model fine-tuning.

language by subsampling the larger class. Together with the BeSLS data, there are ≈970K instances.

• Linguistic acceptability: all train and dev instances were taken from MELA v1.0 (Zhang et al., 2024), as well as Dutch COLA[14] and HuCOLA (Ligeti-Nagy et al., 2024). Positive instances were subsampled to maintain the class balance. Together with the BelaCoLA data, there are ≈42K instances.

• Word in context: all train and dev instances were taken from XL-WiC (Raganato et al., 2020) and RUSSE (Shavrina et al., 2020, section 3.1.2). Negative instances in RUSSE were subsampled to maintain the class balance. Together with the BeWiC data, there are ≈145K instances.

• Winograd schema challenge: all training instances were taken from WinoGrande (Sakaguchi et al., 2021) and XWINO (Tikhonov and Ryabinin, 2021), and all German, French, Russian instances — from the folder `lm_wino_x` of Wino-X (Emelin and Sennrich, 2021). To deal with three different formats in the data, we brought all instances to WNLI structure by constructing the hypotheses automatically from the original sentences: the pronoun or the _ sign is replaced with one of the two coreference candidate spans, and the label is 1 with the correct candidate and 0 otherwise. Note this is a very crude procedure, so that in morphologically richer languages it often produces ungrammatical (although comprehensible) sentences. Together with the BeWSC data, there are ≈110K instances.

• Textual entailment: all instances in the folds `dev_matched` and `dev_mismatched` were taken from MNLI (Williams et al., 2018), and all dev instances — from XNLI (Conneau et al., 2018). Since there are three balanced classes (entailment / neutral / contradiction), we subsampled half of the neutral / contradictory instances to represent non-entailment. Together with the BeRTE-WD data, there are ≈40K instances.

Transfer learning allows to beat the trivial baseline on BeWSC and BeRTE-WD (see Table 3). Improvement is also observed in other datasets.

### 4.3 LLM baselines

We added configurations for BelarusianGLUE tasks to a fork of `lm-evaluation-harness` (Gao et al., 2024). Four types of prompts are examined:

• instructions in Belarusian, zero-shot or few-shot (11 instances, the same as those provided to humans to establish their baseline performance);

• instructions in English, zero-shot or few-shot (first 10 instances in the dev set of each dataset).

Predictions are estimated from log probabilities of the tokens `0` / `1` to follow the target instance.

We measured the performance of local LLMs below 15B parameters on BelarusianGLUE. Among recent (as of December 2024) models and model families with multilingual capabilities, we evaluated Llama 3.1 and 3.2 (Dubey et al., 2024), Phi 3 and 3.5 (Abdin et al., 2024), Gemma 2 (Riviere et al., 2024), Qwen 2 and 2.5 (Yang et al., 2024), Mistral Nemo and Ministral (Jiang et al., 2023), Aya 23 8B (Aryabumi et al., 2024). Additionally we tested several models that demonstrate state-

[14] https://huggingface.co/datasets/GroNLP/dutch-cola

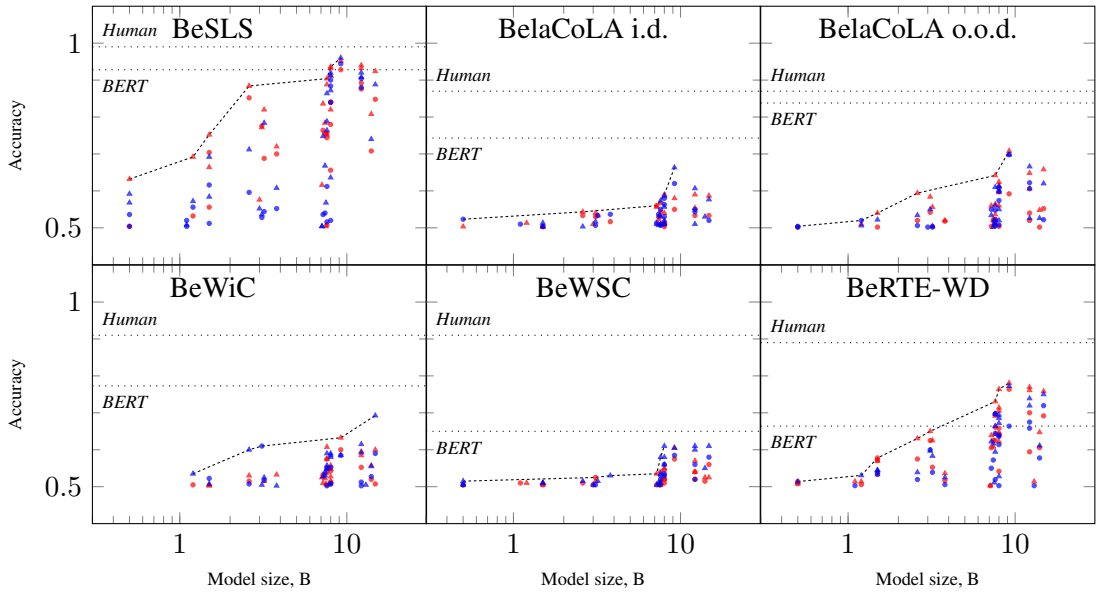


Figure 1: Local LLM size vs. accuracy on BelarusianGLUE. Each point represents a single evaluation run against a local LLM with prompt in Belarusian (blue) or English (red), zero-shot (circle marks) or few-shot (triangle marks). Scores $\leq 0.5$ are not shown. Dotted lines are human scores and the best BERT model scores on each dataset respectively. Dashed line is the Pareto front of optimal LLM performance at given size.

of-the-art performance for Ukrainian and Russian: Sherlock (Boros et al., 2024) and Vikhr (Nikolich et al., 2024), based on various versions of Llama and Mistral, — as well as several older multilingual models: XGLM-7.5B (Lin et al., 2022), mGPT-13B (Shliazhko et al., 2023), BLOOM (Scao et al., 2023) and BLOOMZ (Muennighoff et al., 2023) up to 7B parameters.

Our evaluation of state-of-the-art commercial models was less systematic: we tested GPT-4o and Claude 3.5 Sonnet, but, since their APIs don't support sampling log probabilities for the tokens specified in advance (such as `0` / `1`), we only checked for exact match. This may understate the true performance of these models in comparison with the local models. Due to Claude's tendency to include reasoning in its outputs, an additional post-processing step was required to extract predictions.

Full evaluation results are available in the Tables 6–9 in Appendix B. Among the local LLMs below 15B parameters, Gemma 2 9B is the top competitor, in line with the findings of Thellmann et al. (2024). When a model beats the trivial baseline of 50% accuracy, its few-shot scores tend to be better than zero-shot ones, while the impact of prompting language, Belarusian or English, isn't as clear and varies between datasets.

Figure 1 visualizes the dependency between model size and its scores. Larger models generally perform better, although the Pareto fronts show that the relation isn't linear. Both in zero-shot and in few-shot setting, LLMs outperform supervised BERT baselines on BeSLS and, most impressively, BeRTE-WD, which may be a sign of inherently stronger reasoning capabilities. In other tasks, the supervised baselines are closer to the human level.

Additionally, we tried fine-tuning Gemma 2 9B with PEFT[15] on each of the five training sets separately. LoRA adapters were trained in 4-bit precision for 10 epochs with learning rate 2e-4, batch size 64 on a single GPU. Snapshots were saved once per epoch, and the best snapshot was selected by the dev set loss. As shown in Table 5 in Appendix B, this brings some improvements, especially in BeWiC and BeRTE-WD, although there doesn't seem to be any consistent pattern.

Overall, the highest scores were observed for commercial models with Belarusian prompts. While both GPT-4o and Claude perform well with Belarusian prompts in the zero-shot setting, their behavior diverges with in-context examples. GPT-4o maintains consistent performance with Belarusian prompts and shows improvements in certain tasks with the English ones. Claude 3.5 Sonnet tends to provide reasoning, when prompted in Belarusian, and shows weaker performance on BeWSC with in-context examples.

[15] https://github.com/huggingface/peft

## 5 Discussion

Our findings indicate that a traditional GLUE-type benchmark for a lower-resourced language (Belarusian) may still be challenging for the current generation of LLMs. This offers a complementary perspective to Fenogenova et al. (2024) who assess the benchmarks existing for a related high-resource language (Russian) to be not challenging enough.

Unlike Rybak et al. (2020), we find that a multilingual pretrained BERT model, such as mDeBERTa-v3, is able to outperform a more narrowly focused, monolingual model, such as Belarusian HPLT BERT, possibly due to more advanced architecture or larger scale of the pre-training data.

Our results support the conventional wisdom that the model size is not the only factor that affects the down-stream performance (Hardalov et al., 2023).

## 6 Conclusion

We have introduced BelarusianGLUE and measured how BERT models and LLMs perform on this novel benchmark, as compared with human performance. The easiest dataset, BeSLS, is mostly solved, although some of the instances in it are challenging even for state-of-the-art commercial LLMs: they are struggling to correctly classify instances with implied positivity that relies on domain knowledge or complex pragmatics. For the hardest dataset, BeWSC, there is still a significant gap in performance between human and machine. To obtain higher scores from the LLMs, further prompting tweaks may be required.

We release the datasets[16] and evaluation code.[17]

In the future, we may want to expand the benchmark by adding a diagnostic dataset (present in many of the GLUE-type benchmarks); a perplexity dataset for evaluating generative capabilities of multilingual LLMs, applied to Belarusian; a culture-specific QA dataset that would be hard for the current generation of LLMs but reasonably easy for the native speakers, etc. Converting all datasets to the alternative Belarusian orthographies, *taraškievica* and *łacinka*, may help understand how the model performance on Belarusian language tasks depends on the choice of orthography. Following the common practice, we may want to create a leaderboard to automate evaluation of new models on BelarusianGLUE. For the cutting-edge reasoning models, not covered in our evaluation, it remains to be investigated how the token budget (Wang et al., 2024) affects performance. Finally, the analysis of model errors is also a promising direction of further research.

[16] https://hf.co/datasets/maaxap/BelarusianGLUE

[17] https://github.com/maaxap/BelarusianGLUE


## Acknowledgments

This work was supported by the Technology Agency of the Czech Republic under the project FactDeMice – Fact Verification Based on Textual Evidence Using Fact-Consistent Translation, Fake Review Detection, and Automatic Extraction of Misleading Claims (Project Code: TQ16000028).


## Limitations and ethical considerations

As of yet, we haven't been able to train strong baseline models based on mT5 architecture (Xue et al., 2021) for our datasets, although state-of-the-art performance was reported for similar English datasets, such as WinoGrande (Sakaguchi et al., 2021).

Dataset usage is affected by the political situation in Belarus. Due to sheer amount of sources officially recognized as "extremist materials" by Belarusian courts,[18] we cannot guarantee that BelarusianGLUE is legally safe to use in Belarus — or will be safe in the future, as the list of "extremist materials" is being regularly updated.

## References


Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, and Harkirat Behl et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23:


[18] For context, see Sections 41 and 50 in the UN report on human rights in Belarus.

Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Lukasz Augustyniak, Kamil Tagowski, Albert Sawczyn, Denis Janiak, Roman Bartusiak, Adrian Szymczak, Arkadiusz Janz, Piotr Szymański, Marcin Wątroba, Mikoł aj Morzy, Tomasz Kajdanowicz, and Maciej Piasecki. 2022. This is the way: designing and compiling LEPISZCZE, a comprehensive NLP benchmark for Polish. In *Advances in Neural Information Processing Systems*, volume 35, pages 21805–21818. Curran Associates, Inc.

Łukasz Augustyniak, Szymon Woźniak, Marcin Gruza, Piotr Gramacki, Krzysztof Rajda, Mikołaj Morzy, and Tomasz Kajdanowicz. 2023. Massively multilingual corpus of sentiment datasets and multi-faceted sentiment classification benchmark. *Preprint*, arXiv:2306.07902.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Aleksandrs Berdicevskis, Gerlof Bouma, Robin Kurtz, Felix Morger, Joey Öhman, Yvonne Adesam, Lars Borin, Dana Dannélls, Markus Forsberg, Tim Isbister, Anna Lindahl, Martin Malmsten, Faton Rekathati, Magnus Sahlgren, Elena Volodina, Love Börjeson, Simon Hengchen, and Nina Tahmasebi. 2023. Superlim: A Swedish language understanding evaluation benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8137–8153, Singapore. Association for Computational Linguistics.

Tiberiu Boros, Radu Chivereanu, Stefan Dumitrescu, and Octavian Purcaru. 2024. Fine-tuning and retrieval augmented generation for question answering using affordable large language models. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 75–82, Torino, Italia. ELRA and ICCL.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

James V. Bradley. 1958. Complete counterbalancing of immediate sequential effects in a latin square design. *Journal of the American Statistical Association*, 53(282):525–528.

Christian Buck, Kenneth Heafield, and Bas van Ooyen. 2014. N-gram counts and language models from the Common Crawl. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3579–3584, Reykjavik, Iceland. European Language Resources Association (ELRA).

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.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. 2024. A new massive multilingual dataset for high-performance language technologies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128, Torino, Italia. ELRA and ICCL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, and Angela Fan et al. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A

massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Abdel Rahim Elmadany, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2022. Orca: A challenging benchmark for arabic language understanding. *arXiv preprint arXiv:2212.10758*. Accepted by COLING2020; 10 pages, 4 figures.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. GPTs are GPTs: An early look at the labor market impact potential of large language models. *Preprint*, arXiv:2303.10130.

Denis Emelin and Rico Sennrich. 2021. Wino-X: Multilingual Winograd schemas for commonsense reasoning and coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8517–8532, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alena Fenogenova, Artem Chervyakov, Nikita Martynov, Anastasia Kozlova, Maria Tikhonova, Albina Akhmetgareeva, Anton Emelyanov, Denis Shevelev, Pavel Lebedev, Leonid Sinev, Ulyana Isaeva, Katerina Kolomeytseva, Daniil Moskovskiy, Elizaveta Goncharova, Nikita Savushkin, Polina Mikhailova, Anastasia Minaeva, Denis Dimitrov, Alexander Panchenko, and Sergey Markov. 2024. MERA: A comprehensive LLM evaluation in Russian. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9920–9948, Bangkok, Thailand. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

Daniel Grießhaber, Johannes Maucher, and Ngoc Thang Vu. 2020. Fine-tuning BERT for low-resource natural language understanding via active learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1158–1171, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Serhii Hamotskyi, Anna-Izabella Levbarg, and Christian Hänig. 2024. Eval-UA-tion 1.0: Benchmark for evaluating Ukrainian (large) language models. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 109–119, Torino, Italia. ELRA and ICCL.

Momchil Hardalov, Pepa Atanasova, Todor Mihaylov, Galia Angelova, Kiril Simov, Petya Osenova, Veselin Stoyanov, Ivan Koychev, Preslav Nakov, and Dragomir Radev. 2023. bgGLUE: A Bulgarian general language understanding evaluation benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8733–8759, Toronto, Canada. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

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.

Bernhard Kittel, Diana Lindner, Sviatlana Tesch, and Gerd Hentschel. 2010. Mixed language usage in Belarus: the sociostructural background of language choice. *International Journal of the Sociology of Language*, 2010(206):47–71.

Dimitrios Kotzias, Misha Denil, Nando de Freitas, and Padhraic Smyth. 2015. From group to individual labels using deep features. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 597–606, New York, NY, USA. Association for Computing Machinery.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd schema challenge. In

*Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, pages 552–561. AAAI Press.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Noémi Ligeti-Nagy, Gergő Ferenczi, Enikő Héja, László János Laki, Noémi Vadász, Zijian Győző Yang, and Tamás Váradi. 2024. HuLU: Hungarian language understanding benchmark kit. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8360–8371, Torino, Italia. ELRA and ICCL.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual language models. *Preprint*, arXiv:2112.10668.

Vladislav Mikhailov, Tatiana Shamardina, Max Ryabinin, Alena Pestova, Ivan Smurov, and Ekaterina Artemova. 2022. RuCoLA: Russian corpus of linguistic acceptability. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5207–5227, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sushree Sangita Mohanty, Satya Ranjan Dash, and Shantipriya Parida, editors. 2024. *Applying AI-Based Tools and Technologies Towards Revitalization of Indigenous and Endangered Languages*. Springer, Singapore.

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. *Preprint*, arXiv:2211.01786.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *Preprint*, arXiv:2309.09400.

Aleksandr Nikolich, Konstantin Korolev, Sergei Bratchikov, Igor Kiselev, and Artem Shelmanov. 2024. Vikhr: Constructing a state-of-the-art bilingual open-source instruction-following large language model for Russian. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 189–199, Miami, Florida, USA. Association for Computational Linguistics.

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. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Cardiff, United Kingdom.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. The Edinburgh Twitter corpus. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, pages 25–26, Los Angeles, California, USA. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. 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 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.

Gemma Team: Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and Johan Ferret et al.

2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Viraat Aryabumi, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, Sharad Duwal, Daniil Dzenhaliou, Daniel Fernando Erazo Florez, Fabian Farestam, Joseph Marvin Imperial, Shayekh Bin Islam, Perttu Isotalo, Maral Jabbarishiviari, Börje F. Karlsson, Eldar Khalilov, Christopher Klamm, Fajri Koto, Dominik Krzemiński, Gabriel Adriano de Melo, Syrielle Montariol, Yiyang Nan, Joel Niklaus, Jekaterina Novikova, Johan Samir Obando Ceron, Debjit Paul, Esther Ploeger, Jebish Purbey, Swati Rajwal, Selvan Sunitha Ravi, Sara Rydell, Roshan Santhosh, Drishti Sharma, Marjana Prifti Skenduli, Arshia Soltani Moakhar, Bardia Soltani Moakhar, Ran Tamir, Ayush Kumar Tarun, Azmine Toushik Wasi, Thenuka Ovin Weerasinghe, Serhan Yilmaz, Mike Zhang, Imanol Schlag, Marzieh Fadaee, Sara Hooker, and Antoine Bosselut. 2024. INCLUDE: Evaluating multilingual language understanding with regional knowledge. *Preprint*, arXiv:2411.19799.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik. 2020. KLEJ: Comprehensive benchmark for Polish language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1191–1201, Online. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. WinoGrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

David Samuel, Andrey Kutuzov, Lilja Øvrelid, and Erik Velldal. 2023. Trained on 100 million words and still in shape: BERT meets British National Corpus. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1954–1974, Dubrovnik, Croatia. Association for Computational Linguistics.

BigScience Workshop: Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, and Matthias Gallé et al. 2023. BLOOM: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

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.

Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. RussianSuperGLUE: A Russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726, Online. Association for Computational Linguistics.

Yana Shishkina and Olga Lyashevskaya. 2022. Sculpting enhanced dependencies for Belarusian. In *Analysis of Images, Social Networks and Texts*, pages 137–147, Cham. Springer International Publishing.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2023. mGPT: Few-shot learners go multilingual. *Preprint*, arXiv:2204.07580.

Aleksandr S. Sokolov. 2022. Assessment of reliability of results of the 2019 Belarus population census based on an analysis of changes in the ethnolinguistic composition of population. *Demographic Review*, 9(4):61–103.

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, and Mehdi Ali. 2024. Towards multilingual LLM evaluation for European languages. *Preprint*, arXiv:2410.08928.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Alexey Tikhonov and Max Ryabinin. 2021. It's All in the Heads: Using Attention Heads as a Baseline for Cross-Lingual Transfer in Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3534–3546, Online. Association for Computational Linguistics.

Bibek Upadhayay and Vahid Behzadan. 2024. TaCo: Enhancing cross-lingual transfer for low-resource languages in LLMs through translation-assisted chain-of-thought processes. *Preprint*, arXiv:2311.10797.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Xiaolong Wang, Yile Wang, Yuanchi Zhang, Fuwen Luo, Peng Li, Maosong Sun, and Yang Liu. 2024. Reasoning in conversation: Solving subjective tasks through dialogue simulation for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15880–15893, Bangkok, Thailand. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: A benchmark of linguistic minimal pairs for English. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 409–410, New York, New York. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Yu Xie and Sofia Avila. 2024. The social impact of generative LLM-based AI. *Preprint*, arXiv:2410.21281.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on 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.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, and Fei Huang et al. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Aleš Žagar and Marko Robnik-Šikonja. 2022. Slovene SuperGLUE benchmark: Translation and evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2058–2065, Marseille, France. European Language Resources Association.

Siarhiej Zaprudski. 2007. In the grip of replacive bilingualism: the Belarusian language in contact with Russian. *International Journal of the Sociology of Language*, 2007(183):97–118.

Ziyin Zhang, Yikang Liu, Weifang Huang, Junyu Mao, Rui Wang, and Hai Hu. 2024. MELA: Multilingual evaluation of linguistic acceptability. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2658–2674, Bangkok, Thailand. Association for Computational Linguistics.

# A Sample instances

| Dataset ID | Instance | Label |
|---|---|---|
| BeSLS | *Nie razumieju, čamu niekamu nie spadabałasia kniha.*<br>'I don't understand why someone didn't like the book.' | 1 |
| | *Stolki pafasu, a na vychadzie adzin pšyk.*<br>'So much pathos, and nothing at the end.' | 0 |
| BelaCoLA | *Aleś zaŭvažyŭ na drevie niezvyčajnych ptušak.*<br>'Aleś noticed unusual birds on the tree.' | 1 |
| | *Jany ličyli jaho talenavitymi.*<br>'They considered him talented.PL.' | 0 |
| BeWiC | *U spravazdačnym **dakładzie** staršynia źviarnuŭ uvahu na šerah važnych momantaŭ. \| Uviečary ŭ kłubie, paśla **dakłada**, była mastackaja častka.*<br>'In the status **report**, the chairman drew attention to a number of important points. \| In the evening in the club, after the **report**, there was a performance.' | 1 |
| | *Hordaja hałava alenia, upryhožanaja raskidzistymi **rahami**, była krychu pryŭźniata, vušy naściarožany. \| Prypyniŭšysia na **rahu** vulicy, Natalla Maksimaŭna pačakała, pakul projdzie kałona aŭtamašyn z vajskoŭcami.*<br>'The deer's proud head, decorated with spreading **antlers**, was slightly raised, the ears were alert. \| Having stopped at the **corner** of the street, Natalla Maksimaŭna waited for a column of cars with soldiers to pass.' | 0 |
| BeWSC | *Navat śmieły **voin** byŭ biaśsilny suprać lutaha lva, i navat vostry mieč nie moh dapamahčy **jamu**.*<br>(⇒ *Navat vostry mieč nie moh dapamahčy voinu.*)<br>'Even a brave **warrior** was powerless against the fierce lion, and even a sharp sword could not help **him**.'<br>(⇒ 'Even a sharp sword could not help the warrior.') | 1 |
| | *Ja apuściŭ haračuju dałoń u **vadu**, i **jana** astyła.*<br>(⇒ *Vada astyła.*)<br>'I dipped my hot palm into **water**, and **it** cooled down.'<br>(⇒ 'The water cooled down.') | 0 |
| BeRTE-WD | *Natałka Babina pierakłała kulinarnuju knihu «Litoŭskaja kucharka». ⇒ Natałka Babina maje dośvied u halinie pierakładu.*<br>'Natałka Babina translated the cookbook "Lithuanian Cook". ⇒ Natałka Babina has experience in the field of translation.' | 1 |
| | *Kamianieckaja vieža była pabudavana ŭ 1288 hodzie, a Novy zamak u Hrodnie byŭ pabudavany ŭ 1751 hodzie. ⇒ Kamianieckaja vieža była pabudavana bolš jak na čatyry stahodździ paźniej za Novy zamak u Hrodnie.*<br>'The Kamianiec Tower was built in 1288, and the New Castle in Hrodna was built in 1751.<br>⇒ The Kamianiec Tower was built more than four centuries later than the New Castle in Hrodna.' | 0 |

Table 4: Sample instances (in romanized spelling) with English translations.

# B Detailed results of LLM evaluation

| Prompt language and type | Dataset: / Metric: | BeSLS acc | BelaCoLA i.d. acc / MCC | BelaCoLA o.o.d. acc / MCC | BeWiC acc | BeWSC acc | BeRTE-WD acc |
|---|---|---|---|---|---|---|---|
| Belarusian | zero-shot | 0.944 | 0.620 / 0.247 | 0.698 / 0.397 | 0.585 | 0.585 | 0.664 |
| | few-shot | **0.960** | **0.663 / 0.327** | 0.698 / 0.396 | 0.585 | **0.605** | 0.772 |
| | zero-shot w. fine-tuning | 0.936 | 0.620 / 0.255 | 0.678 / 0.356 | **0.685** | 0.585 | 0.792 |
| English | zero-shot | 0.928 | 0.550 / 0.124 | 0.592 / 0.202 | 0.600 | 0.575 | 0.764 |
| | few-shot | 0.952 | 0.580 / 0.235 | **0.708 / 0.443** | 0.633 | **0.605** | 0.781 |
| | zero-shot w. fine-tuning | 0.956 | 0.643 / 0.287 | 0.686 / 0.379 | 0.635 | **0.605** | **0.814** |

Table 5: Results of Gemma 2 9B fine-tuning.

| Model ID on HuggingFace | Size | Dataset: Metric: Prec. | BeSLS acc | BelaCoLA i.d. acc / MCC | BelaCoLA o.o.d. acc / MCC | BeWiC acc | BeWSC acc | BeRTE-WD acc |
|---|---|---|---|---|---|---|---|---|
| ai-forever/mGPT-13B | 13,0 | 8bit | 0.500 | 0.500 / 0.000 | 0.500 / 0.000 | 0.500 | 0.500 | 0.503 |
| bigscience/bloom-1b1 | 1,1 | full | 0.504 | 0.500 / 0.000 | 0.500 / 0.000 | 0.500 | 0.500 | 0.503 |
| bigscience/bloom-3b | 3,0 | full | 0.488 | 0.497 / -0.014 | 0.502 / 0.010 | 0.500 | 0.505 | 0.500 |
| bigscience/bloom-7b1 | 7,1 | full | 0.500 | 0.500 / 0.000 | 0.500 / 0.000 | 0.500 | 0.500 | 0.500 |
| bigscience/bloomz-1b1 | 1,1 | full | 0.520 | 0.510 / 0.023 | 0.492 / -0.019 | 0.498 | 0.500 | 0.500 |
| bigscience/bloomz-3b | 3,0 | full | 0.500 | 0.500 / 0.000 | 0.498 / -0.026 | 0.500 | 0.500 | 0.500 |
| bigscience/bloomz-7b1 | 7,1 | full | 0.500 | 0.500 / 0.000 | 0.500 / 0.000 | 0.500 | 0.500 | 0.500 |
| CohereForAI/aya-23-8B | 8,0 | full | 0.520 | 0.510 / 0.078 | 0.508 / 0.074 | 0.500 | 0.490 | 0.503 |
| facebook/xglm-7.5B | 7,5 | full | 0.456 | 0.497 / -0.007 | 0.516 / 0.033 | 0.495 | 0.505 | 0.497 |
| google/gemma-2-2b-it | 2,6 | full | 0.596 | 0.497 / -0.034 | 0.506 / 0.060 | 0.508 | 0.500 | 0.519 |
| google/gemma-2-9b-it | 9,2 | full | **0.944** | **0.620 / 0.247** | **0.698 / 0.397** | 0.585 | **0.585** | 0.664 |
| meta-llama/Llama-3.1-8B-Instruct | 8,0 | full | 0.884 | 0.530 / 0.063 | 0.598 / 0.198 | 0.508 | 0.530 | 0.636 |
| meta-llama/Llama-3.2-1B-Instruct | 1,2 | full | 0.556 | 0.500 / 0.000 | 0.520 / 0.040 | 0.493 | 0.495 | 0.492 |
| meta-llama/Llama-3.2-3B-Instruct | 3,2 | full | 0.544 | 0.533 / 0.069 | 0.502 / 0.004 | 0.518 | 0.500 | 0.539 |
| microsoft/Phi-3-medium-4k-instruct | 14,0 | 8bit | 0.496 | 0.493 / -0.019 | 0.494 / -0.020 | 0.528 | 0.495 | 0.578 |
| microsoft/Phi-3-small-8k-instruct | 7,4 | full | 0.540 | 0.537 / 0.091 | 0.492 / -0.021 | 0.538 | 0.510 | 0.572 |
| microsoft/Phi-3.5-mini-instruct | 3,8 | full | 0.552 | 0.537 / 0.133 | 0.498 / -0.009 | 0.495 | 0.500 | 0.506 |
| mistralai/Ministral-8B-Instruct-2410 | 8,0 | full | 0.900 | 0.563 / 0.171 | 0.574 / 0.186 | 0.553 | 0.535 | 0.614 |
| mistralai/Mistral-Nemo-Instruct-2407 | 12,2 | 8bit | 0.880 | 0.550 / 0.164 | 0.606 / 0.297 | 0.503 | 0.520 | 0.658 |
| Qwen/Qwen2-0.5B-Instruct | 0,5 | full | 0.504 | 0.523 / 0.084 | 0.504 / 0.014 | 0.498 | 0.505 | 0.500 |
| Qwen/Qwen2-1.5B-Instruct | 1,5 | full | 0.512 | 0.503 / 0.058 | 0.498 / -0.045 | 0.500 | 0.480 | 0.533 |
| Qwen/Qwen2-7B-Instruct | 7,6 | full | 0.516 | 0.507 / 0.082 | 0.522 / 0.129 | 0.498 | 0.530 | 0.644 |
| Qwen/Qwen2.5-0.5B-Instruct | 0,5 | full | 0.536 | 0.490 / -0.046 | 0.502 / 0.009 | 0.498 | 0.500 | 0.497 |
| Qwen/Qwen2.5-1.5B-Instruct | 1,5 | full | 0.616 | 0.440 / -0.120 | 0.484 / -0.032 | 0.523 | 0.500 | 0.539 |
| Qwen/Qwen2.5-3B-Instruct | 3,1 | full | 0.528 | 0.507 / 0.034 | 0.488 / -0.036 | **0.610** | 0.505 | 0.600 |
| Qwen/Qwen2.5-7B-Instruct | 7,6 | full | 0.400 | 0.550 / 0.115 | 0.548 / 0.123 | 0.550 | 0.520 | 0.697 |
| Qwen/Qwen2.5-14B-Instruct | 14,8 | 8bit | 0.436 | 0.520 / 0.065 | 0.522 / 0.091 | 0.590 | 0.580 | **0.719** |
| SherlockAssistant/Mistral-7B-Instruct-Ukrainian | 7,2 | full | 0.536 | 0.513 / 0.038 | 0.550 / 0.116 | 0.490 | 0.500 | 0.550 |
| Vikhrmodels/Vikhr-7B-instruct_0.4 | 7,6 | full | 0.612 | 0.497 / -0.018 | 0.506 / 0.041 | 0.503 | 0.490 | 0.519 |
| Vikhrmodels/Vikhr-Llama3.1-8B-Instruct-R-21-09-24 | 8,0 | full | 0.840 | 0.580 / 0.169 | 0.608 / 0.218 | 0.508 | 0.560 | 0.639 |
| Vikhrmodels/Vikhr-Nemo-12B-Instruct-R-21-09-24 | 12,2 | 8bit | 0.904 | 0.543 / 0.087 | 0.622 / 0.245 | 0.513 | 0.560 | 0.675 |
| claude-3-5-sonnet-20241022 | — | — | 0.956 | **0.747 / 0.523** | **0.882 / 0.767** | **0.878** | **0.775** | 0.778 |
| gpt-4o-2024-11-20 | — | — | **0.976** | 0.700 / 0.473 | 0.860 / 0.739 | 0.850 | 0.710 | **0.889** |

Table 6: Results of LLM evaluation with zero-shot prompts in Belarusian.

| Model ID on HuggingFace | Size | Dataset: Metric: Prec. | BeSLS acc | BelaCoLA i.d. acc / MCC | BelaCoLA o.o.d. acc / MCC | BeWiC acc | BeWSC acc | BeRTE-WD acc |
|---|---|---|---|---|---|---|---|---|
| ai-forever/mGPT-13B | 13,0 | 8bit | 0.500 | 0.500 / 0.000 | 0.500 / 0.000 | 0.505 | 0.500 | 0.500 |
| bigscience/bloom-1b1 | 1,1 | full | 0.508 | 0.490 / -0.101 | 0.500 / 0.000 | 0.495 | 0.500 | 0.497 |
| bigscience/bloom-3b | 3,0 | full | 0.552 | 0.490 / -0.038 | 0.498 / -0.007 | 0.483 | 0.500 | 0.494 |
| bigscience/bloom-7b1 | 7,1 | full | 0.504 | 0.500 / 0.000 | 0.500 / 0.000 | 0.500 | 0.500 | 0.497 |
| bigscience/bloomz-1b1 | 1,1 | full | 0.500 | 0.500 / 0.000 | 0.500 / 0.000 | 0.500 | 0.500 | 0.500 |
| bigscience/bloomz-3b | 3,0 | full | 0.500 | 0.500 / 0.000 | 0.500 / 0.000 | 0.500 | 0.500 | 0.500 |
| bigscience/bloomz-7b1 | 7,1 | full | 0.504 | 0.500 / 0.000 | 0.500 / 0.000 | 0.460 | 0.500 | 0.500 |
| CohereForAI/aya-23-8B | 8,0 | full | 0.636 | 0.510 / 0.028 | 0.520 / 0.051 | 0.500 | 0.500 | 0.528 |
| facebook/xglm-7.5B | 7,5 | full | 0.568 | 0.533 / 0.077 | 0.510 / 0.025 | 0.500 | 0.500 | 0.511 |
| google/gemma-2-2b-it | 2,6 | full | 0.712 | 0.503 / 0.010 | 0.534 / 0.105 | 0.600 | 0.515 | 0.539 |
| google/gemma-2-9b-it | 9,2 | full | **0.960** | **0.663 / 0.327** | **0.698 / 0.396** | 0.585 | 0.605 | **0.772** |
| meta-llama/Llama-3.1-8B-Instruct | 8,0 | full | 0.912 | 0.520 / 0.067 | 0.550 / 0.133 | 0.580 | 0.580 | 0.672 |
| meta-llama/Llama-3.2-1B-Instruct | 1,2 | full | 0.572 | 0.490 / -0.046 | 0.506 / 0.038 | 0.535 | 0.500 | 0.531 |
| meta-llama/Llama-3.2-3B-Instruct | 3,2 | full | 0.784 | 0.497 / -0.008 | 0.504 / 0.009 | 0.525 | 0.515 | 0.583 |
| microsoft/Phi-3-medium-4k-instruct | 14,0 | 8bit | 0.740 | 0.530 / 0.065 | 0.526 / 0.058 | 0.555 | 0.550 | 0.611 |
| microsoft/Phi-3-small-8k-instruct | 7,4 | full | 0.668 | 0.507 / 0.019 | 0.502 / 0.006 | 0.530 | 0.495 | 0.617 |
| microsoft/Phi-3.5-mini-instruct | 3,8 | full | 0.608 | 0.483 / -0.035 | 0.490 / -0.022 | 0.503 | 0.530 | 0.525 |
| mistralai/Ministral-8B-Instruct-2410 | 8,0 | full | 0.872 | 0.587 / 0.201 | 0.612 / 0.249 | 0.585 | 0.560 | 0.694 |
| mistralai/Mistral-Nemo-Instruct-2407 | 12,2 | 8bit | 0.908 | 0.607 / 0.213 | 0.666 / 0.334 | 0.615 | 0.600 | 0.719 |
| Qwen/Qwen2-0.5B-Instruct | 0,5 | full | 0.568 | 0.470 / -0.065 | 0.492 / -0.017 | 0.435 | 0.505 | 0.514 |
| Qwen/Qwen2-1.5B-Instruct | 1,5 | full | 0.692 | 0.503 / 0.034 | 0.500 / 0.000 | 0.508 | 0.510 | 0.544 |
| Qwen/Qwen2-7B-Instruct | 7,6 | full | 0.864 | 0.530 / 0.071 | 0.560 / 0.138 | 0.558 | 0.535 | 0.664 |
| Qwen/Qwen2.5-0.5B-Instruct | 0,5 | full | 0.592 | 0.500 / 0.000 | 0.472 / -0.060 | 0.443 | 0.515 | 0.500 |
| Qwen/Qwen2.5-1.5B-Instruct | 1,5 | full | 0.584 | 0.513 / 0.028 | 0.522 / 0.046 | 0.468 | 0.510 | 0.544 |
| Qwen/Qwen2.5-3B-Instruct | 3,1 | full | 0.536 | 0.517 / 0.033 | 0.550 / 0.105 | 0.505 | 0.480 | 0.597 |
| Qwen/Qwen2.5-7B-Instruct | 7,6 | full | 0.788 | 0.573 / 0.148 | 0.610 / 0.220 | 0.590 | 0.580 | 0.697 |
| Qwen/Qwen2.5-14B-Instruct | 14,8 | 8bit | 0.888 | 0.577 / 0.201 | 0.620 / 0.286 | **0.693** | **0.610** | 0.750 |
| SherlockAssistant/Mistral-7B-Instruct-Ukrainian | 7,2 | full | 0.748 | 0.510 / 0.024 | 0.534 / 0.087 | 0.528 | 0.505 | 0.622 |
| Vikhrmodels/Vikhr-7B-instruct_0.4 | 7,6 | full | 0.764 | 0.510 / 0.026 | 0.494 / -0.014 | 0.548 | 0.510 | 0.594 |
| Vikhrmodels/Vikhr-Llama3.1-8B-Instruct-R-21-09-24 | 8,0 | full | 0.920 | 0.557 / 0.136 | 0.600 / 0.227 | 0.590 | **0.610** | 0.686 |
| Vikhrmodels/Vikhr-Nemo-12B-Instruct-R-21-09-24 | 12,2 | 8bit | 0.920 | 0.510 / 0.039 | 0.526 / 0.136 | 0.593 | **0.610** | 0.739 |
| claude-3-5-sonnet-20241022 | — | — | 0.964 | **0.773 / 0.565** | **0.874 / 0.748** | 0.755 | 0.570 | 0.806 |
| gpt-4o-2024-11-20 | — | — | **0.980** | 0.733 / 0.513 | 0.864 / 0.733 | **0.843** | **0.740** | **0.883** |

Table 7: Results of LLM evaluation with few-shot prompts in Belarusian.

| Model ID on HuggingFace | Size | Dataset: Metric: Prec. | BeSLS acc | BelaCoLA i.d. acc / MCC | BelaCoLA o.o.d. acc / MCC | BeWiC acc | BeWSC acc | BeRTE-WD acc |
|---|---|---|---|---|---|---|---|---|
| ai-forever/mGPT-13B | 13,0 | 8bit | 0.500 | 0.500 / 0.000 | 0.500 / 0.000 | 0.500 | 0.500 | 0.500 |
| bigscience/bloom-1b1 | 1,1 | full | 0.500 | 0.500 / 0.000 | 0.500 / 0.000 | 0.500 | 0.500 | 0.500 |
| bigscience/bloom-3b | 3,0 | full | 0.500 | 0.500 / 0.000 | 0.500 / 0.000 | 0.500 | 0.500 | 0.500 |
| bigscience/bloom-7b1 | 7,1 | full | 0.500 | 0.500 / 0.000 | 0.500 / 0.000 | 0.500 | 0.500 | 0.500 |
| bigscience/bloomz-1b1 | 1,1 | full | 0.500 | 0.500 / 0.000 | 0.500 / 0.000 | 0.500 | 0.510 | 0.500 |
| bigscience/bloomz-3b | 3,0 | full | 0.500 | 0.500 / 0.000 | 0.500 / 0.000 | 0.500 | 0.500 | 0.500 |
| bigscience/bloomz-7b1 | 7,1 | full | 0.500 | 0.500 / 0.000 | 0.500 / 0.000 | 0.500 | 0.500 | 0.503 |
| CohereForAI/aya-23-8B | 8,0 | full | 0.656 | 0.503 / 0.034 | 0.488 / -0.064 | 0.498 | 0.500 | 0.542 |
| facebook/xglm-7.5B | 7,5 | full | 0.500 | 0.500 / 0.000 | 0.500 / 0.000 | 0.500 | 0.500 | 0.500 |
| google/gemma-2-2b-it | 2,6 | full | 0.852 | 0.533 / 0.083 | 0.520 / 0.057 | 0.515 | 0.510 | 0.575 |
| google/gemma-2-9b-it | 9,2 | full | **0.928** | 0.550 / **0.124** | **0.592 / 0.202** | **0.600** | **0.575** | **0.764** |
| meta-llama/Llama-3.1-8B-Instruct | 8,0 | full | 0.840 | 0.497 / -0.026 | 0.498 / -0.014 | 0.493 | 0.520 | 0.642 |
| meta-llama/Llama-3.2-1B-Instruct | 1,2 | full | 0.532 | 0.500 / 0.000 | 0.500 / 0.000 | 0.505 | 0.500 | 0.506 |
| meta-llama/Llama-3.2-3B-Instruct | 3,2 | full | 0.688 | 0.500 / 0.000 | 0.504 / 0.037 | 0.483 | 0.500 | 0.553 |
| microsoft/Phi-3-medium-4k-instruct | 14,0 | 8bit | 0.708 | 0.493 / -0.041 | 0.502 / 0.020 | 0.520 | 0.515 | 0.606 |
| microsoft/Phi-3-small-8k-instruct | 7,4 | full | 0.756 | 0.530 / 0.070 | 0.522 / 0.045 | 0.473 | 0.480 | 0.625 |
| microsoft/Phi-3.5-mini-instruct | 3,8 | full | 0.700 | 0.517 / 0.041 | 0.518 / 0.049 | 0.498 | 0.490 | 0.517 |
| mistralai/Ministral-8B-Instruct-2410 | 8,0 | full | 0.840 | 0.487 / -0.074 | 0.500 / 0.000 | 0.513 | 0.510 | 0.658 |
| mistralai/Mistral-Nemo-Instruct-2407 | 12,2 | 8bit | 0.892 | 0.477 / -0.063 | 0.520 / 0.048 | 0.553 | 0.540 | 0.700 |
| Qwen/Qwen2-0.5B-Instruct | 0,5 | full | 0.504 | 0.500 / 0.000 | 0.500 / 0.000 | 0.498 | 0.500 | 0.514 |
| Qwen/Qwen2-1.5B-Instruct | 1,5 | full | 0.556 | 0.503 / 0.018 | 0.490 / -0.052 | 0.500 | 0.505 | 0.578 |
| Qwen/Qwen2-7B-Instruct | 7,6 | full | 0.744 | 0.543 / 0.096 | 0.522 / 0.044 | 0.575 | 0.515 | 0.661 |
| Qwen/Qwen2.5-0.5B-Instruct | 0,5 | full | 0.500 | 0.500 / 0.000 | 0.502 / 0.045 | 0.500 | 0.505 | 0.508 |
| Qwen/Qwen2.5-1.5B-Instruct | 1,5 | full | 0.704 | 0.500 / 0.000 | 0.502 / 0.045 | 0.495 | 0.500 | 0.572 |
| Qwen/Qwen2.5-3B-Instruct | 3,1 | full | 0.776 | 0.530 / 0.064 | 0.542 / 0.090 | 0.500 | 0.525 | 0.625 |
| Qwen/Qwen2.5-7B-Instruct | 7,6 | full | 0.756 | 0.500 / 0.000 | 0.508 / 0.037 | 0.505 | 0.540 | 0.697 |
| Qwen/Qwen2.5-14B-Instruct | 14,8 | 8bit | 0.848 | 0.533 / 0.067 | 0.552 / 0.105 | 0.508 | 0.560 | 0.692 |
| SherlockAssistant/Mistral-7B-Instruct-Ukrainian | 7,2 | full | 0.764 | **0.560** / 0.121 | 0.504 / 0.008 | 0.533 | 0.505 | 0.606 |
| Vikhrmodels/Vikhr-7B-instruct_0.4 | 7,6 | full | 0.508 | 0.510 / 0.027 | 0.516 / 0.045 | 0.538 | 0.505 | 0.625 |
| Vikhrmodels/Vikhr-Llama3.1-8B-Instruct-R-21-09-24 | 8,0 | full | 0.780 | 0.497 / -0.034 | 0.504 / 0.023 | 0.543 | 0.545 | 0.644 |
| Vikhrmodels/Vikhr-Nemo-12B-Instruct-R-21-09-24 | 12,2 | 8bit | 0.876 | 0.533 / 0.068 | 0.540 / 0.087 | 0.495 | 0.520 | 0.594 |
| claude-3-5-sonnet-20241022 | — | — | 0.828 | 0.670 / 0.389 | 0.808 / 0.632 | 0.585 | **0.710** | 0.506 |
| gpt-4o-2024-11-20 | — | — | **0.964** | **0.680 / 0.450** | **0.814 / 0.674** | **0.845** | 0.685 | **0.911** |

Table 8: Results of LLM evaluation with zero-shot prompts in English.

| Model ID on HuggingFace | Size | Dataset: Metric: Prec. | BeSLS acc | BelaCoLA i.d. acc / MCC | BelaCoLA o.o.d. acc / MCC | BeWiC acc | BeWSC acc | BeRTE-WD acc |
|---|---|---|---|---|---|---|---|---|
| ai-forever/mGPT-13B | 13,0 | 8bit | 0.500 | 0.497 / -0.058 | 0.470 / -0.070 | 0.483 | 0.500 | 0.514 |
| bigscience/bloom-1b1 | 1,1 | full | 0.496 | 0.497 / -0.034 | 0.470 / -0.070 | 0.490 | 0.500 | 0.514 |
| bigscience/bloom-3b | 3,0 | full | 0.576 | 0.510 / 0.025 | 0.462 / -0.076 | 0.478 | 0.505 | 0.494 |
| bigscience/bloom-7b1 | 7,1 | full | 0.616 | 0.497 / -0.058 | 0.496 / -0.014 | 0.525 | 0.500 | 0.500 |
| bigscience/bloomz-1b1 | 1,1 | full | 0.500 | 0.500 / 0.000 | 0.496 / -0.063 | 0.500 | 0.500 | 0.500 |
| bigscience/bloomz-3b | 3,0 | full | 0.500 | 0.500 / 0.000 | 0.480 / -0.094 | 0.500 | 0.500 | 0.500 |
| bigscience/bloomz-7b1 | 7,1 | full | 0.496 | 0.500 / 0.000 | 0.498 / -0.045 | 0.500 | 0.500 | 0.503 |
| CohereForAI/aya-23-8B | 8,0 | full | 0.820 | 0.513 / 0.033 | 0.560 / 0.120 | 0.528 | 0.500 | 0.622 |
| facebook/xglm-7.5B | 7,5 | full | 0.504 | 0.473 / -0.056 | 0.488 / -0.030 | 0.508 | 0.505 | 0.494 |
| google/gemma-2-2b-it | 2,6 | full | 0.884 | 0.543 / 0.157 | 0.594 / 0.205 | 0.530 | 0.475 | 0.631 |
| google/gemma-2-9b-it | 9,2 | full | **0.952** | 0.580 / **0.235** | **0.708 / 0.443** | **0.633** | **0.605** | **0.781** |
| meta-llama/Llama-3.1-8B-Instruct | 8,0 | full | 0.912 | 0.540 / 0.160 | 0.596 / 0.215 | 0.558 | 0.525 | 0.706 |
| meta-llama/Llama-3.2-1B-Instruct | 1,2 | full | 0.692 | 0.513 / 0.031 | 0.510 / 0.020 | 0.500 | 0.510 | 0.514 |
| meta-llama/Llama-3.2-3B-Instruct | 3,2 | full | 0.820 | 0.533 / 0.096 | 0.556 / 0.112 | 0.515 | 0.495 | 0.625 |
| microsoft/Phi-3-medium-4k-instruct | 14,0 | 8bit | 0.808 | 0.487 / -0.068 | 0.548 / 0.096 | 0.558 | 0.525 | 0.647 |
| microsoft/Phi-3-small-8k-instruct | 7,4 | full | 0.784 | 0.523 / 0.084 | 0.520 / 0.058 | 0.523 | 0.495 | 0.608 |
| microsoft/Phi-3.5-mini-instruct | 3,8 | full | 0.720 | 0.490 / -0.028 | 0.520 / 0.040 | 0.533 | 0.490 | 0.536 |
| mistralai/Ministral-8B-Instruct-2410 | 8,0 | full | 0.932 | 0.517 / 0.068 | 0.624 / 0.297 | 0.515 | 0.520 | 0.764 |
| mistralai/Mistral-Nemo-Instruct-2407 | 12,2 | 8bit | 0.940 | 0.553 / 0.154 | 0.648 / 0.343 | 0.595 | 0.540 | 0.769 |
| Qwen/Qwen2-0.5B-Instruct | 0,5 | full | 0.632 | 0.500 / 0.000 | 0.494 / -0.018 | 0.500 | 0.500 | 0.500 |
| Qwen/Qwen2-1.5B-Instruct | 1,5 | full | 0.664 | 0.503 / 0.009 | 0.496 / -0.023 | 0.505 | 0.505 | 0.569 |
| Qwen/Qwen2-7B-Instruct | 7,6 | full | 0.888 | 0.510 / 0.046 | 0.564 / 0.171 | 0.608 | 0.505 | 0.692 |
| Qwen/Qwen2.5-0.5B-Instruct | 0,5 | full | 0.500 | 0.503 / 0.008 | 0.496 / -0.045 | 0.500 | 0.500 | 0.508 |
| Qwen/Qwen2.5-1.5B-Instruct | 1,5 | full | 0.752 | 0.487 / -0.028 | 0.540 / 0.094 | 0.503 | 0.510 | 0.533 |
| Qwen/Qwen2.5-3B-Instruct | 3,1 | full | 0.772 | 0.540 / 0.082 | 0.584 / 0.190 | 0.500 | 0.515 | 0.650 |
| Qwen/Qwen2.5-7B-Instruct | 7,6 | full | 0.904 | 0.567 / 0.179 | 0.642 / 0.285 | 0.575 | 0.490 | 0.731 |
| Qwen/Qwen2.5-14B-Instruct | 14,8 | 8bit | 0.924 | 0.587 / 0.174 | 0.658 / 0.339 | 0.600 | 0.525 | 0.758 |
| SherlockAssistant/Mistral-7B-Instruct-Ukrainian | 7,2 | full | 0.836 | 0.557 / 0.114 | 0.560 / 0.122 | 0.510 | 0.535 | 0.639 |
| Vikhrmodels/Vikhr-7B-instruct_0.4 | 7,6 | full | 0.752 | 0.500 / 0.000 | 0.536 / 0.098 | 0.550 | 0.510 | 0.642 |
| Vikhrmodels/Vikhr-Llama3.1-8B-Instruct-R-21-09-24 | 8,0 | full | 0.936 | **0.590** / 0.233 | 0.614 / 0.256 | 0.548 | 0.550 | 0.714 |
| Vikhrmodels/Vikhr-Nemo-12B-Instruct-R-21-09-24 | 12,2 | 8bit | 0.932 | **0.590** / 0.186 | 0.606 / 0.246 | 0.585 | 0.570 | 0.761 |
| claude-3-5-sonnet-20241022 | — | — | 0.888 | 0.667 / 0.383 | 0.754 / 0.540 | 0.585 | 0.515 | 0.656 |
| gpt-4o-2024-11-20 | — | — | **0.976** | **0.720 / 0.506** | **0.864 / 0.742** | **0.848** | **0.740** | **0.914** |

Table 9: Results of LLM evaluation with few-shot prompts in English.