Creation of the Russian PIQA Dataset: Technical Report

Daria Pugacheva
AIRI
d.pugacheva@airi.net

Olga Popova
AIRI
popova@airi.net

Elena Tutubalina
AIRI
tutubalina@airi.net

Abstract

We present the construction of a Russian PIQA-inspired dataset. To develop this resource, we collected instructional and everyday reasoning examples from Russian school textbooks across multiple grade levels. Large language models (LLMs) were then employed to generate questions and candidate solutions, which were subsequently refined through manual editing to ensure adherence to PIQA dataset creation guidelines and to preserve naturalness in Russian phrasing. The resulting dataset of 100 questions provides a high-quality benchmark for evaluating commonsense physical reasoning in the Russian language. We also performed a zero-shot evaluation of several LLMs on this dataset and reported performance metrics, highlighting both the challenges and opportunities of extending physical common-sense reasoning tasks beyond English.

1 Creation Methodology

Commonsense reasoning about physical interactions is a key challenge for large language models (LLMs). While benchmarks such as PIQA (Bisk et al., 2020) exist for English, comparable datasets for Russian remain scarce. To address this gap, we constructed a Russian adaptation of the PIQA dataset consisting of 100 examples drawn from Russian school textbooks across multiple grade levels (Chudinova and Bukvaryova, 2020; Tereshchuk et al., 2013; Khodzitskaya et al., 2013; Ayubov et al., 2022a,b; Gendenshteyn et al., 2020; Glozman and Kozhina, 2024). Each example presents a natural-language prompt with two candidate solutions, only one of which is physically plausible.

The methodology for compiling the dataset combines the utilization of large language models (LLMs) and manual curation by subject-matter experts. A total of three models were utilized: GPT 5, GPT 4.1, and O4-mini. Each model was provided with prompts in Russian, resulting in the generation of 50 examples for each textbook. Additionally, 50 examples were generated by the O4-mini model using a prompt without a source. Subsequently, experts manually selected the most relevant examples, corrected any errors or cultural mismatches, and ensured conformity with the template. When necessary, the incorrect answer was manually edited to minimally differ from the correct answer in alignment with PIQA requirements. Statistics on the number of selected examples and their sources for the models are provided in Table 3. The distribution of the number of examples between school grades and individual sources is shown in Tables 4 and 1. Examples of the data set are presented in Table 2. These statistics indicate a deliberate selection strategy to cover diverse educational topics and grade levels.

Source (Russian)	Source (English)	#
Технология. 5 класс	Technology, Grade 5	40
Трудовое обучение (для девочек), 5 класс	Labor training (for girls), Grade 5	23
Окружающий мир, 3 класс	The World Around Us, Grade 3	7
Трудовое обучение (для мальчиков), 5 класс	Labor training (for boys), Grade 5	7
ОБЖ, 10 класс	Basics of Life Safety, Grade 10	6
Физика, 7 класс	Physics, Grade 7	5
ОБЖ, 11 класс	Basics of Life Safety, Grade 11	4
Other		8
Total		**100**

Table 1: Counts per source (Russian books with English translation).

Prompt	Solution 0	Solution 1	Label
При обнаружении пожара в здании, если вы находитесь на верхнем этаже, In case of a fire in the building, if you are on the top floor,	нужно использовать лестницу для эвакуации Use the stairs for evacuation	нужно использовать лифт для эвакуации Use the elevator for evacuation	0
Для остановки кровотечения To stop bleeding	следует приподнять повреждённую конечность One should elevate the injured limb	не следует приподнять повреждённую конечность One should not elevate the injured limb	0
При пожаре нужно закрыть окна и двери, During a fire, you should close windows and doors,	чтобы обеспечить доступ свежего воздуха To ensure access to fresh air	чтобы ограничить доступ воздуха To limit access to air	1

Table 2: Examples in Russian with English translations.

System and user prompts for generating dataset examples

You are an expert with native proficiency in Russian. Your task is to contribute to the collection of data for evaluating physical commonsense reasoning in Russian. The dataset format is similar to the PIQA (Physical Interaction: Question Answering) test, where each example consists of a question ("prompt") with two answer options ("solution0" and "solution1") and a label indicating the correct solution (0 if "solution0" is correct, and 1 if "solution1" is correct). Only one of the solutions should be correct. For each example, the solution should relate to the physical properties of one or more objects. The answer should be known to an average person with Russian as their native language. Culturally significant examples for Russian are encouraged. Some expressions may be difficult to translate into English and may have regional and/or cultural significance, such as examples related to local products, places, everyday objects, customs, traditions, religions, literature, folklore, or art forms. Use examples of varying lengths. Avoid including too many short examples, as they may be too simple for large language models. If possible, most examples (question + solution) should be longer than 25 words. Some questions should consist of multiple sentences. The two solution options should be as similar as possible (e.g., differing by only one or two words or the order of two phrases). One solution should be unequivocally correct, while the other should be incorrect. To ensure the test is not too easy, the incorrect solution should not be absurd. Strive to begin all examples differently.

Example from the dataset:
{ "prompt": "To prepare a honey-chipotle marinade for chicken: In a glass bowl, use a metal spoon to mix three tablespoons of canola oil, two teaspoons of minced garlic from a jar, three tablespoons of honey, and two tablespoons of chopped canned chipotle peppers in adobo sauce until smooth.", "solution0": "Transfer the marinade to a zip-lock bag, place the chicken pieces in the bag, squeeze out the excess air, and seal the bag. Mix the chicken in the marinade and let it sit overnight.", "solution1": "Transfer the marinade to a paper bag, place the chicken pieces in the bag, squeeze out the excess air, and seal the bag. Mix the chicken in the marinade and let it sit overnight.", "label": 0 }

Now, generate 50 examples based on knowledge from the attached file.

LLM	Source	Count
O4-mini	Technology, Grade 5 Basics of Life Safety, Grade 10 without source	22
GPT 4.1	Physics, Grade 7 The World Around Us, Grade 3 Labor training (for girls), Grade 5 Labor training (for boys), Grade 5	22
GPT 5	Labor training (for girls), Grade 5 Technology, Grade 5 Basics of Life Safety, Grade 11	56

Table 3: LLM sources and counts of selected examples per model

Grade (Russian)	Grade (English)	Count
3 класс	Grade 3	7
5 класс	Grade 5	70
7 класс	Grade 7	5
10 класс	Grade 10	6
11 класс	Grade 11	4
Total		**92**

Table 4: Counts per grade.

Model	Ver.	Accuracy	# Questions
YandexGPT Pro	5	0.99	100
YandexGPT Lite	5	0.93	100
Gemma3 27B		0.95	100
Llama 70B	3.3	0.97	100
Llama 8B	3.1	0.74	96
Random baseline		0.41	100

Table 5: Zeso-shot evaluation results. For each model, we omitted questions if a model refuses to answer.

2 Experiments

This report presents our evaluation of several LLMs on this dataset in a zero-shot setting. We used the Yandex SDK[1] to query different LLMs. Each model received the prompt and the two candidate solutions and was asked to choose the correct one. We computed the accuracy as an evaluation metric. We used Russan prompt in Table 6 for all models. The evaluation results are presented in Table 5.

The evaluation results in Table 5 show that all large-scale LLMs substantially outperform the random baseline (0.41). Among them, YandexGPT Pro (v. 5) achieves the highest accuracy of 0.99, nearly perfect performance across the 100 evaluation examples. YandexGPT Lite (v. 5) reaches 0.93, slightly below its larger counterpart but still well above the baseline. In contrast, Llama 8B (v. 3.1) attains a lower accuracy of 0.74 and refused to answer four questions. Overall, the results indicate that model scale has a substantial impact on performance.

[1] https://yandex.cloud/ru/docs/foundation-models/concepts/generation/models

References

E.N. Ayubov, D.Z. Proshchepov, and et al. 2022a. *Basics of Life Safety: Textbook for Grade 10 of General Educational Organizations. Basic Level. (in russian)*. Russkoe slovo – uchebnik.

E.N. Ayubov, D.Z. Proshchepov, and et al. 2022b. *Basics of Life Safety: Textbook for Grade 11 of General Educational Organizations. Basic Level. (in russian)*. Russkoe slovo – uchebnik.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

E.B. Chudinova and E.N. Bukvaryova. 2020. *The World Around Us. Grade 3 (in russian)*. Prosveshenie.

L.E. Gendenshteyn, A. A. Bulatova, and et al. 2020. *Physics. Grade 7 (in russian)*. Prosveshenie.

E.S. Glozman and O. A. Kozhina. 2024. *Technology. Grade 5. Textbook. Federal State Educational Standard (in russian)*. Prosveshenie.

I.Yu. Khodzitskaya, N.N. Pavich, and et al. 2013. *Labor training (for girls). Grade 5 (in russian)*. Aksioma.

B.N. Tereshchuk, V.K. Zagornyy, and et al. 2013. *Labor training (for boys). Grade 5 (in russian)*. Geneza.

Russian	**English**
Ты — эксперт по здравому смыслу и физическому взаимодействию с объектами. Твоя задача — выбирать правильное решение в бытовых ситуациях. Отвечай строго числом: 0 или 1.	You are an expert in common sense and physical interaction with objects. Your task is to choose the correct solution in everyday situations. Answer strictly with a number: 0 or 1.
Задача: {question} Вариант 0: {solution0} Вариант 1: {solution1} Выбери правильный вариант ответа. Отвечай строго числом: 0 или 1.	Task: {question} Option 0: {solution0} Option 1: {solution1} Choose the correct option. Answer strictly with a number: 0 or 1.

Table 6: System and user prompts in Russian with English translation.