Long Input Benchmark for Russian Analysis

Anonymous ACL submission

Abstract

Recent advancements in Natural Language Processing (NLP) have fostered the development of Large Language Models (LLMs) that can solve an immense variety of tasks. One of the key aspects of their application is their ability to work with long text documents and to process long sequences of tokens. This has created a demand for proper evaluation of long-context understanding. To address this need for the Russian language, we propose LIBRA (Long Input Benchmark for Russian Analysis), which comprises 21 adapted datasets to study the LLM's abilities to understand long texts thoroughly. The tests are divided into four complexity groups and allow the evaluation of models across various context lengths ranging from 4k up to 128k tokens. We provide the open-source datasets, codebase, and public leaderboard for LIBRA to guide forthcoming research.

1 Introduction

Large Language Models (LLMs) have demonstrated impressive abilities in many NLP applications. Interacting with people through free-form text instructions, they serve as versatile tools for multiple scenarios, transforming the landscape of AI systems. One direction where LLM usage is developing rapidly includes tasks requiring long text processing, such as summarization and information extraction, where their applications alleviate the handling of long texts for humans.

However, until recently, most LLMs had difficulties in handling long sequences of tokens and were only able to work with a limited context length of several thousand tokens. In recent years, new methods have enabled the models to increase their context significantly, empowering them to solve a new variety of tasks. This, in turn, and the community's demand for automatic systems solving such tasks at a good level has created a need for a thorough evaluation of LLM long context understanding.

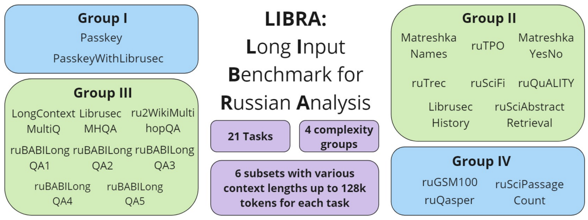

Figure 1: The illustration of the LIBRA benchmark.

To address this demand in English, several long context understanding benchmarks have been created recently with LongBench (Bai et al., 2023)[1] and L-Eval (An et al., 2023)[2] heading the list. However, the Russian language, at this point, lacks a fair instrument for transparent evaluation of long context understanding.

Our work addresses this problem and presents a new benchmark, which we call **L**ong **I**nput **B**enchmark for **R**ussian **A**nalysis, or LIBRA, for the evaluation of LLM long context understanding abilities in Russian (see Figure 1 for LIBRA general structure).

Thus, the contribution of our work can be summarized as follows:

- we present a methodology for the evaluation of long-context abilities of LLMs for the Russian language;
- we publicly release a set of 21 datasets of various skills and complexities in Russian which form the LIBRA benchmark;
- we provide a codebase as long as the number of baseline solutions and public leaderboard[3].

[1] https://huggingface.co/datasets/THUDM/LongBench
[2] https://huggingface.co/datasets/L4NLP/LEval
[3] The link was removed to preserve anonymity during the review period.

2 Related Work

2.1 Long Context Large Language Models

One of the important tasks in the development of LLMs is to increase the length of the context that the model can understand. This problem has two key points: the complexity of calculations for long sequences and the ability of the model to extract important data in a long context. The solution of the first problem can be attributed to research on the effective processing of the self-attention as in Longformer (Beltagy et al., 2020), LongNet (Ding et al., 2023) and FlashAttention (Dao et al., 2022; Dao, 2023), using caches for previously calculated outputs such as Transformer-XL (Dai et al., 2019), Unlimiformer (Bertsch et al., 2024) and LongLLaMA (Tworkowski et al., 2024) or replacing it with another mechanism with more effective inference as in RetNet (Sun et al., 2023) and Mamba (Gu and Dao, 2023). The solution to the second problem is to improve positional encoding techniques such as ALiBi (Press et al., 2021) and RoPE-based approaches (Sun et al., 2022; Peng et al., 2023).

2.2 Long Context Benchmarks

Until recently, most LMs had relatively small context lengths limited by a few thousand tokens. Thus, standard Natural Language Understanding (NLU) benchmarks (Wang et al., 2018, 2019; Shavrina et al., 2020) contained tasks within this size.

Even today, many "new generation" benchmarks created recently, such as HELM (Bommasani et al., 2023), MT-Bench (Zheng et al., 2023), and Russian-oriented benchmark MERA (Fenogenova et al., 2024) follow this pattern, limiting their tasks by relatively small context window size to simplify the evaluation procedure and reducing its cost.

The pioneers of long context processing benchmarks have been ZeroSCROLLS (Shaham et al., 2023)[4], designed to test zero-shot model capabilities for NLU over long texts; L-eval (An et al., 2023)[5], focused on a standardized evaluation methodology for long context LMs addressing two key aspects: dataset construction and evaluation metrics; and LongBench (Bai et al., 2023), the bilingual multi-task benchmark for long context understanding, comprising 21 tasks in English and Chinese. The tasks in LongBench can be divided into 6 big categories and cover key long-text application scenarios, including multi-document QA, single-document QA, summarization, few-shot learning, code completion, and synthesis tasks.

However, the limitation of the long context benchmarks mentioned above is that they are mainly oriented at the English language (and the Chinese language for LongBench). As for the Russian language, there is an urgent need for a reliable system able to evaluate LLM long context understanding abilities. To address this problem, we propose LIBRA, which brings a methodology and 21 tasks for a long context understanding evaluation in Russian.

3 LIBRA

3.1 Benchmark Overview

In this section, we introduce LIBRA (Long Input Benchmark for Russian Analysis), a new benchmark for long context understanding in Russian, which includes 21 tasks for LLM evaluation. LIBRA aims to evaluate a large scope of LLMs, including pretrain models and models with supervised finetuning (SFT) with any system prompt that can be picked up.

The main purpose of the benchmark is to create a reliable instrument for the long context understanding evaluation, enabling the study of the model's ability to solve various tasks of different complexity with respect to the input context length. For this purpose, all tasks in the LIBRA benchmark are divided into 4 complexity groups, and the datasets have several subsets of various context lengths ranging from 4k up to 128k tokens[6]. The latter makes it possible to explore the influence of the context length on the model results.

3.2 Complexity group description

In this section, we describe each of the complexity groups of tasks.

The first complexity group (I) consists of tasks that require finding a short text fragment in long textual paragraphs containing irrelevant information. This group includes Passkey and PasskeyWithLibrusec datasets.

The second complexity group (II) includes tasks that require answering the question based on a relevant context. The following types of tasks are related to this group: question answering (QA) such as MatreshkaNames, MatreshkaYesNo, LibrusecHistory, ruTREC, ruSciFi, ruSciAbstractRe-

[4] https://www.zero.scrolls-benchmark.com/

[5] https://huggingface.co/papers/2307.11088

[6] See explanation on token length calculation in Section 3.3

	Task Name	Data Origin	Skills	Metric	Dataset Size
I	Passkey	Translated	Reasoning	EM	1200
	PasskeyWithLibrusec	New	Reasoning	EM	1200
II	MatreshkaNames	New	Dialogue Context, Reasoning	EM	900
	MatreshkaYesNo	New	Dialogue Context, Reasoning	EM	1799
	LibrusecHistory	New	Reasoning	EM	128
	ruTREC	Translated	Reasoning	EM	300
	ruSciFi	Translated	World Knowledge, Reasoning	EM	64
	ruSciAbstractRetrieval	New	Reasoning	EM	1240
	ruTPO	Translated	Exam, Reasoning	EM	251
	ruQuALITY	Translated	Reasoning	EM	202
III	LongContextMultiQ	New	Reasoning	EM	1200
	LibrusecMHQA	New	Reasoning	EM	384
	ru2WikiMultihopQA	Translated	Reasoning	EM	300
	ruBABILongQA1	Adapted	Reasoning	EM	600
	ruBABILongQA2	Adapted	Reasoning	EM	600
	ruBABILongQA3	Adapted	Reasoning	EM	600
	ruBABILongQA4	Adapted	Reasoning	EM	600
	ruBABILongQA5	Adapted	Reasoning	EM	600
IV	ruSciPassageCount	New	Reasoning	EM	600
	ruQasper	Translated	Reasoning	F1	203
	ruGSM100	Translated	Math, Logic	EM	100

Table 1: The LIBRA tasks outline. The numbers **I**, **II**, **III**, and **IV** in the left column indicate the complexity group of the tasks described in Subsection 3.2. The **Skills** column defines the skills to be tested on a specific task. **Data Origin** discloses the source of the dataset. The **Dataset Size** column shows the number of items in the whole dataset.

trieval and multiple choice QA tasks, which are presented by ruTPO and ruQuALITY.

The natural development of tasks from the second class of complexity are tasks with questions, the answers to which are not explicitly contained in the text but require the analysis of fragments of input data and the generation of an answer based on it. Such tasks in our classification belong to **the third complexity group (III)** and represent a multi-hop question answering (MHQA) type. This group includes the following tasks: ruBABILongQA1, ruBABILongQA2, ruBABILongQA3, ruBABILongQA4, ruBABILongQA5, LongContextMultiQ, LibrusecMHQA and ru2WikiMultihopQA.

Finally, to **the fourth complexity group (IV)** belongs to the tasks that require understanding the whole context, solving mathematical problems, and QA tasks within complex domains. This group includes ruSciPassageCount, ruGSM100 and ruQasper datasets.

It should also be mentioned that we do not include code generation and analysis tasks in LIBRA as most of the software code in the world is written in languages based on English.

3.3 Context Length Estimation

In the LIBRA benchmark, we divide all datasets into subsets of various context lengths. We measure context length in tokens; however, it may vary across different models and tokenizers. In our work, we used the fertility of tokenizers to distribute samples across different context lengths, which indicates the average number of tokens in which one word is tokenized. Thus, the average length in tokens for the text can be approximated by the number of words multiplied by the fertility number.

For the fertility approximation, we calculate the average fertility of the classic LLM tokenizers, which we further evaluate as baselines (see Subsection 4.1 for model description) on a complete list of datasets. The fertility of each model is shown in Table 2. The average fertility is 2.8. However, we decided to choose it with a margin so that the multilingual model with the highest fertility can be tested on the entire benchmark. As a result, we set the standard fertility to 3.

Finally, using the selected fertility value, we divided all datasets into subsets of various context lengths ranging from 4k to 128k tokens. The resulting dataset sizes and the average sample context lengths are given in Table 3.

3.4 Datasets

This section describes the datasets and data collection process in detail. We decided to create a combined benchmark that will include 1) transla-

Model Name	Fertility
GLM4-9B-Chat	2.15
Saiga-LLaMA-3-8B	2.40
LLaMA-3-8B	2.40
LLaMA-3-8B-Instruct	2.40
LLaMA-2-7B-32K	2.83
LongAlpaca-7B	2.83
LongChat-7B-v1.5-32k	2.83
Mistral-7B-v0.1	3.08
Mistral-7B-v0.3	3.08
Mistral-7B-Instruct-v0.3	3.08
ChatGLM2-6B-32k	3.50

Table 2: The table presents the average model's fertility. **Model Name** shows the name of a model. The **Fertility** shows the fertility.

tions of English datasets by using Google translator API[7], 2) adaptations to long input tasks in Russian and 3) entirely new datasets based on open data. We decided not to generate samples using LLMs and instead used annotators to mark up the samples. This helps reduce bias from using models like GPT-4, which are also part of the assessment. However, it does have some drawbacks, as full annotation can be costly and time-consuming in certain cases.

The exact dataset format can be found in Appendix B.

Passkey The Passkey is a synthetic QA dataset based on original passkey dataset from LongLLaMA's GitHub repository[8]. The main idea of the task is to extract a relevant piece of code number from a long text fragment that was created by repeating short sentence template containing noise. The model must find this code among the irrelevant information.

PasskeyWithLibrusec The PasskeyWithLibrusec is a more complicated version of Passkey QA dataset, in which we use randomly selected texts from the Librusec dataset as noise to make this dataset more difficult for LLMs.

ruGSM100 The ruGSM100 dataset is a translation of gsm100[9] one from L-Eval. It contains 100 math problems to be solved using Chain-of-Thought in a few-shot mode. This dataset aims to evaluate the model's reasoning and logical skills in maths. The context for all tasks is a prompt of 16 examples with problem descriptions and answers.

ru2WikiMultihopQA The ru2WikiMultihopQA was created by translating the dataset 2WikiMultihopQA[10] from LongBench, which consists of selected samples with a long context from the original multi-hop QA dataset 2WikiMultihopQA (Ho et al., 2020). This Wikipedia-based dataset tests reasoning skills by requiring a model to combine information from multiple texts to answer a question. The format of this dataset, which consists of up to 5-hop questions, makes it difficult for LLMs.

ruQasper The ruQasper was created by translating the Qasper[11] dataset from LongBench, which consists of selected samples with a long context from the original questions answering dataset over academic research papers called Qasper (Dasigi et al., 2021). The goal of the task is to find the answer to the question in one of the parts of the article. The context for samples is drawn from scientific articles to make the task more difficult.

ruTREC The ruTREC was created by translating the TREC[12] from LongBench. The dataset consists of selected samples with a long context from the original TREC (Li and Roth, 2002). This dataset is a type of few-shot in-context learning, in which the model is given several examples to understand the context, and then it has to answer which topic the question relates to.

ruQuALITY The ruQuALITY was created by translating QuALITY[13] from L-Eval, which consists of selected samples with a long context from the original multiple choice QA dataset called QuALITY (Pang et al., 2021). The model must find relevant information in the text and answer by choosing one of the four suggested options.

ruTPO The ruTPO was created by translating TPO[14] from L-Eval. The original dataset in the L-Eval benchmark consists of 15 samples, that are sourced from the TOEFL Practice Online and the dataset TOEFL-QA (Tseng et al., 2016). The TPO is a multiple-choice QA dataset, and, therefore, the model must find relevant information in the text and answer by choosing one of the four suggested options.

ruSciFi The ruSciFi was created by translating

[7] https://pypi.org/project/googletrans/

[8] https://github.com/CStanKonrad/long_llama/blob/main/examples/passkey.py

[9] https://huggingface.co/datasets/L4NLP/LEval/viewer/gsm100

[10] https://huggingface.co/datasets/THUDM/LongBench/viewer/2wikimqa_e

[11] https://huggingface.co/datasets/THUDM/LongBench/viewer/qasper_e

[12] https://huggingface.co/datasets/THUDM/LongBench/viewer/trec_e

[13] https://huggingface.co/datasets/L4NLP/LEval/viewer/quality

[14] https://huggingface.co/datasets/L4NLP/LEval/viewer/tpo

	Dataset Name	4k size / avg len	8k size / avg len	16k size/ avg len	32k size / avg len	64k size / avg len	128k size / avg len
I	Passkey	200 / 2790	200 / 5450	200 / 10996	200 / 21730	200 / 43391	200 / 87974
	PasskeyWithLibrusec	200 / 2705	200 / 5563	200 / 10835	200 / 22215	200 / 44682	200 / 88189
II	MatreshkaNames	150 / 3190	150 / 6314	150 / 12128	150 / 24168	150 / 48184	150 / 96135
	MatreshkaYesNo	299 / 3200	300 / 6317	300 / 12134	300 / 24173	300 / 48189	300 / 96142
	LibrusecHistory	-	32 / 4515	32 / 9003	32 / 17976	32 / 35924	-
	ruTREC	32 / 2870	50 / 6292	91 / 11886	122 / 22357	-	-
	ruSciFi	-	-	-	36 / 19397	28 / 40065	-
	ruSciAbstractRetrieval	210 / 3264	210 / 7260	210 / 15245	210 / 31231	200 / 63594	200 / 127777
	ruTPO	-	251 / 7651	-	-	-	-
	ruQuALITY	-	41 / 6380	161 / 12387	-	-	-
III	LongContextMultiQ	200 / 2940	200 / 6360	200 / 12240	200 / 26572	200 / 37482	200 / 68239
	LibrusecMHQA	-	384 / 4574	-	-	-	-
	ru2WikiMultihopQA	-	49 / 6378	128 / 11633	123 / 25523	-	-
	ruBABILongQA1	100 / 4002	100 / 8001	100 / 16002	100 / 32001	100 / 64002	100 / 128001
	ruBABILongQA2	100 / 4002	100 / 8001	100 / 16002	100 / 32001	100 / 64002	100 / 128001
	ruBABILongQA3	100 / 4011	100 / 8010	100 / 16011	100 / 32010	100 / 64011	100 / 128010
	ruBABILongQA4	100 / 4014	100 / 8013	100 / 16014	100 / 32013	100 / 64014	100 / 128013
	ruBABILongQA5	100 / 4006	100 / 8005	100 / 16006	100 / 32005	100 / 64006	100 / 128005
IV	ruSciPassageCount	100 / 3528	100 / 7128	100 / 13616	100 / 27160	100 / 53108	100 / 105949
	ruQasper	-	48 / 5768	134 / 11071	21 / 25185	-	-
	ruGSM100	-	-	100 / 9083	-	-	-

Table 3: Sizes and average sample lengths for the task subsets of various context lengths. **Dataset Name** shows the name of the dataset. The columns **4k**, **8k**, **16k**, **32k**, **64k**, **128k** show the number of samples and average sample lengths in tokens for the corresponding context length.

SciFi[15] from L-Eval, which consists of selected samples with a long context from the original SF-Gram[16] dataset, that contains thousands of science-fiction books, novels and movie information. The dataset aims to test the model's ability to follow contextual knowledge instead of parametric knowledge gained at the pretraining stage. The model needs to answer whether the information provided is true or false based on the information from the context and true or false based on the general world knowledge.

MatreshkaNames To create this dataset, we utilized two sets: Matreshka[17] and a Russian names[18] dataset. The Matreshka dataset comprises brief interactions involving "user" and "bot" roles, along with a brief description of the topic being discussed by each participant. To form longer contextual samples, we combined multiple interactions and replaced the names "user" and "bot" with the pull of names taken from the dataset of Russian names. Subsequently, we randomly selected a topic from the combined interactions and the name of the person discussing that topic. The dataset requires the model to identify the individual who discussed the selected topic.

MatreshkaYesNo The MatreshkaYesNo is based on the two datasets: Matreshka and Russian names, similar to the MatreshkaNames dataset. Instead of predicting names in the MatreshkaNames, the model is supposed to indicate whether this topic was mentioned in the dialog. The dataset is balanced across answers.

LongContextMultiQ The LongContextMultiQ is a multi-hop QA long context dataset for Russian that is based on data used for the MultiQ (Taktasheva et al., 2022)[19] dataset creation. The original MultiQ dataset is created by multi-hop dataset generation based on Wikidata[20] and Wikipedia, and consists of samples with different length. We selected 200 samples from these generated sources with a long context for each context length.

ruBABILong We adapted the methodology from (Kuratov et al., 2024) to create the Russian Benchmark for Artificial Intelligence for Long (ruBABILong)-context evaluation. It contains five long-context reasoning tasks for QA using facts hidden among distractor facts and irrelevant back-

[15] https://huggingface.co/datasets/L4NLP/LEval/viewer/sci_fi

[16] https://github.com/nschaetti/SFGram-dataset

[17] https://huggingface.co/datasets/zjkarina/matreshka

[18] https://www.kaggle.com/datasets/rai220/russian-cyrillic-names-and-sex/data

[19] https://huggingface.co/datasets/ai-forever/MERA/viewer/multiq

[20] https://www.wikidata.org/wiki/Wikidata:Introduction

ground text. The **ruBABILongQA1** task requires answering a question about a person's location using a single supporting fact. The **ruBABILongQA2** and **ruBABILongQA3** tasks introduce the challenge of differentiating subjects and objects, utilizing two and three supporting facts, respectively. The **ruBABILongQA4** task tackles spatial reasoning through two-argument relations, while the **ruBABILongQA5** task involves tracking multiple objects to solve the three-argument relation problem. Each task contains 100 samples, scaled to six sequence lengths from 4k to 128k. We obtained the task facts by translating the bAbI dataset (Weston et al., 2016), while the background texts were sampled using books from Librusec.

LibrusecHistory This dataset was created in question-answering (QA) format using Librusec[21]. Each sample in the LibrusecHistory dataset includes a text paragraph and a corresponding question. To create tasks with different input lengths, we initially selected large texts from various books in different domains and styles, divided them into fragments of several thousand tokens, and created the annotation (see Appendix A). These fragments and their respective questions and answers became the dataset's samples. Longer samples, with lengths up to 64,000 tokens, were created by supplementing these fragments with neighboring paragraphs from the original large text on both sides, resulting in longer inputs for the task.

LibrusecMHQA This dataset was created in multi-hop Question Answering (QA) format, also using Librusec as a LibrusecHistory. The main difference between these datasets is that in the LibrusecMHQA dataset, the necessary information for the answer is distributed in several parts of the context, making the task more difficult and allowing us to evaluate the model's reasoning skills better. The generation procedure for samples of different lengths remains the same.

ruSciAbstractRetrieval The ruSciAbstractRetrieval is a QA dataset ideologically similiar to the PassageRetrieval (Bai et al., 2023)[22] dataset from LongBench, that aims to evaluate model's reasoning skills. Each element of the dataset consists of a summary description of the topic and a set text paragraphs created from abstracts of scientific articles from ruSciBench[23]. The goal is to identify the paragraph where the specified topic is discussed. To create this dataset, we randomly choose some abstracts and generate descriptions of their topics using human annotators to acquire targets.

ruSciPassageCount The ruSciPassageCount dataset uses the basic idea of the original PassageCount[24] from LongBench. This QA dataset requires the model to use the full context to solve the problem. To generate the data, we randomly select abstracts from the ruSciBench dataset. We then choose a number of repeats and an ID for the paragraph to repeat. Next, we add the remaining non-repeated paragraphs to the repeated paragraph until we reach the desired context length. The resulting sequence of paragraphs is randomly shuffled. The ground truth for each sample is the number of unique paragraphs.

4 Evaluation Methodology

4.1 Baseline models

We evaluate 12 popular LLMs that feature long context capability, including GPT-4o[25], GLM4-9B-Chat (Zeng et al., 2022)[26], ChatGLM2-6B-32k (Zeng et al., 2022)[27], Saiga-LLaMA-3-8B[28], LLaMA-3-8B[29], LLaMA-3-8B-Instruct[30], LLaMA-2-7B-32K[31], LongAlpaca-7B[32], LongChat-7B-v1.5-32k, Mistral-7B-v0.1[33], Mistral-7B-v0.3[34], Mistral-7B-Instruct-v0.3[35]. A detailed information about the baseline models is given in Appendix C.

4.2 Experimental setup

Since the tasks themselves are long, in order not to go beyond the context window we fixed the evalua-

[21] https://huggingface.co/datasets/IlyaGusev/librusec

[22] https://huggingface.co/datasets/THUDM/LongBench/viewer/passage_retrieval_en

[23] https://huggingface.co/datasets/mlsa-iai-msu-lab/ru_sci_bench

[24] https://huggingface.co/datasets/THUDM/LongBench/viewer/passage_count

[25] Due to resource constraints, we evaluated GPT-4o on only 10% of each dataset of our benchmark, including each context length. Therefore, the results may not be precise.

[26] https://huggingface.co/THUDM/glm-4-9b-chat

[27] https://huggingface.co/THUDM/chatglm2-6b-32k

[28] https://huggingface.co/IlyaGusev/saiga_llama3_8b

[29] https://huggingface.co/meta-llama/Meta-Llama-3-8B

[30] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[31] https://huggingface.co/togethercomputer/LLaMA-2-7B-32K

[32] https://huggingface.co/Yukang/LongAlpaca-7B

[33] https://huggingface.co/mistralai/Mistral-7B-v0.1

[34] https://huggingface.co/mistralai/Mistral-7B-v0.3

[35] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

tion of tasks in zero-shot, except for tasks ruTREC and ruGSM100 in which the few-shot examples provided as a part of long context input. When the input length of the sample surpasses the maximum model context length, we truncate the input sequence from the right. The baselines were evaluated with greedy decoding (temperature= 1.0, num_beams = 1, do_sample = False) for reproducibility.

For each task, we fixed a natural language prompt unified for all the models (see Appendix B for the exact formulation). The prompts were estimated from an empirical analysis of the tasks through a series of experiments. However, it should be noted that further study of this subject is still required.

We run all the experiments on a double NVIDIA A100 GPU.

5 Results

The baseline results with respect to context length are shown in Table 4 and with respect to tasks are shown in Tables 5, 6, 7. Detailed results for each model are given in Appendix D. Based on the obtained results we can draw the following conclusions for each group of tasks.

Group I The tasks from this group are relatively simple, and almost all models pass them well within their maximum input length. The only exception is the LongAlpaca-7B model.

Group II MatreshkaYesNo, turns out to be the most straightforward task in the group, which all models cope with naturally. The ruTPO and ruQuALITY tasks are of medium complexity; several models achieved good scores in them.

The classic QA task LibrusecHistory is effectively handled by modern models; however, the quality decreases with the input length increase (e.g. for ruSciAbstractRetrieval). Nevertheless, in some cases, a larger context is advantageous, as seen in ruTREC, where increasing the input length helps the model handle the task better because this task is designed in a few-shot format.

The most complex tasks in this group can be considered MatreshkaNames and ruSciFi. For the first, several models (e.g., ChatGLM2-6B-32k, LLaMA-2-7B-32K, and LongAlpaca-7B) show low results for any input length. ruSciFi with a 64K context is beyond the capabilities of most models. At the same time, the strongest models (GPT-4o and GLM4-9B-Chat) not only show promising results but also improve the score with the length increase.

Group III For tasks from ruBABILong, an increase in context leads to worse results. ruBABILongQA2 and ruBABILongQA3 turn out to be significantly more complex than others, which coincides with results from (Kuratov et al., 2024). The length of the context plays a significant role; with its growth, the quality immediately begins to decline for all but the strongest models.

LibrusecMHQA turns out to be a complex dataset; the maximum quality of the models for solving this problem is only 50 for 8k tokens.

Group IV ruSciPassageCount is the most difficult task created from scratch. All models except GPT-4o handle it poorly, even with a 4K input length; the result's sensitivity to the context's size is high. Besides, all open models fail to cope with ruQasper for complex tasks and domains. A similar result is obtained when measuring the quality of solutions to mathematical problems from ruGSM100. Our conclusions are similar to those obtained in (An et al., 2023); the only exception is the LLaMA-2 family of models, which performs worse in our experiments, most likely due to translating tasks into the less familiar Russian language.

Overall, SFT models perform better than the pretrain once. In most cases, an increase in the input length negatively affects the capabilities of all models. The results indicate that our prior division of tasks into groups is highly correlated with their complexity.

6 Conclusion

The rapid development of LLMs has posed new challenges for evaluating their ability to process long texts. To address this problem, we have introduced LIBRA (Long Input Benchmark for Russian Analysis). This benchmark evaluates LLM long context understanding abilities through 21 long-context textual tasks. The tasks enable model evaluation across various context lengths ranging from 4k to 128k tokens based on the analysis of dataset context lengths of the models' tokenizers. Our contribution encompasses a benchmark methodology with open-sourced datasets of different lengths and domains, a codebase for model evaluation, and baseline solution scoring. The datasets are published under the MIT license, and the leaderboard is publicly accessible on HuggingFace[36].

[36]The link has been removed to maintain anonymity during the review period.

Model Name	4k	8k	16k	32k	64k	128k	Overall
GPT-4o	**73.3**	**73.1**	**73.5**	**62.0**	**65.3**	**54.8**	**70.2**
GLM4-9B-Chat	61.5	59.8	53.4	50.6	48.7	43.8	52.3
Mistral-7B-Instruct-v0.3	48.3	44.7	37.3	32.3	-	-	29.9
Mistral-7B-v0.3	46.6	42.9	37.9	32.8	-	-	27.4
LLaMA-2-7B-32K	45.2	43.7	36.6	33.0	-	-	27.1
LongChat-7B-v1.5-32k	38.7	36.0	30.4	24.5	-	-	22.1
ChatGLM2-6B-32k	28.6	24.9	22.5	14.5	-	-	15.7
LongAlpaca	26.0	22.3	18.8	13.8	-	-	13.7
LLaMA-3-8B-Instruct	58.1	56.9	-	-	-	-	21.9
Saiga-LLaMA-3-8B	58.7	55.0	-	-	-	-	21.0
LLaMA-3-8B	54.6	49.4	-	-	-	-	18.4
Mistral-7B-v0.1	47.2	42.8	-	-	-	-	17.3

Table 4: The table presents the model evaluation scores for different context lengths. **Model Name** shows the name of the model. The columns **4k**, **8k**, **16k**, **32k**, **64k**, **128k** present evaluation scores averaged over all tasks. The **Overall** score is obtained by averaging the results over all lengths. The best score is put in bold, the second best is underlined.

Model Name	Passkey	MatreshkaYesNo	MatreshkaNames	PasskeyWithLibrusec	LibrusecHistory	ruGSM100	ruSciPassageCount	ru2WikiMultihopQA
GPT-4o	**100.0**	**80.0**	**51.7**	**100.0**	**97.5**	**100.0**	**35.0**	**76.7**
GLM4-9B-Chat	100.0	68.0	47.3	100.0	82.0	8.0	7.5	48.8
Mistral-7B-Instruct-v0.3	66.7	35.3	16.3	66.6	50.8	11.0	8.2	43.2
Mistral-7B-v0.3	66.7	32.0	10.0	66.7	68.0	9.0	0.0	41.0
LLaMA-2-7B-32K	66.7	33.4	3.4	65.5	40.6	7.0	4.7	37.2
LongChat-7B-v1.5-32k	66.5	33.4	5.9	66.0	26.6	5.0	4.8	35.2
ChatGLM2-6B-32k	63.7	33.4	1.3	65.0	8.6	5.0	3.7	17.5
LongAlpaca	42.4	30.5	0.4	40.6	13.3	2.0	3.8	30.3
LLaMA-3-8B-Instruct	33.3	27.3	16.6	33.3	22.7	0.0	6.5	17.7
Saiga-LLaMA-3-8B	33.3	28.0	15.6	33.2	24.2	0.0	3.8	17.7
LLaMA-3-8B	33.3	20.2	10.0	33.3	22.7	0.0	3.3	18.4
Mistral-7B-v0.1	35.0	16.8	8.1	38.3	23.4	13.0	1.3	23.0

Table 5: The table presents the evaluation results. **Model Name** shows the name of the model. The score for each task is averaged by the context length. The best score is put in bold, the second best is underlined.

Model Name	LongContextMultiQ	ruSciAbstractRetrieval	ruTREC	ruSciFi	LibrusecMHQA	ruBABILongQA1	ruBABILongQA2	ruBABILongQA3
GPT-4o	**36.7**	76.9	**75.0**	**75.0**	**50.0**	**78.3**	**36.7**	21.4
GLM4-9B-Chat	7.8	**77.8**	69.9	40.9	44.5	54.1	29.8	**22.3**
Mistral-7B-Instruct-v0.3	4.8	43.6	42.5	15.3	33.6	14.3	2.8	6.0
Mistral-7B-v0.3	5.2	30.5	5.4	0.0	39.1	37.3	16.7	15.7
LLaMA-2-7B-32K	7.9	39.1	23.8	5.6	27.6	40.3	16.6	16.3
LongChat-7B-v1.5-32k	3.2	41.1	7.4	2.8	24.7	17.5	7.2	4.0
ChatGLM2-6B-32k	1.2	13.6	4.5	0.0	6.8	12.2	1.5	2.5
LongAlpaca	0.8	23.5	0.5	1.4	7.8	3.8	0.3	3.5
LLaMA-3-8B-Instruct	4.9	31.4	27.4	0.0	46.1	23.7	4.1	4.5
Saiga-LLaMA-3-8B	4.8	31.7	26.3	0.0	45.1	25.4	4.4	6.1
LLaMA-3-8B	7.0	30.9	19.0	0.0	41.4	20.8	7.7	9.1
Mistral-7B-v0.1	4.4	28.5	4.0	1.4	34.1	21.0	7.7	9.0

Table 6: The table presents the evaluation results. **Model Name** shows the name of the model. The score for each task is averaged by the context length. The best score is bold, the second best is underlined.

Model Name	ruBABILongQA4	ruBABILongQA5	ruQuALITY	ruTPO	ruQasper	Overall
GPT-4o	**79.0**	**90.0**	**83.3**	**100.0**	**31.7**	**70.2**
GLM4-9B-Chat	52.8	70.3	74.1	86.9	5.0	52.3
Mistral-7B-Instruct-v0.3	27.6	37.6	30.6	66.4	5.4	29.9
Mistral-7B-v0.3	23.6	47.1	15.2	39.7	5.8	27.4
LLaMA-2-7B-32K	16.7	43.0	15.5	54.3	4.7	27.1
LongChat-7B-v1.5-32k	12.7	33.3	23.1	39.6	5.0	22.1
ChatGLM2-6B-32k	0.6	8.8	49.2	29.0	2.6	15.7
LongAlpaca	0.2	29.4	44.0	6.8	2.0	13.7
LLaMA-3-8B-Instruct	19.6	25.3	34.6	78.1	2.2	21.9
Saiga-LLaMA-3-8B	20.3	25.2	17.9	75.7	2.5	21.0
LLaMA-3-8B	19.1	22.6	8.5	58.2	2.2	18.5
Mistral-7B-v0.1	12.4	23.2	17.3	39.6	2.5	17.3

Table 7: The table presents the evaluation results. **Model Name** shows the name of the model. The score for each task is averaged by the context length. The **Overall** score is obtained by averaging the results over each task. The best score is put in bold, the second best is underlined.

Limitations

Although the LIBRA was created to solve the absence of the long context benchmark for Russian and provides significant advancements in evaluating language models with long contexts, it still has a number of limitations that need to be acknowledged.

Data Representation. The texts included in the benchmark are gathered from specific domains, which might not cover the full range of Russian language usage. This can raise concerns about data privacy, representation, and potential biases within the benchmark. It is important to consider that dialects, regional variations, and sociolects may not be adequately represented, potentially leading to biased performance metrics. As a result, models may excel in benchmark tasks but struggle with texts outside these domains, limiting their generalization ability. The corpus used for the benchmark may become outdated over time. New words, phrases, and usage patterns could emerge, making the benchmark less relevant for future model evaluations.

Methodology limitations. When creating the datasets, we hypothesized that synthetically augmentation of the context length of the datasets, such as LibrusecHistory, would not affect the results. Our experiments show that these tasks are pretty challenging for many models. We made this methodological assumption due to the limitations of human data annotation; it is difficult for people to read large texts and concentrate enough to create questions and search for information within them. This data creation method may result in task errors, particularly when a newly extended text fragment contains conflicting information that could impact the answer. However, we found this approach acceptable due to the increased speed and cost-effectiveness.

The current methodology also restricts the number of tasks, and many of them are translated only due to the high cost of data creation.

Length context. The benchmark focuses on evaluating long contexts, but the definition of "long context" can differ based on the application and the model. The chosen context lengths may not be ideal for all usage scenarios, and models could exhibit varying performance. In this paper, we have measured the average fertility of baseline model tokenizers on a full list of datasets from our benchmark to sample different contexts and analyzed the models' results on our datasets across various context lengths. LMs with more parameters may inherently perform better, but this does not necessarily reflect improvements in long context understanding.

Data leakage is a critical concern for modern benchmarks because current models are trained on a significant amount of text from the Internet. Long context benchmarks are particularly risky, as their texts are based on web sources and books. This could potentially lead to data leakage and inaccurate evaluation. However, creating original long texts from scratch not found on the web is exceptionally costly. As a result, we use open sources to develop our benchmark, acknowledging the potential risks. Nevertheless, we firmly believe this will make a valuable contribution to the Russian community, as no long context datasets are currently available.

Ethical Considerations. The data used in the benchmark was created from open data sources. When annotating the data, we obtained transparent permission from all users and made efforts to maintain the confidentiality and anonymity of participants. As the benchmark develops, ongoing efforts are required to identify and minimize biases in the benchmark datasets and evaluation metrics. The benchmark does not currently contain the datasets covering the ethical or AI safety skill evaluation, but this is a space for future work.

References

Chenxin An, Shansan Gong, Ming Zhong, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2023. L-eval: Instituting standardized evaluation for long context language models. *arXiv preprint arXiv:2307.11088*.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Amanda Bertsch, Uri Alon, Graham Neubig, and Matthew Gormley. 2024. Unlimiformer: Long-range transformers with unlimited length input. *Advances in Neural Information Processing Systems*, 36.

Rishi Bommasani, Percy Liang, and Tony Lee. 2023. Holistic Evaluation of Language Models. *Annals* *of the New York Academy of Sciences*, 1525(1):140–146.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. *arXiv preprint arXiv:2105.03011*.

Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. 2023. Longnet: Scaling transformers to 1,000,000,000 tokens. *arXiv preprint arXiv:2307.02486*.

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

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. 2024. In search of needles in a 10m haystack: Recurrent memory finds what llms miss. *arXiv preprint arXiv:2402.10790*.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, et al. 2021. Quality: Question answering with long input texts, yes! *arXiv preprint arXiv:2112.08608*.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*.

Ofir Press, Noah A Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. Zeroscrolls: A zero-shot benchmark for long text understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7977–7989.

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)*. Association for Computational Linguistics.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. 2023. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*.

Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. 2022. A length-extrapolatable transformer. *arXiv preprint arXiv:2212.10554*.

Ekaterina Taktasheva, Tatiana Shavrina, Alena Fenogenova, Denis Shevelev, Nadezhda Katricheva, Maria Tikhonova, Albina Akhmetgareeva, Oleg Zinkevich, Anastasiia Bashmakova, Svetlana Iordanskaia, et al. 2022. Tape: Assessing few-shot russian language understanding. *arXiv preprint arXiv:2210.12813*.

Bo-Hsiang Tseng, Sheng-Syun Shen, Hung-Yi Lee, and Lin-Shan Lee. 2016. Towards machine comprehension of spoken content: Initial toefl listening comprehension test by machine. In *INTERSPEECH*.

Szymon Tworkowski, Konrad Staniszewski, Mikołaj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Miłoś. 2024. Focused transformer: Contrastive training for context scaling. *Advances in Neural Information Processing Systems*, 36.

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.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomás Mikolov. 2016. Towards ai-complete question answering: A set of prerequisite toy tasks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *37th Conference on Neural Information Processing Systems (NeurIPS 2023) Datasets and Benchmarks Track*.

Appendix

A Data Annotation Details

The datasets LibrusecHistory, LibrusecMHQA, and ruSciAbstractRetrieval were created via the crowd-sourced platform.

In the LibrusecHistory, annotators were instructed to read a lengthy text and generate four questions based on the text and answer them. Guidelines were provided regarding the type of questions to ask: 1) Questions should be answerable using information present in the text 2) The questions must not be about widely known information but should be related to the text 3) Questions can cover various aspects such as character actions, appearance, thoughts, events, and scene descriptions 4) Logical deductions are not required to answer the questions 5) Each question should have a single, clear, unambiguous answer from the text.

The design of the dataset LibrusecMHQA project follows a similar structure to LibrusecHistory, but the question criteria were more complex. In this dataset, the questions were answered by expert editors rather than through crowd-sourcing. The main distinction in the criteria for annotators is the multi-hop questions, where simply reading the sentence containing the answer is insufficient. Instead, reading at least a paragraph of 2-5 sentences, or the entire relevant fragment, is necessary to gather information and generate a complete answer.

The ruSciAbstractRetrieval was collected by crowd-sourced annotators. These annotators were asked to read a long text annotation and briefly describe the contents. The criteria for the description were as follows: 1) The description must start with the word "Describes". 2) It must be a single sentence, which can be complex. 3) The description should not exceed 30 words, including conjunctions, particles, and prepositions. 4) It should include the main general ideas identified in the abstract but should not include details.

Training examples were available for all projects. The contributions of human annotators are amassed and stored in a manner that ensures anonymity. The average hourly compensation exceeds the minimum wage per hour in Russia. Each annotator is informed about topics that may be sensitive in the data, such as politics, societal minorities, and religion. Table 8 summarizes general details concerning the creation of the datasets via crowd-source on ABC[37] data labeling platform.

B Dataset Examples

This section provides examples of the task format for the benchmark datasets. The exact prompts for the benchmark are not fixed. Here we provide prompts used in our experiments[38].

Passkey: *You are provided with a long text that contains the access key. Just remember the access key.*
Context: {*context*}
You only need to specify the access key in the response.
Question: {*input*}
Answer:

PasskeyWithLibrusec: *You are provided with a long text that contains the access key. Just remember the access key.*
Context: {*context*}
You only need to specify the access key in the response.
Question: {*input*}
Answer:

MatreshkaNames: *You are provided with several dialogues. Remember the names of the people and the topics they talked about.*
Context: {*context*}

[37] https://elementary.activebc.ru

[38] All examples are presented in English for transparency and are given and are for illustrative purposes only to clarify the idea of a given task. The examples are not necessarily a direct translation of specific examples from the dataset. The exact prompts in their original formulation in Russian can be found in our repository [The link has been removed to preserve anonymity during the review period].

Task Name	Total	Pay Rate	Example Number	Overlap
LibrusecHistory	84$	6.25$/hr	32	1
LibrusecMHQA	458$	6.25$/hr	40	3
ruSciAbstractRetrieval	290$	6.25$/hr	100	3

Table 8: The details of datasets collection. **Total** is the budget spent to annotate the tasks employed for metric evaluation. **Pay Rate** is the hourly rate computed as a simple average of pay rates based on time spent annotating one row and the reward for this row. **Example Number** refers to the total number of samples processed while collecting or verifying the dataset. **Overlap** is the median number of votes per dataset sample averaged across all annotation tasks for the same dataset (if more than 1 task is provided).

In the answer, specify only the name of the interlocutor who spoke on the topic from the next question.
Question: {*input*}
Answer:

MatreshkaYesNo: *You are provided with several dialogues. Remember the names of the topics that the interlocutors talked about.*
Context: {*context*}
In the answer, you only need to specify 'Yes' if there was such a topic and 'No' if there was no such topic in the dialogues.
Question: {*input*}
Answer:

LibrusecHistory: *You are given a long text in which you need to find the answer to the question.*
Context: {*context*}
Find the answer in the text to the following question.
Question: {*input*}
Answer:

ruTREC: *Define the type of question below. Here are some examples:*
Context: {*context*}
Define the type of question below.
Question: {*input*}
Answer:

ruSciFi: *You are given a long text in which you need to find the answer to the question.*
Context: {*context*}
You need to answer the following question with one of the options: 'False [in the real world: False]', 'True [in the real world: False]', 'True [in the real world: True]' or 'False [in the real world: True]'.
Question: {*input*}
Answer:

ruSciAbstractRetrieval: *Below are a few paragraphs. Determine which paragraph the short description corresponds to.*
Context: {*context*}
Determine which paragraph the short description corresponds to. The response must contain the paragraph number.
Question: {*input*}
Answer:

ruTPO: *You are given a long text in which you need to find the answer to the question.*
Context: {*context*}
You will be given several answers to the question in the text; choose only one correct one and specify the letter A, B, C, or D.
Question: {*input*}
Answer:

ruQuALITY: *You are given a long text in which you need to find the answer to the question.*
Context: {*context*}
You will be given several answers to the question in the text; choose only one correct one.
Question: {*input*}
Answer:

LongContextMultiQ: *You are given a long text where you need to find the answer to the question.*
Context: {*context*}
Find the answer in the text to the following question.
Question: {*input*}
Answer:

LibrusecMHQA: *You are given a long text where you need to find the answer.*

Context: {*context*}
Find the answer in the text to the following question.
Question: {*input*}
Answer:

ru2WikiMultihopQA: *The answer to the question is based on the above excerpts.*
Context: {*context*}
Answer the question briefly, based on the above excerpts.
Question: {*input*}
Answer:

ruBABILongQA1: *I'm giving you a context with facts about the location of different people. You need to answer the question based only on information obtained from the facts. If the person was in different places, use the last location to answer the question.*
Context: {*context*}
Answer the question as briefly as possible.
Question: {*input*}
Answer:

ruBABILongQA2: *I'm giving you a context with facts about the location and actions of different people. You need to answer the question based only on factual information. If a person took an item in one place and went to another, that item is also in the second place. If a person leaves an item in the first place and moves to the second place, the item remains in the first place.*
Context: {*context*}
Answer the question as briefly as possible.
Question: {*input*}
Answer:

ruBABILongQA3: *I'm giving you a context with facts about the location and actions of different people. You need to answer the question based only on factual information. If a person took an item in one place and went to another, that item is also in the second place. If a person leaves an item in the first mets and moves to the second place, the item remains in the first place.*
Context: {*context*}
Answer the question as briefly as possible.
Question: {*input*}
Answer:

ruBABILongQA4: *I'm giving you a context with facts about the location and actions of different people. You need to answer the question based only on factual information.*
Context: {*context*}
Answer the question as briefly as possible.
Question: {*input*}
Answer:

ruBABILongQA5: *I'm giving you a context with facts about the location and actions of different people. You need to answer the question based only on factual information.*
Context: {*context*}
Answer the question as briefly as possible.
Question: {*input*}
Answer:

ruSciPassageCount: *Below are a few paragraphs. Read them and determine the number of unique paragraphs.*
Context: {*context*}
Determine the number of unique paragraphs. The answer must contain only one number.
Question: {*input*}
Answer:

ruQasper: *You are provided with a scientific article and a question.*
Context: {*context*}
Answer the question as briefly as possible, using a single phrase or sentence if possible. Don't give any explanations.
Question: {*input*}
Answer:

ruGSM100: *Examples of mathematical problems are given below. Think step by step and answer the question.*
Context: {*context*}
Think step by step and answer the question.
Question: {*input*}
Answer:

C Detailed Model Information

The baseline model specifics are presented in Table 9.

D Detailed Model Results

This section presents the detailed results of model evaluation. The results are shown for the following models: GPT-4o (Table 10), GLM4-9B-Chat

Model Name	**Type**	**Parameters**	**Max Context Length**
GPT-4o	Commercial	-	128k
GLM4-9B-Chat	Open-source	9B	128k
Mistral-7B-Instruct-v0.3	Open-source	7B	32k
Mistral-7B-v0.3	Open-source	7B	32k
LLaMA-2-7B-32K	Open-source	7B	32k
LongChat-7B-v1.5-32k	Open-source	7B	32k
ChatGLM2-6B-32k	Open-source	6B	32k
LongAlpaca-7B	Open-source	7B	32k
LLaMA-3-8B-Instruct	Open-source	8B	8k
Saiga-LLaMA-3-8B	Open-source	8B	8k
LLaMA-3-8B	Open-source	8B	8k
Mistral-7B-v0.1	Open-source	7B	8k

Table 9: The models evaluated as baselines. **Model Name** shows the name of the model. The **Max Context Length** shows maximal context lengths.

(Table 11), Mistral-7B-Instruct-v0.3 (Table 12), Mistral-7B-v0.3 (Table 13), LLaMA-2-7B-32K (Table 14), LongChat-7B-v1.5-32k (Table 15), ChatGLM2-6B-32K (Table 16), LongAlpaca (Table 17), LLaMA-3-8B-Instruct (Table 18), Saiga-LLaMA-3-8B (Table 19), LLaMA-3-8B (Table 20) and Mistral-7B-v0.1 (Table 21).

	Dataset Name	4k	8k	16k	32k	64k	128k	Overall
I	Passkey	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	PasskeyWithLibrusec	100.0	100.0	100.0	100.0	100.0	100.0	100.0
II	MatreshkaNames	60.0	60.0	50.0	40.0	50.0	50.0	51.7
	MatreshkaYesNo	80.0	60.0	100.0	80.0	70.0	90.0	80.0
	LibrusecHistory	-	100.0	100.0	100.0	90.0	-	97.5
	ruTREC	60.0	80.0	90.0	70.0	-	-	75.0
	ruSciFi	-	-	-	60.0	90.0	-	75.0
	ruSciAbstractRetrieval	99.0	95.4	92.5	95.6	59.1	19.8	76.9
	ruTPO	-	100.0	-	-	-	-	100.0
	ruQuALITY	-	80.0	86.7	-	-	-	83.3
III	LongContextMultiQ	30.0	100.0	70.0	0.0	10.0	10.0	36.7
	LibrusecMHQA	-	50.0	-	-	-	-	50.0
	ru2WikiMultihopQA	-	80.0	80.0	70.0	-	-	76.7
	ruBABILongQA1	90.0	80.0	70.0	90.0	80.0	60.0	78.3
	ruBABILongQA2	40.0	30.0	40.0	40.0	50.0	20.0	36.7
	ruBABILongQA3	20.0	30.0	10.0	20.0	20.0	28.7	21.4
	ruBABILongQA4	88.0	80.0	80.0	57.1	88.6	80.0	79.0
	ruBABILongQA5	86.7	86.7	93.3	96.7	86.7	90.0	90.0
IV	ruSciPassageCount	100.0	50.0	30.0	0.0	20.0	10.0	35.0
	ruQasper	-	28.7	31.8	34.7	-	-	31.7
	ruGSM100	-	-	100.0	-	-	-	100.0

Table 10: The table presents the evaluation results of GPT-4o. **Dataset Name** shows the name of the dataset. The rows **4k**, **8k**, **16k**, **32k**, **64k**, **128k** show evaluation scores of datasets for each context length, respectively. The **Overall** score is obtained by averaging the results over each length.

	Dataset Name	4k	8k	16k	32k	64k	128k	Overall
I	Passkey	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	PasskeyWithLibrusec	100.0	100.0	100.0	100.0	100.0	100.0	100.0
II	MatreshkaNames	64.7	50.7	52.0	47.3	37.3	32.0	47.3
	MatreshkaYesNo	79.3	75.0	71.3	67.0	59.7	56.0	68.0
	LibrusecHistory	-	84.4	84.4	84.4	75.0	-	82.0
	ruTREC	56.8	70.0	75.8	77.0	-	-	69.9
	ruSciFi	-	-	-	38.9	42.9	-	40.9
	ruSciAbstractRetrieval	98.2	92.3	91.2	81.9	64.1	39.1	77.8
	ruTPO	-	86.9	-	-	-	-	86.9
	ruQuALITY	-	82.9	65.2	-	-	-	74.1
III	LongContextMultiQ	5.5	26.5	3.5	0.5	0.5	10.0	7.8
	LibrusecMHQA	-	44.5	-	-	-	-	44.5
	ru2WikiMultihopQA	-	55.1	55.5	35.8	-	-	48.8
	ruBABILongQA1	69.9	59.0	60.0	50.8	42.9	42.0	54.1
	ruBABILongQA2	38.9	33.0	29.9	26.9	26.8	23.5	29.8
	ruBABILongQA3	24.6	27.9	21.4	22.6	18.7	18.5	22.3
	ruBABILongQA4	62.1	59.6	56.6	58.0	43.0	37.7	52.8
	ruBABILongQA5	73.0	73.5	72.0	66.8	69.7	67.0	70.3
IV	ruSciPassageCount	27.0	8.0	9.0	0.0	1.0	0.0	7.5
	ruQasper	-	6.5	5.9	2.6	-	-	5.0
	ruGSM100	-	-	8.0	-	-	-	8.0

Table 11: The table presents the evaluation results of GLM4-9B. **Dataset Name** shows the name of the dataset. The rows **4k**, **8k**, **16k**, **32k**, **64k**, **128k** show evaluation scores of datasets for each context length, respectively. The **Overall** score is obtained by averaging the results over each length.

	Dataset Name	4k	8k	16k	32k	64k	128k	Overall
I	Passkey	100.0	100.0	100.0	100.0	0.0	0.0	66.7
	PasskeyWithLibrusec	100.0	100.0	100.0	99.5	0.0	0.0	66.6
II	MatreshkaNames	38.0	32.0	16.7	11.3	0.0	0.0	16.3
	MatreshkaYesNo	56.5	50.7	54.7	50.0	0.0	0.0	35.3
	LibrusecHistory	-	71.9	62.5	68.8	0.0	-	50.8
	ruTREC	56.8	38.0	40.7	34.4	-	-	42.5
	ruSciFi	-	-	-	30.6	0.0	-	15.3
	ruSciAbstractRetrieval	98.2	86.9	71.1	5.1	0.0	0.0	43.6
	ruTPO	-	66.4	-	-	-	-	66.4
	ruQuALITY	-	38.2	23.0	-	-	-	30.6
III	LongContextMultiQ	3.5	22.0	3.5	0.0	0.0	0.0	4.8
	LibrusecMHQA	-	33.6	-	-	-	-	33.6
	ru2WikiMultihopQA	-	55.1	46.9	27.6	-	-	43.2
	ruBABILongQA1	25.0	15.0	22.0	24.0	0.0	0.0	14.3
	ruBABILongQA2	8.0	5.0	2.0	2.0	0.0	0.0	2.8
	ruBABILongQA3	10.0	8.0	10.0	8.0	0.0	0.0	6.0
	ruBABILongQA4	51.8	44.3	39.3	30.3	0.0	0.0	27.6
	ruBABILongQA5	54.7	62.0	55.3	53.3	0.0	0.0	37.6
IV	ruSciPassageCount	26.0	14.0	7.0	2.0	0.0	0.0	8.2
	ruQasper	-	6.6	6.6	2.9	-	-	5.4
	ruGSM100	-	-	11.0	-	-	-	11.0

Table 12: The table presents the evaluation results of Mistral-7B-v0.3-Instruct. **Dataset Name** shows the name of the dataset. The rows **4k**, **8k**, **16k**, **32k**, **64k**, **128k** show evaluation scores of datasets for each context length, respectively. The **Overall** score is obtained by averaging the results over each length.

	Dataset Name	4k	8k	16k	32k	64k	128k	Overall
I	Passkey	100.0	100.0	100.0	100.0	0.0	0.0	66.7
	PasskeyWithLibrusec	100.0	100.0	100.0	100.0	0.0	0.0	66.7
II	MatreshkaNames	28.7	16.0	10.7	4.7	0.0	0.0	10.0
	MatreshkaYesNo	44.8	47.0	50.0	50.0	0.0	0.0	32.0
	LibrusecHistory	-	93.8	93.8	84.4	0.0	-	68.0
	ruTREC	0.0	8.0	4.4	9.0	-	-	5.4
	ruSciFi	-	-	-	0.0	0.0	-	0.0
	ruSciAbstractRetrieval	87.4	56.6	36.9	1.9	0.0	0.0	30.5
	ruTPO	-	39.7	-	-	-	-	39.7
	ruQuALITY	-	23.6	6.8	-	-	-	15.2
III	LongContextMultiQ	4.0	24.0	3.5	0.0	0.0	0.0	5.2
	LibrusecMHQA	-	39.1	-	-	-	-	39.1
	ru2WikiMultihopQA	-	46.9	49.2	26.8	-	-	41.0
	ruBABILongQA1	60.0	63.0	58.0	43.0	0.0	0.0	37.3
	ruBABILongQA2	35.0	23.0	18.0	24.0	0.0	0.0	16.7
	ruBABILongQA3	29.0	23.0	23.0	19.0	0.0	0.0	15.7
	ruBABILongQA4	46.3	34.4	36.2	24.9	0.0	0.0	23.6
	ruBABILongQA5	70.3	68.7	75.3	68.3	0.0	0.0	47.1
IV	ruSciPassageCount	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	ruQasper	-	8.9	6.5	1.9	-	-	5.8
	ruGSM100	-	-	9.0	-	-	-	9.0

Table 13: The table presents the evaluation results of Mistral-7B-v0.3. **Dataset Name** shows the name of the dataset. The rows **4k**, **8k**, **16k**, **32k**, **64k**, **128k** show evaluation scores of datasets for each context length, respectively. The **Overall** score is obtained by averaging the results over each length.

	Dataset Name	4k	8k	16k	32k	64k	128k	Overall
I	Passkey	100.0	100.0	100.0	100.0	0.0	0.0	66.7
	PasskeyWithLibrusec	100.0	97.5	98.5	97.0	0.0	0.0	65.5
II	MatreshkaNames	8.0	6.7	2.0	4.0	0.0	0.0	3.4
	MatreshkaYesNo	50.2	50.0	50.0	50.0	0.0	0.0	33.4
	LibrusecHistory	-	68.8	50.0	43.8	0.0	-	40.6
	ruTREC	24.3	18.0	24.2	28.7	-	-	23.8
	ruSciFi	-	-	-	11.1	0.0	-	5.6
	ruSciAbstractRetrieval	85.2	76.1	46.8	26.7	0.0	0.0	39.1
	ruTPO	-	54.3	-	-	-	-	54.3
	ruQuALITY	-	17.1	13.9	-	-	-	15.5
III	LongContextMultiQ	4.5	33.0	10.0	0.0	0.0	0.0	7.9
	LibrusecMHQA	-	27.6	-	-	-	-	27.6
	ru2WikiMultihopQA	-	44.9	39.8	26.8	-	-	37.2
	ruBABILongQA1	60.0	66.0	66.0	50.0	0.0	0.0	40.3
	ruBABILongQA2	25.0	30.0	25.9	19.0	0.0	0.0	16.6
	ruBABILongQA3	22.9	28.9	26.0	20.0	0.0	0.0	16.3
	ruBABILongQA4	31.0	34.0	23.0	12.0	0.0	0.0	16.7
	ruBABILongQA5	59.0	66.0	64.0	69.0	0.0	0.0	43.0
IV	ruSciPassageCount	18.0	5.0	5.0	0.5	0.0	0.0	4.7
	ruQasper	-	5.8	6.0	2.2	-	-	4.7
	ruGSM100	-	-	7.0	-	-	-	7.0

Table 14: The table presents the evaluation results of LLaMA-2-32K. **Dataset Name** shows the name of the dataset. The rows **4k**, **8k**, **16k**, **32k**, **64k**, **128k** show evaluation scores of datasets for each context length, respectively. The **Overall** score is obtained by averaging the results over each length.

	Dataset Name	4k	8k	16k	32k	64k	128k	Overall
I	Passkey	99.5	100.0	100.0	99.5	0.0	0.0	66.5
	PasskeyWithLibrusec	100.0	100.0	98.5	97.5	0.0	0.0	66.0
II	MatreshkaNames	17.3	6.7	8.0	3.3	0.0	0.0	5.9
	MatreshkaYesNo	50.2	50.0	50.0	50.0	0.0	0.0	33.4
	LibrusecHistory	-	56.2	34.4	15.6	0.0	-	26.6
	ruTREC	5.4	10.0	7.7	6.6	-	-	7.4
	ruSciFi	-	-	-	5.6	0.0	-	2.8
	ruSciAbstractRetrieval	87.4	76.2	60.6	22.2	0.0	0.0	41.1
	ruTPO	-	39.6	-	-	-	-	39.6
	ruQuALITY	-	28.5	17.8	-	-	-	23.1
III	LongContextMultiQ	2.5	14.0	2.5	0.0	0.0	0.0	3.2
	LibrusecMHQA	-	24.7	-	-	-	-	24.7
	ru2WikiMultihopQA	-	42.9	39.8	22.8	-	-	35.2
	ruBABILongQA1	26.0	29.0	31.0	19.0	0.0	0.0	17.5
	ruBABILongQA2	11.0	8.0	16.0	8.0	0.0	0.0	7.2
	ruBABILongQA3	9.0	5.0	4.0	6.0	0.0	0.0	4.0
	ruBABILongQA4	25.2	29.2	15.6	5.9	0.0	0.0	12.7
	ruBABILongQA5	51.3	50.0	48.3	50.0	0.0	0.0	33.3
IV	ruSciPassageCount	18.0	8.0	1.0	2.0	0.0	0.0	4.8
	ruQasper	-	6.1	6.5	2.4	-	-	5.0
	ruGSM100	-	-	5.0	-	-	-	5.0

Table 15: The table presents the evaluation results of LongChat. **Dataset Name** shows the name of the dataset. The rows **4k**, **8k**, **16k**, **32k**, **64k**, **128k** show evaluation scores of datasets for each context length, respectively. The **Overall** score is obtained by averaging the results over each length.

	Dataset Name	4k	8k	16k	32k	64k	128k	Overall
I	Passkey	100.0	100.0	100.0	82.0	0.0	0.0	63.7
	PasskeyWithLibrusec	99.0	99.5	98.5	93.0	0.0	0.0	65.0
II	MatreshkaNames	4.7	2.7	0.7	0.0	0.0	0.0	1.3
	MatreshkaYesNo	50.2	50.0	50.0	50.0	0.0	0.0	33.4
	LibrusecHistory	-	21.9	9.4	3.1	0.0	-	8.6
	ruTREC	5.4	4.0	4.4	4.1	-	-	4.5
	ruSciFi	-	-	-	0.0	0.0	-	0.0
	ruSciAbstractRetrieval	41.7	21.9	18.1	0.0	0.0	0.0	13.6
	ruTPO	-	29.0	-	-	-	-	29.0
	ruQuALITY	-	54.5	43.9	-	-	-	49.2
III	LongContextMultiQ	0.5	5.0	1.5	0.0	0.0	0.0	1.2
	LibrusecMHQA	-	6.8	-	-	-	-	6.8
	ru2WikiMultihopQA	-	18.4	21.9	12.2	-	-	17.5
	ruBABILongQA1	27.0	23.0	23.0	0.0	0.0	0.0	12.2
	ruBABILongQA2	5.0	4.0	0.0	0.0	0.0	0.0	1.5
	ruBABILongQA3	7.0	3.0	4.0	1.0	0.0	0.0	2.5
	ruBABILongQA4	2.0	1.8	0.0	0.0	0.0	0.0	0.6
	ruBABILongQA5	20.0	18.0	15.0	0.0	0.0	0.0	8.8
IV	ruSciPassageCount	9.0	6.0	7.0	0.0	0.0	0.0	3.7
	ruQasper	-	3.5	3.3	0.9	-	-	2.6
	ruGSM100	-	-	5.0	-	-	-	5.0

Table 16: The table presents the evaluation results of GLM2-6B-32K. **Dataset Name** shows the name of the dataset. The rows **4k**, **8k**, **16k**, **32k**, **64k**, **128k** show evaluation scores of datasets for each context length, respectively. The **Overall** score is obtained by averaging the results over each length.

	Dataset Name	4k	8k	16k	32k	64k	128k	Overall
I	Passkey	77.5	82.5	57.5	37.0	0.0	0.0	42.4
	PasskeyWithLibrusec	71.0	70.0	56.0	46.5	0.0	0.0	40.6
II	MatreshkaNames	1.3	0.7	0.7	0.0	0.0	0.0	0.4
	MatreshkaYesNo	47.8	39.3	48.0	47.7	0.0	0.0	30.5
	LibrusecHistory	-	18.8	15.6	18.8	0.0	-	13.3
	ruTREC	0.0	2.0	0.0	0.0	-	-	0.5
	ruSciFi	-	-	-	2.8	0.0	-	1.4
	ruSciAbstractRetrieval	65.0	44.7	20.4	11.2	0.0	0.0	23.5
	ruTPO	-	6.8	-	-	-	-	6.8
	ruQuALITY	-	39.8	48.2	-	-	-	44.0
III	LongContextMultiQ	3.0	1.5	0.0	0.0	0.0	0.0	0.8
	LibrusecMHQA	-	7.8	-	-	-	-	7.8
	ru2WikiMultihopQA	-	40.8	28.9	21.1	-	-	30.3
	ruBABILongQA1	9.0	6.0	6.0	2.0	0.0	0.0	3.8
	ruBABILongQA2	1.0	1.0	0.0	0.0	0.0	0.0	0.3
	ruBABILongQA3	5.0	9.0	4.0	2.9	0.0	0.0	3.5
	ruBABILongQA4	0.0	1.0	0.0	0.0	0.0	0.0	0.2
	ruBABILongQA5	44.2	44.0	47.5	40.7	0.0	0.0	29.4
IV	ruSciPassageCount	13.0	5.0	2.0	3.0	0.0	0.0	3.8
	ruQasper	-	2.3	2.2	1.6	-	-	2.0
	ruGSM100	-	-	2.0	-	-	-	2.0

Table 17: The table presents the evaluation results of LongAlpaca. **Dataset Name** shows the name of the dataset. The rows **4k**, **8k**, **16k**, **32k**, **64k**, **128k** show evaluation scores of datasets for each context length, respectively. The **Overall** score is obtained by averaging the results over each length.

	Dataset Name	4k	8k	16k	32k	64k	128k	Overall
I	Passkey	100.0	100.0	0.0	0.0	0.0	0.0	33.3
	PasskeyWithLibrusec	100.0	100.0	0.0	0.0	0.0	0.0	33.3
II	MatreshkaNames	53.3	46.0	0.0	0.0	0.0	0.0	16.6
	MatreshkaYesNo	83.9	80.0	0.0	0.0	0.0	0.0	27.3
	LibrusecHistory	-	90.6	0.0	0.0	0.0	-	22.7
	ruTREC	59.5	50.0	0.0	0.0	-	-	27.4
	ruSciFi	-	-	-	0.0	0.0	-	0.0
	ruSciAbstractRetrieval	96.6	91.5	0.0	0.0	0.0	0.0	31.4
	ruTPO	-	78.1	-	-	-	-	78.1
	ruQuALITY	-	69.1	0.0	-	-	-	34.6
III	LongContextMultiQ	5.0	24.5	0.0	0.0	0.0	0.0	4.9
	LibrusecMHQA	-	46.1	-	-	-	-	46.1
	ru2WikiMultihopQA	-	53.1	0.0	0.0	-	-	17.7
	ruBABILongQA1	68.6	73.4	0.0	0.0	0.0	0.0	23.7
	ruBABILongQA2	14.0	10.9	0.0	0.0	0.0	0.0	4.1
	ruBABILongQA3	9.0	18.0	0.0	0.0	0.0	0.0	4.5
	ruBABILongQA4	57.3	60.3	0.0	0.0	0.0	0.0	19.6
	ruBABILongQA5	76.7	75.2	0.0	0.0	0.0	0.0	25.3
IV	ruSciPassageCount	31.0	8.0	0.0	0.0	0.0	0.0	6.5
	ruQasper	-	6.5	0.0	0.0	-	-	2.2
	ruGSM100	-	-	0.0	-	-	-	0.0

Table 18: The table presents the evaluation results of LLaMA-3-8B-Instruct. **Dataset Name** shows the name of the dataset. The rows **4k**, **8k**, **16k**, **32k**, **64k**, **128k** show evaluation scores of datasets for each context length, respectively. The **Overall** score is obtained by averaging the results over each length.

	Dataset Name	4k	8k	16k	32k	64k	128k	Overall
I	Passkey	100.0	100.0	0.0	0.0	0.0	0.0	33.3
	PasskeyWithLibrusec	100.0	99.5	0.0	0.0	0.0	0.0	33.2
II	MatreshkaNames	53.3	40.0	0.0	0.0	0.0	0.0	15.6
	MatreshkaYesNo	87.3	81.0	0.0	0.0	0.0	0.0	28.0
	LibrusecHistory	-	96.9	0.0	0.0	0.0	-	24.2
	ruTREC	51.4	54.0	0.0	0.0	-	-	26.3
	ruSciFi	-	-	-	0.0	0.0	-	0.0
	ruSciAbstractRetrieval	97.7	92.6	0.0	0.0	0.0	0.0	31.7
	ruTPO	-	75.7	-	-	-	-	75.7
	ruQuALITY	-	35.8	0.0	-	-	-	17.9
III	LongContextMultiQ	5.5	23.5	0.0	0.0	0.0	0.0	4.8
	LibrusecMHQA	-	45.1	-	-	-	-	45.1
	ru2WikiMultihopQA	-	53.1	0.0	0.0	-	-	17.7
	ruBABILongQA1	76.3	75.8	0.0	0.0	0.0	0.0	25.4
	ruBABILongQA2	19.6	6.9	0.0	0.0	0.0	0.0	4.4
	ruBABILongQA3	14.7	21.6	0.0	0.0	0.0	0.0	6.1
	ruBABILongQA4	63.5	58.2	0.0	0.0	0.0	0.0	20.3
	ruBABILongQA5	74.7	76.3	0.0	0.0	0.0	0.0	25.2
IV	ruSciPassageCount	19.5	3.5	0.0	0.0	0.0	0.0	3.8
	ruQasper	-	7.4	0.0	0.0	-	-	2.5
	ruGSM100	-	-	0.0	-	-	-	0.0

Table 19: The table presents the evaluation results of Saiga-LLaMA-3. **Dataset Name** shows the name of the dataset. The rows **4k**, **8k**, **16k**, **32k**, **64k**, **128k** show evaluation scores of datasets for each context length, respectively. The **Overall** score is obtained by averaging the results over each length.

	Dataset Name	4k	8k	16k	32k	64k	128k	Overall
I	Passkey	100.0	100.0	0.0	0.0	0.0	0.0	33.3
	PasskeyWithLibrusec	100.0	100.0	0.0	0.0	0.0	0.0	33.3
II	MatreshkaNames	40.0	20.0	0.0	0.0	0.0	0.0	10.0
	MatreshkaYesNo	62.2	59.0	0.0	0.0	0.0	0.0	20.2
	LibrusecHistory	-	90.6	0.0	0.0	0.0	-	22.7
	ruTREC	37.8	38.0	0.0	0.0	-	-	19.0
	ruSciFi	-	-	-	0.0	0.0	-	0.0
	ruSciAbstractRetrieval	97.1	88.1	0.0	0.0	0.0	0.0	30.9
	ruTPO	-	58.2	-	-	-	-	58.2
	ruQuALITY	-	17.1	0.0	-	-	-	8.5
III	LongContextMultiQ	9.5	32.5	0.0	0.0	0.0	0.0	7.0
	LibrusecMHQA	-	41.4	-	-	-	-	41.4
	ru2WikiMultihopQA	-	55.1	0.0	0.0	-	-	18.4
	ruBABILongQA1	68.0	57.0	0.0	0.0	0.0	0.0	20.8
	ruBABILongQA2	27.0	19.0	0.0	0.0	0.0	0.0	7.7
	ruBABILongQA3	28.5	25.9	0.0	0.0	0.0	0.0	9.1
	ruBABILongQA4	58.4	56.4	0.0	0.0	0.0	0.0	19.1
	ruBABILongQA5	67.2	68.7	0.0	0.0	0.0	0.0	22.6
IV	ruSciPassageCount	15.0	5.0	0.0	0.0	0.0	0.0	3.3
	ruQasper	-	6.5	0.0	0.0	-	-	2.2
	ruGSM100	-	-	0.0	-	-	-	0.0

Table 20: The table presents the evaluation results of LLaMA-3-3B. **Dataset Name** shows the name of the dataset. The rows **4k**, **8k**, **16k**, **32k**, **64k**, **128k** show evaluation scores of datasets for each context length, respectively. The **Overall** score is obtained by averaging the results over each length.

	Dataset Name	4k	8k	16k	32k	64k	128k	Overall
I	Passkey	100.0	97.5	12.5	0.0	0.0	0.0	35.0
	PasskeyWithLibrusec	100.0	100.0	30.0	0.0	0.0	0.0	38.3
II	MatreshkaNames	32.7	16.0	0.0	0.0	0.0	0.0	8.1
	MatreshkaYesNo	50.2	50.0	0.3	0.0	0.0	0.0	16.8
	LibrusecHistory	-	78.1	15.6	0.0	0.0	-	23.4
	ruTREC	2.7	10.0	3.3	0.0	-	-	4.0
	ruSciFi	-	-	-	2.8	0.0	-	1.4
	ruSciAbstractRetrieval	94.8	76.1	0.0	0.0	0.0	0.0	28.5
	ruTPO	-	39.6	-	-	-	-	39.6
	ruQuALITY	-	22.8	11.8	-	-	-	17.3
III	LongContextMultiQ	4.0	22.0	0.5	0.0	0.0	0.0	4.4
	LibrusecMHQA	-	34.1	-	-	-	-	34.1
	ru2WikiMultihopQA	-	42.9	18.0	8.1	-	-	23.0
	ruBABILongQA1	63.0	63.0	0.0	0.0	0.0	0.0	21.0
	ruBABILongQA2	21.0	25.0	0.0	0.0	0.0	0.0	7.7
	ruBABILongQA3	29.0	25.0	0.0	0.0	0.0	0.0	9.0
	ruBABILongQA4	42.9	31.6	0.0	0.0	0.0	0.0	12.4
	ruBABILongQA5	70.0	69.3	0.0	0.0	0.0	0.0	23.2
IV	ruSciPassageCount	4.0	4.0	0.0	0.0	0.0	0.0	1.3
	ruQasper	-	6.3	1.1	0.1	-	-	2.5
	ruGSM100	-	-	13.0	-	-	-	13.0

Table 21: The table presents the evaluation results of Mistral-7B-V0.1. **Dataset Name** shows the name of the dataset. The rows **4k**, **8k**, **16k**, **32k**, **64k**, **128k** show evaluation scores of datasets for each context length, respectively. The **Overall** score is obtained by averaging the results over each length.