DOI: 10.14529/jsfi250301

RuParam: a Russian Parametric Dataset for LLM Evaluation

Pavel V. Grashchenkov[1], *Lada I. Pasko*[1], *Regina R. Nasyrova*[1]

We introduce RuParam, a parametric dataset designed to evaluate the acquisition of Russian by large language models (LLMs). This corpus mirrors the structure of the BLiMP family of datasets by containing minimal pairs of sentences. However, our goal was to expand its scope as much as possible by incorporating diverse phenomena from several domains of Russian grammar. A significant portion of the data originates from the Tests of Russian as a Foreign Language (TORFL); similar sources were not previously used for linguistic evaluation of LLMs. Additionally, this study details experimental findings involving six LLMs. These LLMs, sourced from multiple developers, vary in size and pretraining data, which affects their proficiency in Russian. We investigate how effectively these models handle universal, typological, and Russian-specific grammatical features. Our results indicate that while most of the models demonstrate relatively high performance, they struggle significantly with some of the Russian-specific categories.

Keywords: Large Language Models, linguistic evaluation, minimal pairs, Russian, linguistic parameters, language acquisition.

Introduction

In theoretical linguistics, the ability to differentiate between grammatical and ungrammatical sentences (i.e. ones that conform to the rules of grammar of a certain language and ones that do not) has long been considered as a core part of human linguistic competence. A common way of illustrating grammaticality are minimal pairs – sentences that are identical in terms of their lexical content, but differ in whether they violate some grammatical constraint, cf. (1):

(1) *The bird is singing* vs **The bird are singing.*

In recent years, such minimal pairs have moved beyond papers on theoretical linguistics into the field of large language model (LLM) evaluation, with BLiMP [28] being the pioneer in this area. This seems to be a logical step, since if LLMs are to accurately replicate human linguistic behavior, they should be tested on tasks that speakers of natural language are known to succeed at.

Since BLiMP, a lot of work has been done: similar corpora have been developed for various languages. We present **RuParam** – a Russian dataset of minimal pairs[2]. Although there have already been some successful attempts to create grammaticality datasets for Russian – RuCoLA [14] and RuBLiMP [21], we believe that our corpus makes a significant contribution to the field. RuParam addresses some of the shortcomings of other grammaticality corpora, such as semi-automatic data generation, only a small range of linguistic phenomena covered, scarce linguistic annotation, and natural variability in sentence acceptability.

Our main contributions are as follows:

- We introduce a new grammaticality dataset of 11,336 minimal pairs in Russian.
  - Our data are classified into 150 linguistic categories covering both universal and language-specific phenomena. The phenomena range from basic concepts of grammar, such as standard cases of subject-verb agreement and case assignment, to more nuanced cases, e.g. clitic placement, licensing of negative polarity items, allomorph distribution and so on.

[1] Lomonosov Moscow State University, Moscow, Russian Federation

[2] https://github.com/grapaul/RuParam

  - One part of our dataset (8,039 pairs) originates from an independent source of grammaticality minimal pairs – multiple choice tasks from the Test of Russian as a Foreign Language (TORFL). To our knowledge, materials of language proficiency tests for non-native speakers had not previously been used for linguistic evaluation of LLMs.
  - The other part of the dataset (3,297 pairs) was taken from Russian corpora (Russian National Corpus, RuConst) and manually modified by trained linguists.
- We evaluate six LLMs on our data using the method of metalinguistic prompting. Although none of the models reach 100% accuracy, some of them are very close to this threshold.

The article is structured as follows. Firstly, we provide some essential background on evaluating large language models using linguistic benchmarks. Secondly, we introduce the RuParam dataset and offer a comprehensive description thereof. Next, we present an experimental study, where RuParam was used to evaluate six different language models – those extensively trained on Russian data and others that were not. Lastly, we analyze and summarize our findings.

1. Related Work

BLiMP (Benchmark of Linguistic Minimal Pairs) [28] was the first wide-range dataset to use the grammaticality minimal pair format. BLiMP covers 12 linguistic phenomena of English, for which 67 minimal pair templates were created by linguists. The data were automatically generated using these templates, which enabled the massive size of the dataset (67K minimal pairs). However, this approach has certain limitations due to differences between generated and naturally occurring data. For example, automatically generated data fall short of corpus sentences in both length [5] and structural diversity [26], which makes the evaluation results not entirely representative. Recently, BLiMP-style datasets have been developed for a variety of languages: Chinese (CLiMP [30], SLING [19]), Dutch (BLiMP-NL [20]), Japanese (JBLiMP [16]), Russian (RuBLiMP [21]), Turkish (TurBLiMP [2]), and Urdu (UrBLiMP [1]). MultiBLiMP [12], in turn, covers 101 languages, focusing on just one linguistic phenomenon, namely verb-subject agreement. Some of these benchmarks use the original method of data generation (e.g. CLiPM), while others employ a more naturalistic approach, using examples from corpora as a starting point (e.g. MultiBLiMP, SLING, RuBLiMP, UrBLiMP).

Linguistic acceptability was used in LLM evaluation before the minimal pair approach. The predecessor of BLiMP, CoLA (Corpus of Linguistic Acceptability) [29], created for English, contains individual sentences tagged as either grammatical or ungrammatical; the sentences do not have a counterpart differing in grammaticality. All the sentences in CoLA, along with grammaticality judgements, come from works on theoretical linguistics such as articles and textbooks. As in the case of BLiMP, equivalents for CoLA have been created for many other languages, including Catalan (CatCoLA [3]), Chinese (CoLAC [10]), Danish (DaLAJ [27]), Japanese (JCoLA [17]), Hungarian (HuCoLA [13]), Italian (ItaCoLA [25]), Norwegian (NoCoLA [11]), Russian (RuCoLA [14]) and Spanish (EsCoLA [4]). Importantly for us, some of these datasets – DaLAJ and NoCoLA – use data from the field of second language (L2) acquisition. In both of them, the data come from L2 learner corpora. The ungrammatical sentences are those where L2 learners made mistakes, while grammatical ones are those corrected by native speakers.

In general, sources originating from the educational field have been extensively used for the task of LLM evaluation. MMLU [9] is one prominent example. It aims to evaluate the LLM's knowledge of factual information covering a wide range of subjects; the data stem from multiple-choice questions found in textbooks and examination materials from diverse fields. Among grammaticality datasets, RuCoLA includes ungrammatical sentences derived from tasks of Unified State Exam in Russian, aimed at high school graduates. However, the status of such sentences as ungrammatical is doubtful. They rather represent prescriptive norms that are not necessarily part of a native speaker's grammar – otherwise, there would be no point in using them as part of an exam for Russian schoolchildren.

Regarding the procedure of model evaluation for BLiMP and its equivalents, the preference of a model is standardly defined as the difference in the probability of sentences forming a minimal pair. This experimental design serves as a solution to the limitations of CoLA-style evaluation. Since CoLA views acceptability judgement as a binary classification task, it is necessary to train a supervised classifier prior to LLM evaluation. Other factors, besides grammaticality, such as sentence length and word frequency, are inevitably involved. On the contrary, BLiMP allows for separating the grammatical contrast from these additional factors. Furthermore, other approaches have been recently proposed for BLiMP-style data, such as metalinguistic prompting [18]. In this case, LLMs are treated as human subjects of linguistic experiments (see e.g. [6] for an overview of human acceptability judgement task): the prompt consists of an explicit verbal instruction to choose the more acceptable sentence out of the minimal pair. This method allows researchers to investigate whether LLMs have acquired human-like linguistic introspection.

2. RuParam

2.1. Corpus Structure

RuParam includes 11,336 minimal pairs[3]. The dataset consists of two parts: the first part (8,039 pairs, 70.92%) is based on data from the Test of Russian as a Foreign Language[4] (TORFL); the second part (3,297 pairs, 29.08%) represents a parametric dataset created by linguists and based on naturally occurring sentences. The dataset covers the wide range of linguistic phenomena corresponding to 150 smaller categories. One of the goals of creating RuParam was to maximize the number of diagnostic grammatical features and to diversify the methods for obtaining contrast within each feature. This approach aims to enable not only an overall assessment of linguistic proficiency, but also a detailed analysis of the level of acquisition of specific grammar points.

As the data in the first part come from TORFL materials, it covers phenomena specific (although not necessarily unique) to Russian. The tasks in TORFL were independently created by experts in acquisition of Russian by learners with different native languages. Therefore, this part addresses crucial grammatical phenomena and is particularly relevant for testing multilingual LLMs. The purpose of the second part is twofold. First, it covers universal features (such as projectivity and island constraints) that are characteristic of natural language in general and are therefore absent from TORFL tasks. Second, it includes phenomena that are specific to Russian, but underrepresented in TORFL because of methodological reasons. In the next sections, we discuss the data generation procedure and the phenomena in more detail.

[3] The earlier version of RuParam, containing 4,382 minimal pairs, along with the results of LLM evaluation, was presented in [8].

[4] `https://testingcenter.spbu.ru/ru/materials.html`
`https://www.pushkin.institute/certificates/cct/tests-online/?ysclid=meqsg6x5b6171434445`
`https://test.tsu.ru/ru/trki`

Part 1: TORFL

This part of the data originates from multiple-choice "Vocabulary/Grammar" tasks of TORFL. Each task consists of a sentence (or several sentences) with a gap. There are 3 or 4 options for filling in the gap, only one of which is correct. For example, the task in (2) tests the acquisition of adjective agreement. Only option B forms a grammatical sentence in Russian.

(2) *Это очень ... здание.*
this very building.N
'It is a very ... building.'
A. высокая
tall.F
B. высокое
tall.N
C. высокий
tall.M

The minimal pairs were generated by filling in the gap with each option. The correct answer produces the grammatical member of a minimal pair, while the incorrect answers create the ungrammatical ones. The grammatical sentence was paired with all the ungrammatical options, generating two or three minimal pairs per task. Some of the original tasks from TORFL were excluded from the data because they required the use of prior context to choose the correct option (e.g. when a series of tasks represents a coherent text). TORFL covers all CEFR (Common European Framework of Reference) levels of language proficiency: A1, A2, B1, B2, C1, and C2. A1 corresponds to basic knowledge, and C2 is closest to native speaker proficiency. The complexity tags of the original tasks are included in the dataset. The quantity of data per level is shown in Tab. 1.

Table 1. Level distribution in TORFL subset of RuParam

Level	A1	A2	B1	B2	C1	C2
Number of pairs	1755	932	2413	1588	1147	204
%	21.84	11.59	30.00	19.75	14.27	2.54

While the TORFL tasks cover a wide variety of the Russian language phenomena, there is no linguistic annotation available to users. We identified the phenomena used in the tasks and provided annotation for each minimal pair. This was manually done by trained linguists, and each tag was verified by at least one other expert. The annotation included the following parameters, which are divided into 29 categories. We list the most important categories below:

- attributive and predicative agreement;
- lexical selection of all major parts of speech;
- government of different parts of speech;
- use of non-finite forms;
- use of aspect, tense, modality;
- use of coordinating and subordinating conjunctions;
- use of constructions with numerals;
- correctness of using particular parts of speech;
- use of copulas in nominal predications;

- grammaticality of voice forms;
- use of various types of pronouns;
- grammaticality of negative constructions.

Most of these parameters are found in tasks of different levels. Many categories, such as aspect, attributive agreement, and verbal selection, are present in tasks of all six levels.

For some pairs, more than one tag was appropriate. For example, in (3), the ungrammatical form *начинается* ‘start.IPF.PRS’ differs from the grammatical one *начался* ‘start.PF.PST’ in both aspect and tense. Such minimal pairs were replicated in the dataset, with only one grammatical category present in the annotation of each instance of the pair.

(3)	*gram*	*Спектакль*	***начался***	*давно,*	*вы*	*уже*	*опоздали.*
		performance	start.PF.PST	long ago	you	already	are late
		‘The performance started a long time ago, you are already too late.’					
	ungram	*Спектакль*	***начинается***	*давно,*	*вы*	*уже*	*опоздали.*
		performance	start.IPF.PRS	long ago	you	already	are late
		‘The performance is starting a long time ago, you are already too late.’					

Part 2: Parametric Dataset

While the approach to the first part of the dataset was data-driven (we annotated pairs created independently for TORFL), the starting point for the second part was grammatical parameters. We used our experience in theoretical linguistics to identify categories that are important for Russian and human language in general, but which are insufficiently covered or not covered at all in TORFL tasks.

The total number of categories in the second part of RuParam is 121. The number of examples varies across different parameters, but each one is represented by at least 15 minimal pairs.

Universal phenomena included in the dataset are binding principles; island constraints (coordinate structure, complex NP, adjunct, subject, and others); projectivity.

Other categories are **specific to Russian**. Some of them, such as the directionality of branching, the use of null subjects, or *wh*-word placement (on the left periphery vs *in situ*), represent **parameters of typological variation**. The remaining phenomena covered in the dataset include the following: different types of agreement; clitic placement (P2-clitics, clitics forming conditional clauses); distribution of non-finite forms; licensing of different types of negative polarity items; case of nominal predication; control; voice; depictives; analytical tense forms; analytical comparative and superlative forms; morphophonological variation; double conjunctions; matching in free relatives; constructions with numerals; distribution of the short form of adjectives; genitive marking under negation, and others.

As in many other benchmarks with a similar purpose, grammatical sentences were derived from corpora. We used two corpora of Russian: RuConst [7] and Russian National Corpus, RNC [15]. These sentences were modified by experts in theoretical linguistics to ensure that an ungrammatical counterpart in a minimal pair violated some linguistic constraint. For example, the ungrammatical sentence in (4) is a case of non-projectivity. Although word order in Russian is relatively free, such sentences are ruled out. In (4), the adjective *главного* ‘in.chief.GEN’ was dislocated so that it is separated from the nominal head that it modifies (*редактора* ‘editor.GEN’) by another nominal phrase (*смены* ‘change.GEN’). Each minimal pair was verified by at least one other expert to confirm that the expected grammatical contrast was present. The total number of errors did not exceed 1%.

(4) *gram* *О причинах смены* ***главного*** *редактора не сообщается.*
about reasons change.GEN in chief.GEN editor.GEN not is reported
'The reasons for the change of the editor-in-chief are not reported.'
ungram *О причинах* ***главного*** *смены редактора не сообщается.*
about reasons in chief.GEN change.GEN editor.GEN not is reported
Int. 'The reasons for the change of the editor-in-chief are not reported.'

3. Evaluation

We assessed the abilities of several LLMs to distinguish between grammatical and ungrammatical sentences in our dataset. The following subsections introduce the models that were tested and describe the experimental setup.

3.1. Models

The models for evaluation were chosen based on the following criteria: model size, and accessibility as well as quantity of Russian data used during the training procedure.

The first group included closed-source foundation models of large size that were extensively exposed to Russian during training. Consequently, we expected these models to provide the most reliable judgements:

- **GigaChat-2-Max**[5] – the largest and most efficient version of GigaChat models;
- **YandexGPT 5 Pro**[6] – likewise, the most advanced model of the YandexGPT family.

The second group, as opposed to the first one, consisted of open-source models with 7-8B parameters, based on the fruitful Qwen2.5 model [22]. Using these models, we aimed to investigate whether the ability to differentiate between sentences in a minimal pair is influenced by a smaller model size, multilingual pre-training, and different adaptation techniques:

- **Qwen2.5-7B-Instruct**[7] [22] – the instruction-tuned multilingual model pre-trained on a large-scale dataset and demonstrating high performance on various tasks from language understanding to coding;
- **T-lite-1.0**[8] – an adaptation of Qwen2.5 model for the Russian language, which was pre-trained in two stages using a combination of Russian and English texts, as well as instruction data, then fine-tuned to follow instructions and preferences;
- **RuadaptQwen2.5-7B**[9] [23, 24] – the modification of Qwen2.5-7B with the tokenizer better suited to the morphology of Russian. The model was also trained on Russian data and with Learned Embedding Propagation procedure.

Moreover, we added **Mistral-7B-Instruct-v0.3**[10] to the comparison, as it is also a multilingual 7B model with different pre-training data. However, it is more English-oriented than other models in terms of pretraining data, which could hinder its performance on our task.[11]

[5] `https://developers.sber.ru/docs/ru/gigachat/models/updates`

[6] `https://yandex.cloud/ru/docs/foundation-models/concepts/yandexgpt/models`

[7] `https://huggingface.co/Qwen/Qwen2.5-7B-Instruct`

[8] `https://huggingface.co/t-tech/T-lite-it-1.0`

[9] `https://huggingface.co/RefalMachine/RuadaptQwen2.5-7B-Lite-Beta`

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

[11] Hereafter we will address the models by the first letter in the title: G(igaChat-2-Max), Y(andexGPT 5 Pro), T(-lite-1.0), R(uadaptQwen2.5-7B), Q(wen2.5-7B-Instruct), M(istral-7B-Instruct-v0.3).

Какое из двух предложений является правильным и грамматичным с точки зрения русского языка?
Предложение 1. <sent_one>
Предложение 2. <sent_two>
Ответь только одной цифрой 1 или 2, ничего не добавляя.

Which of the two sentences is correct and grammatical according to the Russian language?
Sentence 1. <sent_one>
Sentence 2. <sent_two>
Respond with only one number 1 or 2, without adding anything.

Figure 1. The prompt used for model evaluation

3.2. Setup

To evaluate a model's ability to make grammaticality judgements, it was prompted with the instruction in Fig. 1.

The model's response should have been either "1" or "2", denoting the number of the grammatical sentence. As our corpus was designed for diagnostic purposes, there was no training sample, so we only tested the models in zero-shot settings and did not study their abilities in the few-shot setup or after fine-tuning.

LLMs are prone to position bias [31], i.e. they tend to choose the first or the second option regardless of their contents. Therefore, each minimal pair was evaluated twice: with sentences given in their default order, and in the opposite one. The model's response was considered correct only if the model preferred the grammatical option in both iterations. The order of examples within the corpus is arbitrary, different grammatical categories are interleaved so that LLMs do not accumulate guesses about the type of ungrammaticality.

4. Results and Discussion

4.1. General Results

Table 2 summarizes the results of LLM evaluation. Some minimal pairs were rejected by the models due to ethical considerations, so no answer was given regarding the grammaticality of sentences. This is explained by the source of our data: some corpora examples, especially those coming from news, may contain discussions on sensitive topics, such as politics and health, cf. (5). However, the amount of filtered data was not significant in most cases. The second column of Tab. 2 presents the percentage of correct answers overall, while the third column shows the percentage after the filtered examples were excluded. The other two columns offer statistics for two parts of the dataset separately.

(5) *Турецкие пограничники задержали судно, перевозившее тонну героина.*
Turkish border guards detained ship carrying ton of heroin
'Turkish border guards detained a ship carrying a ton of heroin.'

We anticipated that the closed-source large-scale models with substantial exposure to Russian would exhibit superior performance. This expectation was confirmed by G, which had the

Table 2. Model evaluation results

Model	Accuracy	Accuracy (filtered)	TORFL accuracy (filtered)	Parametric data accuracy (filtered)
G	97.85	97.92	98.05	97.54
Y	95.98	97.61	97.47	97.94
T	87.79	91.06	90.81	91.66
R	89.13	89.13	89.29	88.75
Q	87.31	87.31	87.24	87.50
M	52.31	61.67	60.24	64.84

best results and was closely followed by Y. Although neither of the models reached 100% accuracy, they came close to this threshold. The open-source Qwen-based models (T, R, Q) achieved lower results and differed from each other by approximately two percentage points. M scored significantly lower than all other models, as expected given that it had less Russian pretraining data.

Regarding the difference between the two parts of RuParam, there seems to be none in terms of LLMs performance. The results for the two parts are approximately the same for all models. The only exception is M, which performs significantly better on the parametric part of the data (64.84% vs 60.24%). This may be due to the fact that the second part of the dataset includes some universal phenomena that do not necessarily require extensive pretraining on Russian data.

4.2. Results by TORFL Levels

The data in the first part of our dataset coming from TORFL was distributed among six CEFR complexity levels, ranging from A1 to C2. We expect that if the linguistic competence of LLMs is similar to the human one, the results of the models will correlate with the complexity levels: the more difficult the tasks, the lower the accuracy is. This prediction is partially borne out. The models' results are mostly in accordance with the levels, but some exceptions are present – this is shown in Tab. 3 (the results before filtration are given).

Table 3. Accuracy by TORFL levels

	G	Y	T	R	Q	M
A1	98.69	98.69	91.00	91.80	90.55	52.05
A2	98.61	97.53	89.06	90.88	88.63	48.28
B1	98.14	97.93	87.53	90.05	88.11	47.91
B2	98.11	97.36	88.41	88.22	85.77	54.97
C1	97.82	95.82	85.35	87.71	84.39	45.16
C2	89.71	89.22	69.12	68.14	69.61	39.22

4.3. Results by Linguistic Phenomena

The models demonstrate some common patterns in error frequency. The most complex phenomena come from the second part of the dataset, which consists of data created from corpus examples. Interestingly, most mistakes are made in Russian-specific categories by both groups of models: industrial (G, Y) and open-source (T, R, Q, M). Table 4 shows the rating of most complex phenomena.

Table 4. The most complex linguistic phenomena.
‘The category is marked “+” if it is among the top-10 error-prone categories for a model

Category/Model	G	Y	T	R	Q	M	sum
GOV_LOC_1	+	+	+	+	+		5
SUPER_3	+	+	+		+	+	5
COND_1			+	+	+	+	4
COND_5	+	+			+	+	4
IMP_VAR	+	+	+			+	4
PREP_VAR	+	+		+	+		4
DISTR	+	+	+				3
FUT_ASP_1			+	+		+	3
FUT_ASP_2				+	+	+	3
GOV_LOC_2	+				+	+	3

Examples and descriptions of the most complex phenomena are given in Tabs. 5, 6, and 7; the data are presented as [*gram*/*ungram*].

Table 5. The description of the most complex linguistic phenomena along with examples illustrating them. Part 1

GOV_LOC_1	Some Russian nouns, such as *шкаф* ‘closet’, *угол* ‘corner’, *нос* ‘nose’ have a special form of the locative case which is required after certain prepositions (e.g. *в* ‘in’, *на* ‘on’). For all other nouns, the prepositional case is used after these prepositions. The grammatical sentences in this category contain a noun in the locative case (e.g. *шкаф-у* ‘closet-LOC’), while in their ungrammatical counterparts the regular prepositional case form is used instead (e.g. *шкаф-е* ‘closet-PREP’).
	26-летний мужчина прятался в ***[шкафу/шкафе]****, где девочка* 26-year-old man was hiding in [closet.LOC/closet.PREP] where girl *хранит свою одежду.* keeps REFL clothes ‘The 26-year-old man was hiding in the closet where the girl keeps her clothes.’
SUPER_3	One of the ways to form the superlative degree form of an adjective in Russian is through the use of the circumfix *наи-...-ейш-*. In ungrammatical sentences, the second part of this circumfix (-*ейш-*) is omitted, while the first part (*наи-*) remains.
	Дебаты – это ***[наиважнейшая/наиважная]*** *часть* debates COP [SUPER-important-SUPER-F.SG/SUPER-important-F.SG] part *избирательного процесса.* electoral process ‘Debates are the most important part of the electoral process.’

As one can see, many complex phenomena are associated with allomorphy. To choose between allomorphs, one needs to know about the properties of individual lexemes (e.g. the presence of a special locative form) and about the context: both the lexical and the phonological features are important. We assume that it is the multifactorial nature of allomorphy that makes the

Table 6. The description of the most complex linguistic phenomena along with examples illustrating them. Part 2

COND_1	In counterfactual conditionals, both clauses must include the particle *бы* responsible for the conditional mood. The ungrammatical sentences are produced by eliminating this particle from subordinate clauses headed by *если* 'if'.
	Все было бы нормально, если ***[бы/∅]*** *разговоры подкреплялись литературой.* everything was COND normal if [COND/∅] conversations were supported by literature 'Everything would be fine if the conversations were supported by literature.'
COND_5	In Russian, there is a construction with conditional semantics that includes a *wh*-word and the negative particle *ни*. The ungrammatical examples are created by omitting *ни*.
	Как это ***[ни/∅]*** *горько признать, мы еще до покупателя не доехали.* how this [PART/∅] bitter admit we yet to buyer not reached 'As much as I hate to admit it, we have not reached the buyer yet.'
IMP_VAR	The singular imperative form is formed by a suffix with two allomorphs: -*и* (*сохран-и* 'save-IMP') and -∅ (*встань-∅* 'get.up-IMP'). The distribution is determined both phonologically and lexically: -*и* is generally used when a verb has stress on its inflection in the present tense, although there are many exceptions (*прыгн-и* 'jump-IMP', *вытян-и* 'draw-IMP'). The modification of sentences in this category consists of changing the required allomorph to the other one.
	[Сохрани/Сохрань] *свою жизнь ради людей, которые в тебя верят!* [save-IMPER.1/save-IMPER.2] REFL life for people that in you believe 'Save your life for the sake of the people who believe in you!'
PREP_VAR	Short prepositions ending with a consonant have an allomorph ending with -*о*, e.g. *с*/*со* 'with'. The choice of allomorph depends on the phonological conditions: if the word following the preposition begins with the same consonant the preposition ends with, -*о* is inserted (*с мнением* 'with an opinion' vs *со ссылкой* 'with reference'). Likewise, the preposition *о* 'about' has an allomorph that ends with a consonant *об* which is used before vowels (*о мнении* 'about the opinion' vs *об этом* 'about this'), and another lexically selective one *обо* (*о мнении* 'about the opinion' vs *обо мне* 'about me'). In the ungrammatical sentences, an incorrect allomorph of a preposition is used.
	Об этом сообщает РИА Новости ***[со/с]*** *ссылкой на режиссера.* about it reports RIA Novosti [with(1)/with(2)] reference to director 'This is reported by RIA Novosti with reference to the director.'
DISTR	This category deals with collective predicates (e.g. *пересекаться* 'cross'). Those require the use of a coordinated noun phrase or a plural noun as their subject. The ungrammatical sentences were altered to contain a semantically inappropriate subject.
	[***Наши пути*** *не* ***пересекались****.* / ***Наш путь*** *не* ***пересекался****.*] [our.PL path.PL not crossed.PL / our.SG path.SG not crossed.SG] 'Our paths did not cross.'

models struggle. For instance, models perform significantly better on agreement even though the difference between correct and incorrect forms is often only one character, just as in examples involving allomorphy. This is surprising given that most factors determining the choice of an

Table 7. The description of the most complex linguistic phenomena along with examples illustrating them. Part 3

FUT_ASP_1	Future tense in Russian is formed synthetically for perfective verbs (*совершит* 'make.PF.FUT') and analytically for imperfective verbs (*будет совершать* 'will make.IPF.INF'). In pairs of this category, the ungrammatical counterpart has a perfective infinitive in an analytical form instead of the imperfective one.
	Планируется, что автомобиль будет ***[совершать/совершить]*** planned that car will [make.IPF.INF/make.PF.INF] *вертикальный взлет и посадку.* vertical take off and landing 'It is planned that the car will take off and land vertically.'
FUT_ASP_2	Just as in the case of FUT_ASP_1, the ungrammatical sentences in this category have an erroneous use of the analytical future form of a perfective verb. The difference is that the grammatical member of the minimal pair includes a correct synthetic form, but not an analytical one.
	За моральный ущерб мужчина ***[получит/будет получить]*** *100 рублей.* for moral damage man [receive.PF.FUT/will receive.PF.INF] 100 rubles 'The man will receive 100 rubles for moral damage.'
GOV_LOC_2	Similarly to GOV_LOC_1, this category deals with the locative/prepositional case distinction. The ungrammatical sentences demonstrate incorrect use of the locative form.
	О красном ***[снеге/снегу]*** *сообщали жители нескольких районов* about red [snow.PREP/snow.LOC] reported residents several districts *области.* region 'Residents of several districts of the region reported red snow.'

allomorph are local (e.g. the first character of the next word), while by agreement the goal and the probe can be located at a considerable linear distance. However, in the case of agreement, it is mostly a single factor that matters: the morphological form of the goal (e.g. the verb form has to correspond to the morphological features of the noun in the nominative case).

Conclusion

We introduced a new dataset for testing the competence of LLMs in Russian. Our corpus uses the minimal pair framework that is now common for the task of linguistic evaluation of LLMs. The data in this benchmark come from two sources: tasks of the Test of Russian as a Foreign Language and corpora of Russian. The first type of data is novel to the field, while the second one has been extensively used in similar datasets and is preferable to automatic data generation. Ungrammatical sentences in the first part of the dataset were independently created by the L2 acquisition experts, while those in the second part were manually generated specifically for the corpus by trained linguists. The dataset contains fine-grained linguistic annotation covering a wide range of categories. Some of them are universal, while others represent Russian-specific phenomena. All annotations were performed by experts in theoretical linguistics.

We evaluated six LLMs on our dataset. To do this, we used the metalinguistic prompting method, treating the models as if they were human subjects in linguistic studies. We found that large-scale models trained extensively on Russian demonstrate very high performance levels, close to the 100% threshold. As it was expected, smaller open-source models achieved lower results. The evaluation results on the TORFL part of our dataset show that the degree of success of LLMs mostly correlates with CEFR complexity levels of the tasks. This finding demonstrates one similarity between linguistic competence of LLMs and humans. Although models generally achieve relatively good results, some categories proved to be problematic. These are Russian-specific phenomena from the part of the dataset generated using corpus data. Many of these phenomena involve morphophonological variation and the constraints on analytical forms. One of the most important findings is that models from different origins converge on exactly the same types of errors. This may not be a coincidence and could reveal insights into differences in linguistic competence of LLMs and human Russian speakers.

To conclude, RuParam is a novel, carefully designed source of data on the Russian language. We hope that it will be useful for further investigation into LLMs' grammar, as well as for model development.

Limitations

- For evaluation we adopt Large Language Models, including foundational models, which constantly undergo modifications. Hence, later evaluations may differ from the results presented in the paper.
- While we present several findings that shed light on the common patterns in the linguistic competence of LLMs, further analysis is required to explain the reasons behind them. For instance, we assume that tokenization may affect the complexity of allomorphy. In addition, the analysis may be enriched by the data from other languages apart from Russian.

Acknowledgements

Lada Pasko and Regina Nasyrova prepared the data and conducted the experiment. This was supported by the MSU Program of Development, Project No 23-SCH02-10 "Linguistic Competence of Natural Language Speakers and Neural Network Models".

Pavel Grashchenkov performed the analysis and organized data collection. His study was conducted under the state assignment of Lomonosov Moscow State University.

We would like to thank Maria Kravchuk, Daniil Burmistrov, Alexandra Anisimova, Anatoly Sergunov, Varvara Slobozhaninova, and Alexandra Trofimova – students at the Department of Theoretical and Applied Linguistics, Lomonosov Moscow State University – for their significant assistance in annotation. We are also grateful to Kseniia Studenikina for her collaborative efforts on the first release of the RuParam dataset.

References

1. Adeeba, F., Dillon, B., Sajjad, H., Bhatt, R.: UrBLiMP: A Benchmark for Evaluating the Linguistic Competence of Large Language Models in Urdu (2025), https://arxiv.org/abs/2508.01006

2. Başar, E., Padovani, F., Jumelet, J., Bisazza, A.: TurBLiMP: A Turkish Benchmark of Linguistic Minimal Pairs. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. p. 16506–16521. Association for Computational Linguistics, Suzhou, China (2025). https://doi.org/10.34810/data1393

3. Bel, N., Punsola, M., Ruiz-Fernández, V.: CatCoLA, Catalan Corpus of Linguistic Acceptability. Procesamiento del Lenguaje Natural 73, 177–190 (2024). https://doi.org/10.34810/data1393

4. Bel, N., Punsola, M., Ruiz-Fernández, V.: EsCoLA: Spanish corpus of Linguistic Acceptability. In: Calzolari, N., Kan, M.Y., Hoste, V., *et al.* (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 6268–6277. ELRA and ICCL, Torino, Italia (May 2024). https://doi.org/10.34810/data1138

5. Daultani, V., Martínez, H.J.V., Okazaki, N.: Acceptability Evaluation of Naturally Written Sentences. Journal of Information Processing 32, 652–666 (2024). https://doi.org/10.2197/ipsjjip.32.652

6. Featherston, S.: Response Methods in Acceptability Experiments, p. 39–61. Cambridge Handbooks in Language and Linguistics, Cambridge University Press (2021). https://doi.org/10.1017/9781108569620

7. Grashchenkov, P.: RuConst: A Treebank for Russian. Lomonosov Philology Journal. Series 9. Philology 3, 94–112 (2024). https://doi.org/10.55959/MSU0130-0075-9-2024-47-03-7, (in Russian)

8. Grashchenkov, P., Pasko, L., Studenikina, K., Tikhomirov, M.: Russian parametric corpus RuParam. Scientific and Technical Journal of Information Technologies, Mechanics and Optics 24(6), 991–998 (2024). https://doi.org/10.17586/2226-1494-2024-24-6-991-998, (in Russian)

9. Hendrycks, D., Burns, C., Basart, S., *et al.*: Measuring Massive Multitask Language Understanding. Proceedings of the International Conference on Learning Representations (ICLR) pp. 11260–11285 (2021). https://doi.org/10.18653/v1/2024.findings-acl.671

10. Hu, H., Zhang, Z., Huang, W., *et al.*: Revisiting acceptability judgements (05 2023). https://doi.org/10.48550/arXiv.2305.14091

11. Jentoft, M., Samuel, D.: NoCoLA: The Norwegian Corpus of Linguistic Acceptability. In: Alumäe, T., Fishel, M. (eds.) Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa). pp. 610–617. University of Tartu Library, Tórshavn, Faroe Islands (May 2023), https://aclanthology.org/2023.nodalida-1.60/

12. Jumelet, J., Weissweiler, L., Bisazza, A.: MultiBLiMP 1.0: A Massively Multilingual Benchmark of Linguistic Minimal Pairs (04 2025). https://doi.org/10.48550/arXiv.2504.02768

13. Ligeti-Nagy, N., Ferenczi, G., Héja, E., *et al.*: HuLU: Hungarian Language Understanding Benchmark Kit. In: Calzolari, N., Kan, M.Y., Hoste, V., *et al.* (eds.) Proceedings of the

2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 8360–8371. ELRA and ICCL, Torino, Italia (May 2024), https://aclanthology.org/2024.lrec-main.733/

14. Mikhailov, V., Shamardina, T., Ryabinin, M., *et al.*: RuCoLA: Russian Corpus of Linguistic Acceptability. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. p. 5207–5227. Association for Computational Linguistics (2022). https://doi.org/10.18653/v1/2022.emnlp-main.348

15. Savchuk, S.O., Arkhangelskiy, T., Bonch-Osmolovskaya, A.A., *et al.*: Russian National Corpus 2.0: New opportunities and development prospects. Voprosy Jazykoznanija 2, 7–34 (2024). https://doi.org/10.31857/0373-658X.2024.2.7-34, (in Russian)

16. Someya, T., Oseki, Y.: JBLiMP: Japanese Benchmark of Linguistic Minimal Pairs. In: Vlachos, A., Augenstein, I. (eds.) Findings of the Association for Computational Linguistics: EACL 2023. pp. 1581–1594. Association for Computational Linguistics, Dubrovnik, Croatia (May 2023). https://doi.org/10.18653/v1/2023.findings-eacl.117

17. Someya, T., Sugimoto, Y., Oseki, Y.: JCoLA: Japanese Corpus of Linguistic Acceptability. In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 9477–9488. ELRA and ICCL, Torino, Italia (May 2024), https://aclanthology.org/2024.lrec-main.828/

18. Song, S., Hu, J., Mahowald, K.: Language Models Fail to Introspect About Their Knowledge of Language (2025), https://arxiv.org/abs/2503.07513

19. Song, Y., Krishna, K., Bhatt, R., Iyyer, M.: SLING: Sino Linguistic Evaluation of Large Language Models. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 4606–4634. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022). https://doi.org/10.18653/v1/2022.emnlp-main.305

20. Suijkerbuijk, M., Prins, Z., Kloots, M.d.H., *et al.*: BLiMP-NL: A Corpus of Dutch Minimal Pairs and Acceptability Judgments for Language Model Evaluation. Computational Linguistics. P. 1–35 (05 2025). https://doi.org/10.1162/coli_a_00559

21. Taktasheva, E., Bazhukov, M., Koncha, K., *et al.*: RuBLiMP: Russian Benchmark of Linguistic Minimal Pairs. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 9268–9299. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.emnlp-main.522

22. Team, Q.: Qwen2.5: A party of foundation models (September 2024), https://qwenlm.github.io/blog/qwen2.5/

23. Tikhomirov, M., Chernyshev, D.: Impact of Tokenization on LLaMa Russian Adaptation. In: 2023 Ivannikov Ispras Open Conference (ISPRAS). pp. 163–168 (2023). https://doi.org/10.1109/ISPRAS60948.2023.10508177

24. Tikhomirov, M., Chernyshev, D.: Facilitating large language model Russian adaptation with Learned Embedding Propagation. Journal of Language and Education 10(4), 130–145 (2024). https://doi.org/10.17323/jle.2024.22224

25. Trotta, D., Guarasci, R., Leonardelli, E., Tonelli, S.: Monolingual and Cross-Lingual Acceptability Judgments with the Italian CoLA corpus. In: Findings of the Association for Computational Linguistics: EMNLP 2021. p. 2929–2940. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.findings-emnlp.250

26. Vázquez Martínez, H.J., Heuser, A., Yang, C., Kodner, J.: Evaluating Neural Language Models as Cognitive Models of Language Acquisition. In: Hupkes, D., Dankers, V., Batsuren, K., *et al.* (eds.) Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP. pp. 48–64. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.genbench-1.4

27. Volodina, E., Mohammed, Y.A., Klezl, J.: DaLAJ - a dataset for linguistic acceptability judgments for Swedish. In: Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning. p. 28–37. LiU Electronic Press (2021), https://aclanthology.org/2021.nlp4call-1.3/

28. Warstadt, A., Parrish, A., Liu, H., *et al.*: BLiMP: The Benchmark of Linguistic Minimal Pairs for English. Transactions of the Association for Computational Linguistics 8, 377–392 (07 2020). https://doi.org/10.1162/tacl_a_00321

29. Warstadt, A., Singh, A., Bowman, S.R.: Neural Network Acceptability Judgments. Transactions of the Association for Computational Linguistics 7, 625–641 (09 2019). https://doi.org/10.1162/tacl_a_00290

30. Xiang, B., Yang, C., Li, Y., *et al.*: CLiMP: A Benchmark for Chinese Language Model Evaluation. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 2784–2790. Association for Computational Linguistics, Online (Apr 2021). https://doi.org/10.18653/v1/2021.eacl-main.242

31. Zheng, L., Chiang, W.L., Sheng, Y., *et al.*: Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 46595–46623. Curran Associates, Inc. (2023). https://doi.org/10.5555/3666122.3668142