RusConText Benchmark: A Russian Language Evaluation Benchmark for Understanding Context

Chirkin A. [1, 2] adchirkin@edu.hse.ru
Kuznetsova S. [1] svkuznetsova_1@edu.hse.ru
Volina M. [1] mavolina@edu.hse.ru
Dengina A. [1] avdengina@edu.hse.ru

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

Abstract

This paper represents an implementation of an approach rather similar to that of Zhu et al. (2024), adapted for the Russian-language data. We introduce the RusConText Benchmark for evaluating short-context understanding in Russian, comprising four distinct yet interrelated tasks: coreference resolution, discourse understanding, idiom interpretation and ellipsis resolution. Each task targets a specific aspect of linguistic processing, challenging a large language model to recover omitted information, resolve referential dependencies, interpret idioms and discourse. The RusConText Benchmark is an additional resource beyond standard benchmarks, designed to assess model performance from a specific perspective. In addition, we present the results of scoring 4 models on our benchmark.

1 Introduction

In the rapidly evolving field of Natural Language Processing (NLP), there is a growing interest in benchmarks as they serve as tools for evaluating the performance and capabilities of large language models (LLMs). Most of the academic LLM benchmarks are designed as a task set that measures LLM efficiency in solving problems, e.g. math or reasoning problems.

As LLMs become increasingly complex and effective in text understanding and generation, assessing their ability to understand context is relevant for ensuring LLM efficiency. Modern models are quite successful at grasping the semantic and logical structure of human-written text; however, their ability to perceive subtle nuances of context remains limited (Zhu et al., 2024). Therefore, benchmarks that evaluate aspects related to contextual understanding are particularly relevant.

Considering the rapid advancement of model capabilities in processing textual information, there is a need to create context-oriented benchmarks that will include more complex and specialized tasks. Although, due to the differences in grammar and discourse across natural languages, it is reasonable to develop unique context understanding benchmarks for evaluating the performance of LLMs across different languages. In this paper, a new context understanding benchmark RusConText is proposed. It is aimed to evaluate LLM performance in processing contextual nuances within the Russian language.

2 Related work

RussianSuperGLUE is considered to be one of the first benchmarks created specifically for the Russian language (Shavrina et al., 2020). It was aimed at evaluating the general language understanding of language models based on the transformer architecture. The main tasks encompass common sense understanding, natural language inference, reasoning, machine reading, and world knowledge. Although it was largely adopted from the SuperGLUE methodology (Wang et al., 2019), some of the tasks were developed from scratch due to the linguistic specificity of Russian. However, this benchmark is mainly intended for smaller transformer models and is not suitable for foundation models that far exceed the capabilities of basic transformers.

To rectify this deficiency, the MERA benchmark has been introduced (Fenogenova et al., 2024). It was aimed at evaluating the performance of the foundation generative models in the Russian language. The benchmark includes 21 evaluation tasks covering a variety of skills including not only reasoning, common sense, mathematics, logic, world knowledge, but also NLI and Dialog System, and as far as language understanding is concerned. So far it can be considered the most reliable tool for the Russian language. However, MERA's focus on a wide range of skills may dilute its sensitivity to specific challenges in parsing sophisticated

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 1158–1170

linguistic structures.

In addition to these benchmarks, there is also TAPE dataset on Russian data, primarily focused on evaluating "intellectual" LLM abilities such as multi-hop reasoning, logical inference, and ethical judgment (Taktasheva et al., 2022) . Another benchmark dataset, RuCoLA, is designed to evaluate language model linguistic competence in the Russian language by classifying sentences as acceptable or unacceptable (Mikhailov et al., 2022). The gold labels are based on native speaker judgments. These datasets complement benchmarks assessing LLM performance in Russian by focusing on more nuanced aspects of language understanding and reasoning abilities.

The need for tools that evaluate how well language models understand complex context has already been addressed in Zhu et al. (2024). The authors have created a benchmark comprising four distinct tasks, namely, coreference resolution, dialogue state tracking, and implicit discourse relation classification, adapting existing datasets for the evaluation of generative models. The choice of tasks is explained by both the growing capabilities of modern LLMs and the real-world applications they are used in. However, it is only available for English and, to the best of our knowledge, does not have any equivalents applicable to Russian.

The BABILong benchmark (Kuratov et al., 2024) is also dedicated to the problem of LLM context understanding. However, the primary objective of this work is to evaluate how effectively LLMs can handle extremely broad contexts. The core focus of this study is to present tasks that require reasoning over lengthy texts in which relevant information is "hidden" among extraneous text. It is a scalable synthetic suite consisting of 20 reasoning tasks, including fact chaining, induction, deduction, counting, and operations involving lists and sets. The principal challenge lies in the extraction and integration of information that is distributed across documents containing up to 10 million tokens or more. So, the main idea of BABILong is to assess how well models utilize their available content window, rather than just a small portion of it. Thus, the emphasis is not on linguistic nuances but rather on the model capacity to manage extensive informational contexts.

Based on the above, there is a need to create a specialized context-oriented benchmark that could be used to evaluate the language capabilities of large language models (LLMs) in Russian in a more comprehensive format. We are guided by initiatives like the work by (Zhu et al., 2024) that demonstrate the possibility to develop a benchmark focused specifically on context processing.

3 RusConText Benchmark: Overview

We formalize the problem of short-context understanding as follows: the model should be able to interpret an entity in the input text using a span of at most one or two sentences or a short paragraph (Zhu et al., 2024). To evaluate the model's performance, we chose a subset of 4 tasks that are closely related to close context understanding: coreference resolution, discourse relation identification, idiomatic expression detection and ellipsis resolution.

Coreference resolution task tests whether a model can identify semantic relations between entities within a given context, a capability essential for maintaining textual coherence and accurately tracking entities across sentences. Discourse relation identification assesses whether the model can recognize logical or text-level semantic connections, such as cause-effect or contrast, which is illustrative for evaluating of the structure and coherence comprehensive understanding. Idiomatic expression detection is a novel approach to LLM deep context understanding evaluation, this perspective is relevant, as the model must integrate information from the immediate and broader context to make a correct judgment, ensuring coherent interpretation of the text parts. Finally, ellipsis resolution evaluates a model's ability to recover information that is implied but not explicitly stated, relying on the immediate context to reconstruct the intended meaning.

3.1 Coreference

Coreference is a linguistic phenomenon that describes the relationship between expressions in a discourse that denote the same entity (or different entities that are semantically related). Coreference resolution is a process of identifying and linking expressions. It is an important and complex NLP problem. The establishment of successful referential connections requires the integration of lexical, syntactic, and discourse-level information, in addition to frequent reliance on extralinguistic common sense. Accurate coreference resolution is essential for thorough text understanding (Poesio et al., 2023).

In addition to the term coreference resolution, the term anaphora resolution can also be found in the literature. Although the terms are often used interchangeably in the NLP-related literature, the tasks they refer to can be distinguished. Anaphora resolution specifically focuses on identifying the antecedents of anaphoric expressions (typically pronouns) (Stylianou and Vlahavas, 2021). Coreference resolution constitutes a broader task that involves identifying both anaphoric and cataphoric connections between a pronoun and its referent, as well as connections between several referential expressions (typically full noun phrases (NPs)) (Kummerfeld and Klein, 2013). In other words, "complete" coreference resolution means finding all mentions that refer to the same real-world entity. Such exhaustive sets of entity mentions are called coreferential chains (Toldova et al., 2016).

There are several common approaches to studying coreference resolution. One such task is the Winograd Schema Challenge (WSC), which was first proposed as an NLP task in the work of (Levesque et al., 2012).

Although this task focused on context understanding, we do not include it in our benchmark, since its variant with Russian data has already been implemented in the Russian SuperGLUE project (Shavrina et al., 2020). Another well-known benchmark for evaluating LLMs on coreference resolution is CRAC (Khosla et al., 2021), which provides tasks on realistic texts than WSC and allows for the assessment of document-level coreference resolution.

Recent research using WSC, CRAC, and CRAC-style benchmarks demonstrates the high performance of modern instruction-tuned LLMs in coreference resolution tasks in few- and zero-shot modes. In the approach described by (Gan et al., 2024), a model is required to identify the antecedent for a given pronoun or referential expression with free-form answers. Open-ended questions provide a comprehensive assessment of a model's effectiveness but require manual verification, which is not suitable for benchmarks. Another approach, outlined in (Le and Ritter, 2023), involves asking a model to tag all entity mentions directly within the text (using different tags for different entities). The authors highlight the issue of unintentional conflation between mention detection and the referential chain annotation.

In this benchmark, we present two distinct tasks. The first task, which focuses on anaphora[1] resolution, is structured in a multiple-choice format. The sets of possible answers are made taking into account the rich morphology of the Russian language (each antecedent option may correlate with the pronoun given in the task). The second task examines the referential relationships between referential expressions (typically NPs). A model answers whether two mentions belong to the same referential chain in True/False mode.

For creating these tasks, we utilized the RuCoCo corpus (Dobrovolskii et al., 2022), a Russian corpus comprised of news texts[2], manually annotated for coreference. The corpus covers a wide range of coreferential and anaphoric relations annotated with a high level of inter-annotator agreement.

To form the tasks, RuCoCo texts were automatically segmented into paragrahs. Then, the paragraphs were filtered to meet task-specific criteria. For the anaphora resolution task (Task 1, corefAnaphs in 1), we selected fragments containing at least one anaphoric pronoun with three or more morphologically compatible non-anaphoric antecedents (within the same fragment). For the coreference detection task (Task 2, corefREs in 1), fragments were required to include at least two referential chains, each with three or more non-anaphoric mentions. Than examples were manually curated from the script output to ensure quality and adherence to linguistic constraints. The first task consists of 500 examples, and the second – 300.

3.2 Discourse

Discourse is a complex term that encompasses a wide range of meanings, generally referring to some kind of connectivity within a text, speech, or other type of linguistic act (Johnstone and Andrus, 2024). Understanding the connection – and, more importantly, the type of such connection -— between two phrases is highly dependent on the context and discourse in which the speech act occurs. This context can depend on knowledge defined outside the text and on common sense.

The study of discourse-related issues of contemporary NLP technologies such as LLMs can improve automatic discourse parsing, highlight

[1]The set of lexemes that we treat as anaphoric pronouns is quite similar to the one described in (Toldova et al., 2016), it is also complemented by some pronominal adverbs with a spatial meaning (such as *zdes'* (*here*), *otkuda* (*from where*))

[2]the corpus contains about a million words drawn from over 3 000 texts, in which 150 000 mentions are annotated

the most problematic types of discourse relations, and help researchers and engineers to make algorithms behave more human-like in the conversation. There are many existing corpora that address discourse-related tasks, available in English Asher et al. (2016) and Russian (Pisarevskaya et al., 2017). In addition, there was an attempt to create a unified discourse corpora that cover multiple languages, frameworks, and domains (Braud et al., 2024).

To evaluate the LLM capabilities on the discourse-related tasks, we employ a set of phrase relation tasks constructed as multi-label choice. The data sources are the Russian language subset of the DISRPT dataset (Braud et al., 2024) and the RuDABank dataset (Elena Vasileva, 2024). Both datasets consists of two sentences and a relation tag that defines the semantic relation between them. The combined corpora consists of 2738 samples (2238 for RuDABank and 500 for DISRPT) and 37 tags (15 for RuDABank and 22 for DISRPT).

3.3 Idioms

Idioms are generally understood as multi-word expressions whose meaning cannot be inferred through the compositional interpretation of constituents. The use of idioms makes the language both more figurative and complex, so that more effort is required for it to be processed even by humans. Thus, in many studies, it has been shown that texts abounding with idiomatic expressions tend to have lower understanding scores, especially among children or learners (Edwards, 1974). As long as idioms can not be processed and understood without sufficient awareness of context we deem it appropriate to use this linguistic phenomenon to evaluate language model capabilities.

The first complexity related to understanding idioms is connected to the fact that certain combinations of words may have literal or idiomatic meaning depending on the context. Expressions of this kind are referred to as Potentially Idiomatic Expressions, or PIEs for short (Haagsma et al., 2020). PIEs have already been used for LLM assessment in English (Mi et al., 2024). To adapt this task to Russian, we have made use of the corpus of 100 Russian PIEs (Aharodnik et al., 2018), previously collected for the task of automatic idiom extraction. From this corpus, we have automatically selected 500 samples. The prompt used to evaluate a model includes, in addition to the base instruction, an idiom, a context, and two options - literal and idiomatic meaning.

The reliance on context while interpreting idioms may be stronger if an idiom has more than one figurative meaning. If so, only in case of thorough understanding of the surrounding context is it possible to deduce the correct meaning of an idiom. To use this suggestion to evaluate LLMs, we have selected 30 idioms possessing between 2 and 4 distinct meanings from the comprehensive dictionary of Russian idioms (Dobrovolskij and Baranov, 2020). The contexts featuring different meanings of the selected idioms were collected with the help of the Russian National Corpus regardless of word insertions, grammatical variations, and omission of non-key components. Thus, we have created a dataset of 500 contexts, labeled with the correct meaning of the idiom used in every entry. The prompt given to a model includes a context, the correct meaning, an alternative meaning of the current idiom, and a meaning of a random idiom from the dataset.

Another version of this task, also requiring from language models an ability to understand and retain larger context, consists of choosing between three texts, all containing the same idiom used in different meanings. The model is given one possible interpretation of the idiom and must identify which text corresponds to that specific meaning. To make the task more challenging, only idioms having three or more meanings were included.

3.4 Ellipsis

Ellipsis is a group of phenomena in which unexpressed information from a discourse can be recovered from the context (Testelets, 2011), distinguishing it from elision, which relies on extralinguistic knowledge rather than context. Since elliptical constructions lack overtly expressed components necessary for understanding, this information must be supplied from the context within which the sentence occurs (Thomas, 1979).

Studying ellipsis resolution is important for improving the accuracy of NLP systems that handle large data with ellipsis constructions (Zhang et al., 2019). However, in the field of NLP, problems related to the phenomenon of ellipsis still cause difficulties, as machines always struggle with the omitted and ambiguous information, and there is still a lack of research, corpus data and materials to solve the problems of ellipsis resolution, especially for the Russian language (Hardt, 2023; Ćavar et al., 2024b). The difficulty of restoring the elided material for the Russian language is that it does

not always coincide with the antecedent in its form. For example, the grammatical features (such as person or number) of the omitted verb do not always correspond to those of the verb in another clause.

To address these challenges, various instruments have been developed for Ellipsis Resolution task, ranging from rule-based parsers to modern machine learning approaches. For the detection of the antecedent of the ellipsis and the ellipsis site itself, SOTA parsers are commonly used. However, Cavar and Holthenrichs (2024) state that "common state-of-the-art NLP pipelines fail", including Stanza, SpaCy, and LFG parsers. For the Ellipsis Resolution task, LLMs remain the best solution, although they still struggle, because they are trained to suggest word chains rather than fill in the omitted words and phrases (Ćavar et al., 2024a).

To assess the performance of LLMs in Ellipsis Resolution, we constructed a specialized corpus containing constructions of various types of ellipsis for Russian language. This corpus consists of 626 sentences, containing such ellipsis constructions as gapping, NP ellipsis, VP ellipsis, sluicing, answer ellipsis, polarity ellipsis (100 sentences each), stripping (14 sentences), verb-stranding (3 sentences) and 9 sentences with a combination of different ellipsis types.

The data for the corpus was taken from existing ellipsis corpora for Russian or from articles about ellipsis in the Russian language, was manually selected by the author from the Russian National Corpus, created or elicited by the author. To find out the source of the sentence, see the source column in the ellipsis corpus[3].

4 Evaluation

The RusConText Benchmark[4] comprises multiple subsets, each represented as JSON or CSV files corresponding to different linguistic tasks:

- `coref__anaph_ref_choice_questions.json` – Question-based anaphora resolution
- `coref__are_NPs_coref_task.json` – Coreference detection for noun phrases
- `disrpt.json` – Discourse relation parsing
- `rudabank.csv` – Discourse relation parsing
- `idiom_literal.json` – Literal vs. idiomatic interpretation
- `idiom_text.json` – Idiom disambiguation across contexts
- `idiom_meaning.json` – Polysemous idiom resolution
- `ellipsis.csv` – Ellipsis identification and resolution

The tasks vary in complexity, ranging from multi-label classification (e.g., coreference resolution) to structured prediction (e.g., ellipsis restoration, requiring models to identify elided content and infer it from context). The examples of these tasks can be found in the Appendix A.

4.1 Evaluation Metrics

We assess model performance using:

- Standard classification metrics: $Precision$, $accuracy$, $recall$, and $F1\ score$ for discrete-label tasks.
- $ROUGE$ (Lin, 2004) for evaluating generated text in ellipsis resolution.

4.2 Models and Implementation

We evaluate a suite of state-of-the-art language models for comparability:

- GPT-4o-mini (OpenAI, 2024)
- GPT-4.1 (OpenAI, 2025)
- Llama-4-Scout (Touvron et al., 2023)
- Qwen-3-30B (Yang et al., 2025)

Models were accessed via the LangChain framework (Chase, 2022) using a unified Python pipeline. Selection criteria included benchmark performance parity and source diversity to include open-source models as well as closed ones. Each model was trained on the mixture of multiple languages including Russian. Each model was asked to return a valid JSON string, the responses that could not be salvaged were considered as wrong answers. Temperature of generation was set to 0, other parameters were default to the models. Prompts that were used for each task can also be observed in Appendix B. The "random baseline" (obtained by uniform random sampling from the possible choices) serves as a lower-bound reference for LLM performance.

[3] https://github.com/NotBioWaste905/RuConText-Bench/blob/main/data/ellipsis.csv

[4] https://github.com/NotBioWaste905/RuConText-Bench/blob/main/data

4.3 Results

The LLM evaluation results for all tasks except for the ellipsis task are shown in table 1, the results for the ellipsis task are displayed in table 2. The first column indicates the evaluated model.

The ellipsis task remained difficult for all models resulting in low $F1\ score$ across all models. It was unexpected that zero-shot prompts slightly improved the results of the models' ellipsis resolution, while in Ćavar et al. (2024b) few-shot prompts gave better results, increasing the accuracy, but their results were consistent with ours in that LLMs still struggle with ellipsis resolution. The improvement in results using zero-shot prompts can be explained by the fact that if the model receives a specific example with a certain type of ellipsis and the position of the ellipsis in the sentence (for example, the models were very sensitive to the position of the ellipsis at the end of the sentence), the model might overfit to these examples. The analysis was conducted based on ROUGE scores, with relatively good results (ROUGE > 0.35) observed for VP ellipsis (65%), polarity ellipsis (55%), and NP ellipsis (54%) types. In contrast, low results (ROUGE < 0.2) were seen for gapping (86%), sluicing (65%), and NP ellipsis (43%) types. Notably, while NP ellipsis appeared in both categories, its performance varied significantly, suggesting that resolution success depends on contextual factors rather than the ellipsis type. The model struggled the most with gapping, indicating a major challenge in handling this type of ellipsis.

The discourse tasks have also posed difficulties to the models, primarily the DISRPT subset. We suppose that the main struggle for the model is juggling more than 20 possible tags in a single prompt, many of which are very similar in their meaning. After evaluating the models on the DISRPT and RuDABank discourse subsets we were able to identify the classes models struggle most with. For the DISRPT subset "sequence" is consistently the best-predicted tag (61.5-92.3% accuracy), indicating strong performance on this common structure. Challenging tags like "cause-effect", "preparation", "interpretation-evaluation", and "solutionhood" show 0% accuracy in most models, highlighting persistent weaknesses. Performance variability is also significant: GPT-4.1 excels overall (e.g., 92.3% "sequence", 78.9% "condition"), while Qwen3-30b uniquely handles "cause-effect" (50%) but fails completely on "evaluation" and "evidence". We also hypothesize that such results can be related to our prompting method (passing all possible labels in one prompt) — the LLM has much harder time choosing between similar tags and can make a decision that discriminates less "prototypical" tag. As for the results for the RuDABank subset all models excel at recognizing "neg_answer" (0.96-1.0) and "apology" (0.9-1.0), with "pos_answer" also strong in most models (0.64-0.91). At the same time "backchanneling" (0.02-0.22) and "yes_no_question" (0.1-0.27) are challenging for all models as well as "other_answers" which is near-zero in three models. The surprisingly low accuracy result in "yes_no_question" can possibly be explained by model mistakes it for "open_question" because Russian language lacks "yes/no" question markers that cannot be easily omitted. The existence of negation and apology markers possibly can explain the high results of the top tags.

In coreference resolution tasks, LLMs demonstrate the highest effectiveness among all segments of the benchmark. In these tasks, the metrics of Accuracy and Precision for all tested models are close to 0.8 or significantly higher.

Finally, in idioms-related tasks, we can observe that, for the most part, models perform relatively similarly. However, the differences in scores still allow us to differentiate between models and select the strongest one for each task. For some models the task of choosing between literal and idiomatic meanings turned out to be the easiest, which can be explained by the fact that potentially idiomatic expressions have typical surrounding contexts depending on whether they are used idiomatically or not, and models may have retained this information during training. As for the tasks involving polysemous idioms, the one including texts proved to be more challenging, evidently because it requires simultaneous processing of all three contexts in order to be solved successfully.

5 Conclusions

The RusConText Benchmark is designed to evaluate LLM short-context understanding for Russian, addressing a gap in existing evaluation frameworks. While many benchmarks focus on broad reasoning tasks or long-context comprehension, our approach specifically targets the model ability to interpret and reason within constrained text intervals — a competence essential for real-world applications

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
	corefAnaphs	0.786	0.786	0.786	0.786
	corefREs	0.81	0.823	0.819	0.81
	idioms_text	0.41	0.407	0.414	0.376
	idioms_literal	0.72	0.716	0.667	0.673
	idioms_meaning	0.65	0.333	0.217	0.263
gpt-4.1	rudabank	0.584	0.642	0.595	0.576
	disrpt	0.388	0.306	0.284	0.258
	corefAnaphs	0.904	0.904	0.905	0.904
	corefREs	0.927	0.929	0.931	0.927
	idioms_text	0.55	0.517	0.539	0.523
	idioms_literal	0.72	0.727	0.685	0.688
	idioms_meaning	0.77	0.5	0.385	0.435
llama-4-scout	rudabank	0.415	0.565	0.426	0.379
	disrpt	0.286	0.205	0.174	0.151
	corefAnaphs	0.79	0.792	0.789	0.79
	corefREs	0.87	0.884	0.862	0.866
	idioms_text	0.495	0.5	0.538	0.49
	idioms_literal	0.55	0.668	0.532	0.422
	idioms_meaning	0.64	0.5	0.32	0.39
qwen-3-30B	rudabank	0.392	0.483	0.4	0.382
	disrpt	0.194	0.147	0.174	0.131
	corefAnaphs	0.93	0.931	0.93	0.93
	corefREs	0.893	0.894	0.891	0.892
	idioms_text	0.495	0.5	0.538	0.49
	idioms_literal	0.55	0.668	0.532	0.422
	idioms_meaning	0.71	0.333	0.237	0.277
random baseline	rudabank	0.076	0.075	0.077	0.075
	disrpt	0.05	0.056	0.048	0.04
	corefAnaphs	0.316	0.315	0.316	0.316
	corefREs	0.515	0.516	0.516	0.515
	idioms_text	0.33	0.318	0.312	0.305
	idioms_literal	0.54	0.542	0.543	0.537
	idioms_meaning	0.36	0.33	0.121	0.178

Table 1: Comparison of LLM performance across tasks.

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

Table 2: Comparison of LLM performance across ellipsis task.

such as conversational AI, summarization, and precise information retrieval. The RusConText Benchmark shows that modern LLMs may still struggle to solve problems related to understanding the close context.

Our results demonstrate that while leading LLMs perform well on established benchmarks for Russian data (even on that are conceptually aligned with some of ours, such as RWSC in (Shavrina et al., 2020) or RCB in (Fenogenova et al., 2024)),

their performance on the RusConText Benchmark reveals key weaknesses in fine-grained understanding of context.

Limitations

The limitations of the RusConText Benchmark are primarily in its scope: the tasks presented in this benchmark — resolution of coreference, metaphor and ellipsis, as well as discourse understanding by the model — do not reflect the full variety of contextual tasks. Additionally, we aim to significantly expand the size of each dataset in the future.

It must also be noted that model scoring results largely depend on prompt engineering (especially for zero-shot question answering approach, which we are mostly following), and although we have selected prompts that helped us achieve maximum accuracy received during the tests, these prompts may not be universal or ideal.

References

Katsiaryna Aharodnik, Anna Feldman, and Jing Peng. 2018. Designing a russian idiom-annotated corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the stac corpus. In *10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2721–2727.

Chloé Braud, Amir Zeldes, Laura Rivière, Yang Janet Liu, Philippe Muller, Damien Sileo, and Tatsuya Aoyama. 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)*.

Damir Cavar and Van Holthenrichs. 2024. On ellipsis in slavic: The ellipsis corpus and natural language processing results. In *Formal Approaches to Slavic Linguistics*.

Damir Ćavar, Ludovic Mompelat, and Muhammad Abdo. 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.

Damir Ćavar, Zoran Tiganj, Ludovic Veta Mompelat, and Billy Dickson. 2024b. Computing ellipsis constructions: Comparing classical nlp and llm approaches. In *Proceedings of the Society for Computation in Linguistics 2024*, pages 217–226.

Harrison Chase. 2022. LangChain.

Vladimir Dobrovolskii, Mariia Michurina, and Alexandra Ivoylova. 2022. Rucoco: a new russian corpus with coreference annotation. *Preprint*, arXiv:2206.04925.

Dmitrij Dobrovolskij and Anatolij Baranov. 2020. Akademicheskij slovar' russkoj frazeologii.

Peter Edwards. 1974. Idioms and reading comprehension. *Journal of Reading Behavior*, 6(3):287–293.

Denis Shcherbatov Elena Vasileva. 2024. Rudabank.

Alena Fenogenova, Artem Chervyakov, Nikita Martynov, Anastasia Kozlova, Maria Tikhonova, Albina Akhmetgareeva, Anton Emelyanov, Denis Shevelev, Pavel Lebedev, Leonid Sinev, and 1 others. 2024. Mera: A comprehensive llm evaluation in russian. *arXiv preprint arXiv:2401.04531*.

Yujian Gan, Juntao Yu, and Massimo Poesio. 2024. Assessing the capabilities of large language models in coreference: An evaluation. In *Joint 30th International Conference on Computational Linguistics and 14th International Conference on Language Resources and Evaluation, LREC-COLING 2024*, pages 1645–1665. European Language Resources Association (ELRA).

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. Magpie: A large corpus of potentially idiomatic expressions. In *12th Language Resources and Evaluation Conference: LREC 2020*, pages 279–287. European Language Resources Association (ELRA).

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.

Barbara Johnstone and Jennifer Andrus. 2024. *Discourse analysis*. John Wiley & Sons.

Sopan Khosla, Juntao Yu, Ramesh Manuvinakurike, Vincent Ng, Massimo Poesio, Michael Strube, and Carolyn Rosé. 2021. The codi-crac 2021 shared task on anaphora, bridging, and discourse deixis in dialogue. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 1–15.

Jonathan K Kummerfeld and Dan Klein. 2013. Error-driven analysis of challenges in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 265–277.

Yury Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. 2024. Babilong: Testing the limits of

llms with long context reasoning-in-a-haystack. *Advances in Neural Information Processing Systems*, 37:106519–106554.

Nghia T Le and Alan Ritter. 2023. Are large language models robust coreference resolvers? *arXiv preprint arXiv:2305.14489*.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. *KR*, 2012:13th.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Maggie Mi, Aline Villavicencio, and Nafise Sadat Moosavi. 2024. Rolling the dice on idiomaticity: How llms fail to grasp context. *arXiv preprint arXiv:2410.16069*.

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*, page 5207–5227. Association for Computational Linguistics.

OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence. Accessed: 2025-05-19.

OpenAI. 2025. Introducing GPT-4.1 in the API. Accessed: 2025-05-19.

Dina Pisarevskaya, Margarita Ananyeva, Maria Kobozeva, Alexander Nasedkin, Sofia Nikiforova, Irina Pavlova, and Alexey Shelepov. 2017. Towards building a discourseannotated corpus of russian. In *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, pages 201–212.

Massimo Poesio, Juntao Yu, Silviu Paun, Abdulrahman Aloraini, Pengcheng Lu, Janosch Haber, and Derya Cokal. 2023. Computational models of anaphora. *Annual Review of Linguistics*, 9(1):561–587.

Tatiana Shavrina, Alena Fenogenova, Anton Emelyanov, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. Russiansuperglue: A russian language understanding evaluation benchmark. *arXiv preprint arXiv:2010.15925*.

Nikolaos Stylianou and Ioannis Vlahavas. 2021. A neural entity coreference resolution review. *Expert Systems with Applications*, 168:114466.

Ekaterina Taktasheva, Tatiana Shavrina, Alena Fenogenova, Denis Shevelev, Nadezhda Katricheva, Maria Tikhonova, Albina Akhmetgareeva, Oleg Zinkevich, Anastasiia Bashmakova, Svetlana Iordanskaia, Alena Spiridonova, Valentina Kurenshchikova, Ekaterina Artemova, and Vladislav Mikhailov. 2022. Tape: Assessing few-shot russian language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, page 2472–2497. Association for Computational Linguistics.

Yakov Testelets. 2011. Ellipsis v russkom yazyke: teoreticheskije i opisatel'nyje podhody. In *Conference « Typology of morphosyntactic parameters»: presentation. — M.: MSU, 2011.*

Andrew L Thomas. 1979. Ellipsis: the interplay of sentence structure and context. *Lingua*, 47(1):43–68.

Svetlana Toldova, Ilya Azerkovich, Alina Ladygina, Anna Roitberg, and Maria Vasilyeva. 2016. Error analysis for anaphora resolution in Russian: new challenging issues for anaphora resolution task in a morphologically rich language. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, pages 74–83, San Diego, California. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Wei-Nan Zhang, Yue Zhang, Yuanxing Liu, Donglin Di, and Ting Liu. 2019. A neural network approach to verb phrase ellipsis resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7468–7475.

Yilun Zhu, Joel Ruben Antony Moniz, Shruti Bhargava, Jiarui Lu, Dhivya Piraviperumal, Site Li, Yuan Zhang, Hong Yu, and Bo-Hsiang Tseng. 2024. Can large language models understand context? *Preprint*, arXiv:2402.00858.

A Task examples

A.1 Coreference

A.1.1 Antecedent search (corefAnaphs)

Task example


```
"paragraph": {
  "filename": "2021_sport_pony.json",
  "index": 3,
  "text": "Отмечается, что ранее в
```

социальных сетях его сыном было опубликовано видео, где да Силва ездит верхом на пони по имени Пикулито, при этом чрезмерно дергает поводьями, причиняя лошади боль. Также решением трибунала было установлено, что за пони бежала собака, причиняя животному стресс на фоне испытываемых болезненных ощущений, а сам да Силва был слишком тяжел для лошади."
},
"anaphoric span": "его",
"variants": [
"да Силва",
"видео",
"пони по имени Пикулито"
],
"gold answer": "1"

A.1.2 NP coreference (corefREs)

The fields {fisrt} and {second} correspond "the first RE (NP) span" and "the second RE (NP) span respectively".

Task example
"first": "совершенно легальный пиратский интернет-сервис", "second": "сайта", "paragraph": { "filename": "2009_hitech_antigua.json", "index": 1, "text": "На острове Антигуа открылся "совершенно легальный пиратский интернет-сервис". Администрация сайта утверждает, что в закромах имеется полторы тысячи кинофильмов и 50 тысяч музыкальных композиций. Желающие их скачать должны оформить подписку стоимостью 9,95 доллара в месяц." }, "gold": true

A.2 Discourse

A.2.1 DISRPT

Task example
Sentence 1: В этой статье решено привести обобщение алгоритмического базиса Sentence 2: которые могут быть описаны одной или несколькими дугами кривых, для всех случаев пространств координат. Label: "elaboration" Choices: preparation, condition, antithesis, solutionhood, restatement, cause, effect, attribution, sequence, evaluation, evidence, interpretation-evaluation, cause-effect, elaboration, background, conclusion, motivation, concession, comparison, purpose, contrast, joint

A.2.2 RuDABank

Task example
Sentence 1: Соответственно, сегодня ночью мы не спим Sentence 2: Пап, это отличная идея. Label: "appreciation" Choices: statement, open_question, other_answers, yes_no_question, pos_answer, neg_answer, appreciation, disapproval, command, avoiding, opening, closing, thanking, back-channeling, apology

A.3 Idioms

A.3.1 idiom_literal

Task example
Idiom: ловить блох Text: На полках стояли кабинетные часы из бронзы и мрамора и современные будильники, а в углу монументально возвышались большие напольные часы. Антон заметил прислоненные к стене костыли. Я б и сам так думал, — сказал часовщик. Что ж блох ловить, если сила есть. Он опустил лупу на глаз и стал копаться в часах. Потом сказал: Ты бы оставил их, я проверю. Label: 1 Meaning: idiomatic

A.3.2 idiom_text

Task example

Idiom: играть в бирюльки
Texts:
1. "И я думаю, что сигналы такого рода... Государство – серьезная штука. Не надо игнорировать государство. Не надо играть с ним в бирюльки и азартные игры. "
2. "Мы у тебя из спины куски кожи будем вырезать и солью посыпать, если соврешь. И еще, много орешь, старый пень! Придется тебе рот заклеить... Петрович, принеси скотч и приступай. Хватит с ним в бирюльки играть."
3. "Не подпадало дела настоящего, да и только! Ну, а в бирюльки играть был он не охотник. Всякий, конечно, норовил охаять... "
Label: 1
Meaning: относиться несерьезно к кому-либо

A.3.3 idiom_meaning

Task example
Idiom: бок о бок Possible meanings: 1. вместе, совместно 2. выражать незнание ответа на заданный вопрос 3. очень близко, один возле другого Label: 0 Meaning: вместе, совместно Example: Ярким примером являются водители и переводчики, которые наряду с военными бок о бок участвуют в Сирии по сути на передовой боевых действий.

A.4 Ellipsis

Task example
Sentence: Работа с двухбайтовыми наборами символов — просто кошмар для программиста, так как часть их состоит из одного байта, а часть — __ из двух. label: состоит ellipsis type: gapping

B Prompts

B.1 Coreference

B.1.1 Antecedent search (corefAnaphs)

Prompt
Ответь на вопрос по этому фрагменту текста: {paragraph}. Тебе нужно понять, к какой сущности относится это упоминание: {anaphoric span}. Из предложенных ниже выбери упоминание, которое тоже относится к этой сущности. Варианты ответа: {variants} Напиши только варинат ответа, 1, 2 или 3, без комментариев и знаков препинания.

Prompt translation
Answer a question about this text fragment: {paragraph}. You need to determine which entity this mention refers to: {anaphoric span}. From the options below, select the mention that refers to the same entity. Answer choices: {variants} Write only the answer option, 1, 2 or 3, without any comments or punctuation marks.

B.1.2 NP coreference (corefREs)

The fields {fisrt} and {second} correspond "the first RE (NP) span" and "the second RE (NP) span respectively".

Prompt
В тексте: {paragraph} упоминания (подстроки) {first} и {second} отсылают к одной и той же сущности? Отвечай True, если да, False если нет, без знаков препинания и дополнительных комментариев.

Prompt translation
In the text: {paragraph} do the mentions (substrings) {first} and {second} refer to the same entity? Answer True if yes, False if no, without punctuation or additional comments.

B.2 Discourse

Relevant for both datasets (DISRPT and RuDABank).

Prompt

Определите связь между двумя предложениями. Возможные следующие варианты ответа:
{options}.
Предложение 1: {sent_1}
Предложение 2: {sent_2}
Дайте только один ответ из предложенных. Используйте JSON для вывода, состоящий из одного поля: "answer".

Дано начальное высказывание и ответное высказывание, определите тип ответа из следующих вариантов:{options}

Начальное высказывание: {initial_utterance}
Ответное высказывание: {tagged_utterance}

Дайте только один ответ из предложенных. Используйте JSON для вывода, состоящий из одного поля: "answer".

Prompt translation

Determine the relationship between two sentences. The following are possible answer options:
{options}.
Sentence 1: {sent_1}
Sentence 2: {sent_2}
Give only one answer from the suggested ones. Use JSON for output, consisting of one field: "answer".

Given an initial statement and a response statement, determine the type of answer from the following options:{options}

Initial statement: {initial_utterance}
Response statement: {tagged_utterance}

Give only one answer from those suggested. Use JSON for output, consisting of one field: "answer".

B.3 Idioms

B.3.1 Idiom_literal

Prompt

Задание: Определи, используется ли выражение в прямом или переносном смысле.
Выражение: {idiom}
Контекст: {example}
Варианты ответа: 0 - буквальное значение, 1 - переносное значение
Ответ:

Prompt translation

Task: Determine whether the expression is used literally or figuratively.
Expression: {idiom}
Context: {example}
Answer options: 0 - literal meaning, 1 - figurative meaning
Answer:

B.3.2 idiom_text

Prompt

Задание: Определи, в каком тексте выражение имеет указанное значение.
Выражение: {idiom}
Значение: {current_meaning}
Тексты: {texts}
Ответ:

Prompt translation

Task: Identify which text contains the expression with the specified meaning.
Expression: {idiom}
Meaning: {current_meaning}
Texts: {texts}
Answer:

B.3.3 idiom_meaning

Prompt

Задание: Определи, какое значение соответствует данному выражению в данном контексте.
Выражение: {idiom}
Контекст: {example}
Варианты ответа: {possible_meanings}
Ответ:

Prompt translation

Task: Determine which meaning

corresponds to the given expression in this context.
Expression: {idiom}
Context: {example}
Answer options: {possible_meanings}
Answer:

B.4 Ellipsis

Prompt
Дано предложение {text}. Оно содержит эллипсис, в нем пропущена часть информации. Постарайся восполнить как можно больше информации, не придумывай и не добавляй того, чего нет в контексте. Определи, 1) в каком месте пропущена информация, обозначь это место нижним подчеркиванием. 2) Восполни информацию и 3) напиши новое предложение с восполненой информацией. Ответ дай в формате: изначальное - ответ на 1, эллипсис - ответ на 2, полное - ответ на 3. Ответ должен быть в формате json. В ответе должен быть только JSON в markdown нотации (начинаться с ``` json и заканчиваться ```) без дополнительных комментариев.

Prompt translation
Given the sentence {text}. It contains ellipsis, some information is omitted. Try to recover as much information as possible without inventing or adding anything beyond the context. Determine: 1) where information is missing (mark this place with an underscore), 2) recover the omitted information, and 3) write a new sentence with the recovered information. Provide the answer in the format: original - answer to 1, ellipsis - answer to 2, complete - answer to 3. The response must be in JSON format. Include only JSON in markdown notation (starting with ```json and ending with ```) without additional comments.