# RusConText Benchmark: A Russian Language Evaluation Benchmark for Understanding Context

Chirkin Andrey[1,2], Kuznetsova Svetlana[1], Volina Maria[1], and Dengina Anna[1]

[1]HSE University, [2]MIPT, Neural Networks and Deep Learning Lab

## Background

The recent LLM benchmarks for Russian, such as MERA (Fenogenova et al., 2024), RussianSuperGLUE (Shavrina et al., 2020) and BABILong (Kuratov et al., 2024), lack short-context understanding tasks, prioritizing general language understanding tasks and reasoning in the long context. RusConText benchmark aims to bridge the gap with tasks tailored to Russian syntactic, discourse and lexical features, enabling precise evaluation of local context interpretation by LLMs.

## Overview

The problem of LLM short-context understanding is that the model should be able to correctly interpret an input text fragment using previous context of at most 1-2 sentences (Zhu et al., 2024). To evaluate LLM performance, four distinct tasks closely related to short context understanding were chosen: coreference resolution, discourse relation identification, idiomatic expression detection and ellipsis resolution. Each task was tested using 4 modern LLMs: GPT-4o-mini, GPT-4.1, Llama-4-Scout, and Qwen-3-30B.

## Coreference

Coreference resolution, involving finding all mentions that refer to the same real-world entity, is a significantly context dependent task. It is particularly complex for the Russian language due to the Russian rich morphology and flexible word order.

Data for coreference task was taken from RuCoCo (Dobrovolskii et al., 2022) corpus and manually annotated. The benchmark coreference task is divided into 2 subtasks:

- **Anaphora Resolution.** Selecting (in multiple choice format) the correct antecedent for pronouns and pronominal adverbs. [500 examples]
- **Coreference Detection**. Determining whether two noun phrases refer to the same entity. [300 examples]

Coreference task is the best performing over benchmark (Accuracy/Precision ≥0.8). The relative difficulty of Anaphora Resolution (task 1) and Coreference Detection (task 2) is unclear.

| Model | Task | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|---|
| gpt-4o-mini | corefAnaphs | 0.786 | 0.786 | 0.786 | 0.786 |
| | corefREs | 0.81 | 0.823 | 0.819 | 0.81 |
| gpt-4.1 | corefAnaphs | 0.904 | 0.904 | 0.905 | 0.904 |
| | corefREs | 0.927 | 0.929 | 0.931 | 0.927 |
| llama-4-scout | corefAnaphs | 0.79 | 0.792 | 0.789 | 0.79 |
| | corefREs | 0.87 | 0.884 | 0.862 | 0.866 |
| qwen-3-30B | corefAnaphs | 0.93 | 0.931 | 0.93 | 0.93 |
| | corefREs | 0.893 | 0.894 | 0.891 | 0.892 |
| random baseline | corefAnaphs | 0.316 | 0.315 | 0.316 | 0.316 |
| | corefREs | 0.515 | 0.516 | 0.516 | 0.515 |

## Ellipsis

Ellipsis Resolution task consists of identifying and reconstructing ellipsis in a sentence to restore its full meaning, and it remains a key NLP challenge, especially in Russian, where elided material often grammatically mismatches its antecedent (Hardt, 2023; Cavar et al., 2024b). Despite advances, SOTA parsers (Stanza, SpaCy) and LLMs still struggle, as they predict word chains rather than reconstruct omissions (Cavar et al., 2024a).

We present a 626-sentence Russian ellipsis corpus, covering different ellipsis types:

- Gapping, NP/VP ellipsis, sluicing, answer/polarity ellipsis (100 each)
- Stripping (14), verb-stranding (3), and mixed cases (9)

The results for the Ellipsis Resolution task evaluation are presented at the table below: all models show low performance overall (F1), gpt-4o-mini outperforms others in accuracy, F1, and ROUGE-1/L, while gpt-4.1 excels in ROUGE-2; qwen-3-30B lags significantly. For ellipsis tasks, gpt-4o-mini is the top performer, but all models show struggle with Ellipsis Resolution task.

Examples with VP/polarity ellipsis (ROUGE >0.35) outperform gapping/sluicing (<0.2). Zero-shot works better than few-shot.

| Model | Accuracy | Precision | Recall | F1 | ROUGE-1 F1 | ROUGE-2 F1 | ROUGE-L F1 |
|---|---|---|---|---|---|---|---|
| gpt-4o-mini | 0.169 | 0.09 | 0.169 | 0.290 | 0.324 | 0.248 | 0.322 |
| gpt-4.1 | 0.139 | 0.064 | 0.139 | 0.244 | 0.394 | 0.297 | 0.390 |
| llama-4-scout | 0.085 | 0.037 | 0.085 | 0.156 | 0.171 | 0.114 | 0.170 |
| qwen-3-30B | 0.02 | 0.012 | 0.012 | 0.012 | 0.101 | 0.075 | 0.101 |

## Idioms

As idiomatic meanings cannot be derived compositionally from the meanings of their individual components, understanding idioms requires significant contextual awareness. Thus, the idiom task was included in the benchmark. It is divided into 3 subtasks:

- Distinguishing between literal and idiomatic uses of potentially idiomatic expressions.
- Determining the specific meaning of polysemous idioms in context.
- Identifying texts that contain a specific meaning of a polysemous idiom.

For the first task, we select examples from an existing corpus of Russian potentially idiomatic expressions (Aharodnik et al., 2018). For the tasks involving polysemous idioms, we use a specifically created dataset of 700 short excerpts, annotated according to its contextual meaning.

The results for all three tasks are presented on the right. Literal/idiomatic distinction is easiest for models; tasks involving polysemous idioms are challenging.

| Model | Task | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|---|
| **gpt-4o-mini** | text | 0.41 | 0.407 | 0.414 | 0.376 |
| | literal/ idiomatic | 0.72 | 0.716 | 0.667 | 0.673 |
| | meaning | 0.65 | 0.333 | 0.217 | 0.263 |
| **gpt-4.1** | text | 0.55 | 0.517 | 0.539 | 0.523 |
| | literal/idiomatic | 0.72 | 0.727 | 0.685 | 0.688 |
| | meaning | 0.77 | 0.5 | 0.385 | 0.435 |
| **llama-4-scout** | text | 0.495 | 0.5 | 0.538 | 0.49 |
| | literal/idiomatic | 0.55 | 0.668 | 0.532 | 0.422 |
| | meaning | 0.64 | 0.5 | 0.32 | 0.39 |
| **qwen-3-30B** | text | 0.495 | 0.5 | 0.538 | 0.49 |
| | literal/idiomatic | 0.55 | 0.668 | 0.532 | 0.422 |
| | meaning | 0.71 | 0.333 | 0.237 | 0.277 |
| **random baseline** | text | 0.33 | 0.318 | 0.312 | 0.305 |
| | literal/idiomatic | 0.54 | 0.542 | 0.543 | 0.537 |
| | meaning | 0.36 | 0.33 | 0.121 | 0.178 |

## Discourse

Identifying discourse relations between sentences reveals LLM ability to recognize logical and semantic connections in text that is crucial for contextual understanding.
Data for the task (2738 samples in total) was collected from 2 datasets, containing manually labeled sentence pairs, and consists of

- 2238 samples from DISRPT (Braud et al., 2024) across 22 possible discourse relation tags,
- 500 samples from RuDABank (Elena Vasileva, 2024) across 15 possible discourse tags.

The results (see Table below) show that best-performing tags are "sequence" (0.62-0.92 accuracy), "neg_answer" (0.96-1.0) and "apology" (0.9-1.0). Tags like "cause-effect", "preparation", "interpretation-evaluation", and "solutionhood" show 0% accuracy in most models, highlighting persistent weaknesses.

| Model | Task | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|---|
| gpt-4o-mini | rudabank | 0.462 | 0.545 | 0.469 | 0.447 |
| | disrpt | 0.272 | 0.178 | 0.206 | 0.166 |
| gpt-4.1 | rudabank | 0.584 | 0.642 | 0.595 | 0.576 |
| | disrpt | 0.388 | 0.306 | 0.284 | 0.258 |
| llama-4-scout | rudabank | 0.415 | 0.565 | 0.426 | 0.379 |
| | disrpt | 0.286 | 0.205 | 0.174 | 0.151 |
| qwen-3-30B | rudabank | 0.392 | 0.483 | 0.4 | 0.382 |
| | disrpt | 0.194 | 0.147 | 0.174 | 0.131 |
| random baseline | rudabank | 0.076 | 0.075 | 0.077 | 0.075 |
| | disrpt | 0.05 | 0.056 | 0.048 | 0.04 |

## Results

The RusConText Benchmark is an automated evaluation tool for assessing LLM short-context understanding on Russian data. While models perform well on standard benchmarks, RusConText reveals specific weaknesses in fine-grained interpretation of compact text segments, which is crucial for real-world applications like dialogue systems and precise information retrieval.

**View on GitHub**

**View on Hugging Face**

## References

1. Alena Fenogenova, et al. 2024. MERA: A comprehensive LLM evaluation in Russian. arXiv preprint arXiv:2401.04531.
2. Yury Kuratov, et al. 2024. BABILong: Testing the limits of LLMs with long context reasoning-in-a-haystack. Advances in Neural Information Processing Systems, 37:106519–106554.
3. Tatiana Shavrina, et al. 2020. RussianSuperGLUE: A russian language understanding evaluation benchmark. arXiv preprint arXiv:2010.15925.
4. Yilun Zhu, et al. 2024. Can large language models understand context? Preprint,arXiv:2402.00858.
5. Vladimir Dobrovolskii, et al. 2022. RuCoCo: a new Russian corpus with coreference annotation. Preprint, arXiv:2206.04925.
6. Daniel Hardt. 2023. Ellipsis-dependent reasoning: a new challenge for large language models. In The 61st Annual Meeting of the Association for Computational Linguistics, pages 39–47. Association for Computational Linguistics.
7. Damir Cavar, et al. 2024a. The typology of ellipsis: a corpus for linguistic analysis and machine learning applications. In Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP, pages 46–54.
8. Damir Cavar, et al. 2024b. Computing ellipsis constructions: Comparing classical NLP and LLM approaches. In Proceedings of the Society for Computation in Linguistics 2024, pages 217–226.
9. Katsiaryna Aharodnik, et al. 2018. Designing a Russian idiom-annotated corpus. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).
10. Chloé Braud, et al. 2024. DISRPT: A multilingual, multi-domain, cross-framework benchmark for discourse processing. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024).
11. Elena Vasileva, et al. 2024. Rudabank. Github: https://github.com/gajka-eva/RuDABank.