Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2025”

April 23–25, 2025

Gradual Acceptability Judgments with LLMs: Evidence from Agreement Variation in Russian

Kseniia Studenikina
Lomonosov Moscow State University,
Moscow, Russia
xeanst@gmail.com

Ekaterina Lyutikova
Lomonosov Moscow State University,
Moscow, Russia
lyutikova2008@gmail.com

Anastasia Gerasimova
Lomonosov Moscow State University
Moscow, Russia
anastasiagerasimova432@gmail.com

Abstract

This study examines the LLMs’ performance on gradual acceptability judgments task. Previously, the linguistic competence of LLMs was evaluated using binary acceptability scales, which contradicts the theoretical concept of acceptability. We present a new benchmark KVaS (***K**orpus **Va**riativnogo **S**oglasovanija* ‘Corpus of Variable Agreement’) derived from syntactic experiments on variable agreement in Russian. Our dataset contains multiple phenomena of agreement variation, ideal for modeling diverse acceptability levels, and compiles 7013 sentences rated by native speakers on a 1–7 Likert scale. We evaluated two LLMs, mainly Russian-trained GigaChat-Pro and multilingual Mistral Large, comparing their capability to treat acceptability as a scale to the reference human scores from KVaS. We used prompting providing benchmark sentences in two modes: zero-shot mode included only instructions while a few-shot mode added training sentences and their scores. The results show that GigaChat-Pro underperformed compared to Mistral Large. GigaChat-Pro improved significantly in a few-shot mode while Mistral Large exhibited more stable behavior. The case study shows that Mistral can detect nearly all significant contrasts in an experiment, whereas GigaChat performed near-randomly. The corpus may be useful for ranking LLMs, fine-tuning, and enhancing Russian text generation quality.

Keywords: linguistic competence, syntax, agreement, LLM, benchmark
DOI: 10.28995/2075-7182-2025-23-XX-XX

Градуальная оценка приемлемости с помощью больших языковых моделей: данные русского вариативного согласования

Студеникина К. А.
МГУ им. М. В. Ломоносова,
Москва, Россия
xeanst@gmail.com

Лютикова Е. А.
МГУ им. М. В. Ломоносова,
Москва, Россия
lyutikova2008@gmail.com

Герасимова А. А.
МГУ им. М. В. Ломоносова,
Москва, Россия
anastasiagerasimova432@gmail.com

Аннотация

В данном исследовании рассматривается способность больших языковых моделей (БЯМ) оценивать приемлемость предложений по градуальной шкале. Ранее способность БЯМ оценивать приемлемость языковых выражений выявлялась в бинарных задачах. Мы представляем новый бенчмарк, данными для которого послужили результаты синтаксических экспериментов, посвященных вариативному согласованию в русском языке.

Датасет включает 7013 предложений с оценками по шкале от 1 до 7, в которых представлены феномены вариативного согласования, соответствующие различным уровням приемлемости. С использованием датасета мы провели тестирование двух БЯМ на задаче градуальной оценки приемлемости в двух режимах диалога. Режим zero-shot включал только инструкцию, в режиме few-shot были добавлены тренировочные предложения и их оценки. Результаты показывают, что результат работы модели GigaChat-Pro, обучавшейся преимущественно на русскоязычных данных, зависит от режима тестирования: качество повышается в режиме few-shot. Качество мультиязычной модели Mistral Large выше и не зависит от типа инструкции. Mistral выявляет почти все значимые контрасты в одном из рассмотренных экспериментов, тогда как ответы модели GigaChat близки к случайным. Представленный корпус может быть использован не только для ранжирования БЯМ, но и для дообучения и улучшения качества русскоязычной генерации.

Ключевые слова: языковая способность, синтаксис, согласование, большие языковые модели, бенчмарк

1 Introduction

Modern large language models (LLMs) exhibit near-human proficiency in natural language processing. They are effectively used to understand and generate texts during interactions with user. The performance of LLMs is assessed through benchmarks that primarily focus on evaluating the semantic and pragmatic capabilities of an LLM, e.g. possessing information correctly, building causal relationships. For example, the SuperGLUE benchmark (Wang et al. 2019) includes tasks such as question answering and recognizing textual entailment, while the MMLU benchmark (Hendrycks et al. 2020) contains multiple choice questions from 57 subject areas.

Besides semantic adequacy, LLMs' linguistic behavior should be approximated to human behaviour based on formal linguistic criteria. Native speakers can not only produce correct linguistic expressions but also distinguish them from incorrect ones. Ideally, similar competencies are expected from LLMs. Prior assessments of LLM linguistic competence were limited to binary modes, such as the CoLA benchmark's binary classification of acceptable vs. unacceptable sentences (Warstadt et al. 2019) or the BLiMP's selection of the more acceptable sentence from a minimal pair (Warstadt et al. 2020). Meanwhile, theoretical linguistics views sentence acceptability as a gradual concept involving two aspects of a linguistic expression: (i) its grammatical well-formedness, which entails adherence to the grammatical constraints; (ii) its processing, which is subject to usage factors like word frequency, syntactic complexity, etc. (Chomsky 1965; see Sprouse 2007; Schütze, Sprouse 2014; Lau, Clark, Lappin 2017 a.o.). The existence of marginal acceptability ratings suggests that treating acceptability as a scale is a key feature of language competence.

This study aims to determine whether LLMs can assess sentence acceptability on a gradual scale. To achieve this goal, we make use of a well-studied phenomenon of Russian morphosyntax — the variable agreement. Agreement rules, being rather straightforward and unequivocal in standard cases, allow for multiple alternative strategies with non-canonical controllers, such as numeral or coordinated phrase. This phenomenon enables us to analyse the contribution of various factors to the choice of a specific agreement strategy and compare their impact in both language models and human speakers.

The rest of the paper is organized as follows. Section 2 presents our benchmark. Section 3 discusses the experiment with LLMs and evaluates its results. Section 4 concludes.

2 Corpus of Variable Agreement in Russian

We created a new benchmark called KVaS (***K**orpus **Va**riativnogo **S**oglasovanija* 'Corpus of Variable Agreement') designed to test language models on the gradual acceptability judgment task in Russian. In standard agreement, the target's grammatical form is unambiguously determined by the controller features (person, number and gender), which makes it possible to distinguish grammatical and ungrammatical variants (1). Variable agreement occurs if the calculation of the target features becomes ambiguous, which happens in case of several potential controllers (constructions with a postpositive relative clauses (2), coordinated subjects (3)) or within a non-canonical controller (quantitative noun phrases (4), quantified subjects (5)).

(1) *Marina* ***poliva-et*** / ****poliva-em*** *cvety* *po subbotam.*
Marina water-3SG / water-1PL flowers on Saturdays
'On Saturdays Marina waters the flowers.'

(2) *Vse, kto* ***zajdet/*** ***zajdut*** *v* *buhgalteriyu,*
all who come-3SG come-3PL in accounting_department
poluchat *zarplatu.*
receive-3PL salary
'Everyone who enters the accounting department will receive a salary.'

(3) *Petya i ya* ***id-em /*** ***?id-u /*** ***?id-ut /*** ****id-et*** *domoj.*
Petya and I go-1PL go -1SG go-3PL go-3SG home
'Petya and I are going home.'

(4) *Na l'du* ***svalil-as'/*** ***?svalil-is' /*** ****svalil-os'*** *ujma lyud-ej.*
On ice fell-SG.F fell-PL fell-SG.N heap people-GEN
'A lot of people fell on the ice.'

(5) *Dvoe iz nas* ***pridut /*** *?**pridet /*** ****pridem*** *v* *gosti.*
Two from us.GEN come-3PL come-3SG come-1PL in guests
'Two of us are coming to visit.'

Data for the corpus was gathered from the experimental studies conducted during the practical course on experimental syntax[1] in 2022-2024 published in the Database of Agreement Variation[2]. Each study employed a factorial design, wherein two or more independent variables (factors) were manipulated to examine their effect on a dependent variable, with one factor always corresponding to agreement pattern and the dependent variable being acceptability rating. Other factors in the experiments described different features of sentences that might influence the choice of agreement, e.g. order, semantics of the controller and target. By combining the factors, researchers investigated interactions between them, i.e. whether acceptability of an agreement pattern changes depending on the grammatical environment.

All the source studies applied the method of 1–7 Likert scale acceptability judgment task, which is a standard tool in the field of experimental syntax providing a quantifiable way to compare acceptability of different sentence structures. Unlike simpler methods, such as using asterisks (*) or question marks (?) commonly employed in generative syntax to indicate degrees of grammaticality, the Likert scale offers a more nuanced and systematic approach. An uneven number of points on the scale allows for non-binary judgments, which is especially important for detecting fine-grained effects of syntactic manipulations. To reduce the impact of lexical content on acceptability, stimuli were arranged in lists such that each sentence (lexicalization) appeared in each list under one condition (a specific combination of factor values). An example of lexicalization from the experiment on constructions with a postpositive relative clause is given in (2) and (6). The experiment examined agreement variation that emerges if there is a conflict in the number feature between the head (*vse / vse sotrudniki* 'all.PL / all.PL employees') and the relative pronoun (*kto* 'who.SG') and comprised two factors: agreement pattern and head structure.

(6) *Vse sotrudniki, kto* ***zajdet/*** ***zajdut*** *v* *buhgalteriyu,*
all employees who come-3SG come-3PL in accounting_department
poluchat *zarplatu.*
receive-3PL salary
'All employees who enters the accounting department will receive a salary.'

Alongside the stimuli, each experiment included filler sentences without any factors: grammatical ones represented fully correct sentences, while ungrammatical ones contained agreement errors. The two types of fillers are used as the threshold of (un)grammaticality when analysing ratings in experiments (Gerasimova 2023). Fillers can also serve as a baseline to demonstrate the model's ability to score non-variable agreement.

[1] The course is taught by Dr. Anastasia Gerasimova and Prof. Dr. Habil. Ekaterina Lyutikova at the Department of Theoretical and Applied Linguistics, Philological Faculty, Lomonosov Moscow State University. See more information at the website of Moscow Experimental Syntax Group, URL: https://expsynt.com/.
[2] URL: https://expsynt.com/table.html/.

The benchmark KVaS is publicly available in the GitHub repository[3]. It contains 7013 sentences from 15 experimental studies (from 66 to 137 participants each, mean number is 96, standard deviation is 20). For all experiments, outliers were excluded based on criteria from (Gerasimova 2021), from 5 to 15 participants in each study. The distribution of contexts in the corpus is shown in Figure 1. The apparent disparity in the number of examples stems from the fact that different contexts may exhibit varying number of relevant factors influencing agreement.

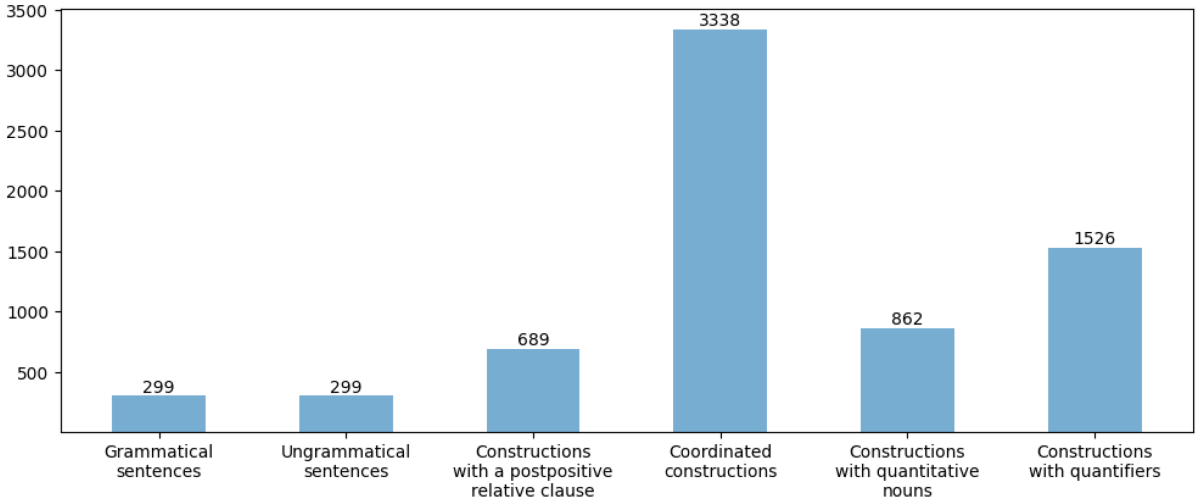

Figure 1. Distribution of contexts in the benchmark KvaS

The dataset includes the following information: the experiment code (author and year), the sentence, the average score for all participants, the type of construction / its grammaticality, the presence of variable agreement, the lexicalization number. A fragment of the corpus is shown in Table 1.

Experiment	**Sentence**	**response**	**subtype**	**type**	**lexicalization**
Davidjuk_2023	*Ya i Maksim progulyaet poslednij urok.* ‘Me and Maxim are skipping the last lesson.’	2.56	coordination	stimulus	32
Belova_2023	*Shestero iz nas spravyatsya s zamenoj lampochek.* ‘Six of us can handle replacing the light bulbs.’	5.27	quantifier	stimulus	11
Krainova_2023	*Eti million marok hranilis' v al'bomah.* ‘These million stamps were stored in albums.’	3.88	quantitative_ nouns	stimulus	28
Dorofeeva_2023	*Kazhduyu mat', kto yavilas' na sobranie roditelej, rassprashivaet klassnyj rukovoditel'.* ‘Every mother who attends the parents' meeting is questioned by the homeroom teacher.’	3.43	relative_ clauses	stimulus	34
Davidjuk_2023	*I Maksim, i ya progulyayut poslednij urok.* ‘Both Maxim and I will skip the last lesson.’	2.70	coordination	stimulus	32
Pasko_2024	*Na aukcione tret' skul'ptur pokupaesh' kollekcioner.* ‘A collector buys a third of the sculptures at auction.’	1.81	bad	filler	505
Danilova_2023	*Klient poprosil vegetarianskoe ili postnoe menyu.* ‘The client asked for a vegetarian or lean menu.’	6.03	good	filler	51

Table 1. Fragment of the benchmark KvaS

Additionally, each experiment has its own dedicated dataset with the markup from an experimental study. Table 2 shows such a dataset for T. Davidyuk’ experiment (2024). This information complements the main corpus by specifying the type of conjuncts (*I* + proper name / proper name + *I*), the agreement

[3] URL: https://github.com/Xeanst/KVaS.

strategy (1st person singular / 3rd person singular / 1st person plural / 3rd person plural), the type of conjunction (*and*, *and..and*, *or*, *or..or*) and the word order (subject + predicate / predicate + subject).

sentence	**response**	**subtype**	**type**	**conjuncts**	**agreement**	**conjunction**	**order**
I ya, i Grisha sygraem shahmatnuyu partiyu. 'Grisha and I will play a chess game.'	4.50	coordination	stimul	I_noun	1pl	and_and	SV
Vlad i ya pokazhu korotkuyu dorogu. 'Vlad and I will show you a shortcut.'	2.44	coordination	stimul	noun_I	1sg	and	SV
Dima ili ya poslushayut eto golosovoe soobshchenie. 'Dima or I will listen to this voice message.'	4.88	coordination	stimul	noun_I	3pl	or	SV
Ili Mitya, ili ya sostavit etot dlinnyj spisok. 'Either Mitya or I make this long list.'	2.20	coordination	stimul	noun_I	3sg	or_or	SV
Lyova oformit etot delovoj dogovor. 'Leva will formalize this business agreement.'	6.44	good	filler	NaN	NaN	NaN	NaN

Table 2. Fragment of grammatical markup for the experiment by T. Davidyuk (2024)

The KvaS benchmark has both strengths and weaknesses compared to other acceptability judgments benchmarks. On the one hand, the existing corpora like CoLA and BLiMP cover a broader array of grammatical phenomena than KVaS, enabling the identification of which phenomena pose greater challenges for LLMs. On the other hand, they only allow for binary evaluations of language competence, which may seem overly simplistic and insufficiently nuanced. KVaS stands out as the first dataset for gradual acceptability judgments on a scale from 1 to 7. The dataset allows us to test how effectively language models trained on unlabelled texts are able to detect subtle morphosyntactic and semantic differences between sentences. Since each example in the KVaS benchmark was rated by many respondents, this markup is reliable and reflects the linguistic knowledge of Russian native speakers. Thus, KVaS may be used to examine whether LLMs' scoring consistency for different lexicalizations matches that of human evaluators.

Our corpus further enables ranking LLMs based on their level of linguistic competence, specifically in the domain of grammatical variation. In the next section, we will outline the experiment involving gradual acceptability judgments with various LLMs. The Likert-scale scores by LLMs will be compared with the reference human scores from KVaS.

3 Experiment on Gradual Acceptability Judgments with LLMs

This section describes the LLMs' evaluation on gradual acceptability judgments task using KVaS benchmark. We first present the method and the choice of models and then discuss the results.

3.1 Method

Previously, evaluation of LLMs linguistic competence mainly required either prior training on binary acceptability classification as seen with CoLA, or direct comparison of probability metrics as in BLiMP. In our study, we investigate LLMs' grammatical preferences by analysing direct responses to verbal instructions. A similar procedure was used with MMLU data on factual information (Hendrycks et al. 2021): the LLM was presented with multiple-choice questions and tasked with generating the correct answer's number. Since LLMs are specifically trained on instructional datasets, this approach aligns naturally with their capabilities. Our method not only tests the model's ability to predict that an unacceptable sentence is less likely than an acceptable one but also explores the models' concept of acceptability. Essentially, we replicate the experimental method used for human data labelling treating the model the same way.

We aimed to compare models trained primarily on Russian data with multilingual ones. The first models were expected to perform better in the gradual acceptability judgments, consistent with prior observations when contrasting the language competence of LLMs unprepared for Russian data versus those extensively trained on Russian texts (Grashchenkov et al. 2024). In this study, we tested two models: GigaChat-Pro[4] as a mainly "Russian-speaking" one and Mistral Large[5] as a multilingual one. Both were selected as top performers according to MERA benchmark (Fenogenova et al. 2024) at the time of the experiment (fall 2024) regardless of the number of parameters. The model choice was limited by API availability.

The final version of the instruction, yielding optimal results, is given in example (7). Two approaches were tested: (i) zero-shot – the prompt contains only the instructions and the target sentence, (ii) few-shot – in addition, the prompt contains two training examples: sentences and their respective scores. An example of a training pair is given in (8).

(7) *Tebe nuzhno ocenit' predlozhenie po shkale ot 1 do 7. Esli predlozhenie zvuchit horosho, tak mozhno skazat', postav' emu vysokuyu ocenku (6 ili 7). Esli predlozhenie zvuchit ploho, "ne po-russki", tak skazat' nel'zya, postav' emu nizkuyu ocenku (1 ili 2). Nekotorye predlozheniya mogut kazat'sya ne ochen' horoshimi, no v principe dopustimymi. Takim predlozheniyam postav' srednyuyu ocenku (ot 3 do 5).*
Oceni predlozhenie po shkale ot 1 do 7: "{ _}". Otvet' odnoj cifroj, nichego ne dobavlyaya.
'You need to grade the sentence on a scale from 1 to 7. If the sentence sounds good, you can say so, give it a high score (6 or 7). If the sentence sounds bad, "not in Russian", it is impossible to say so, give it a low score (1 or 2). Some sentences may not seem very good, but they are acceptable in fact. Give these sentences an average score (from 3 to 5).
Grade the sentence on a scale from 1 to 7: "{Target sentence}". Answer with one digit, without adding anything.'

(8) *Naprimer, predlozhenie "Pered paroj Yaroslav nadeli i pidzhak, i zhiletku." soderzhit oshibku. Emu stoit postavit' ocenku 1 ili 2. Predlozhenie "V aprele Marina poseyala semena i pomidora, i tykvy." yavlyaetsya vpolne estestvennym. Emu mozhno postavit' ocenku 6 ili 7.*
'For example, the sentence "Before the class, Yaroslav wore$_{PL}$ both a jacket and a vest." contains an error. It should be graded with 1 or 2. The sentence "In April, Marina sowed$_{SG}$ the seeds of both tomatoes and pumpkins." is quite natural. It can be given a score of 6 or 7.'

3.2 Results

Figure 2 shows the average human and LMMs' scores of sentences grouped by context type. Human scores for grammatical sentences fall within the upper range of the scale (6 out of 7), while ungrammatical sentences receive scores in the lower range (2 out of 7). GigaChat scores obtained in the zero-shot mode do not reflect this pattern. However, its scores generated in the few-shot mode are close to human judgments, indicating that the inclusion of training sentences enhances GigaChat performance. Conversely, there is no difference between the modes for Mistral. In both zero-shot and few-shot modes, grammatical fillers obtain high scores while the ungrammatical ones are rated low. The average scores for stimulus sentences – both for human and LLMs – lie in the middle of the scale (3-4 out of 7).

[4] URL: https://developers.sber.ru/docs/ru/gigachat/models.
[5] URL: https://docs.mistral.ai/getting-started/models/models_overview/.

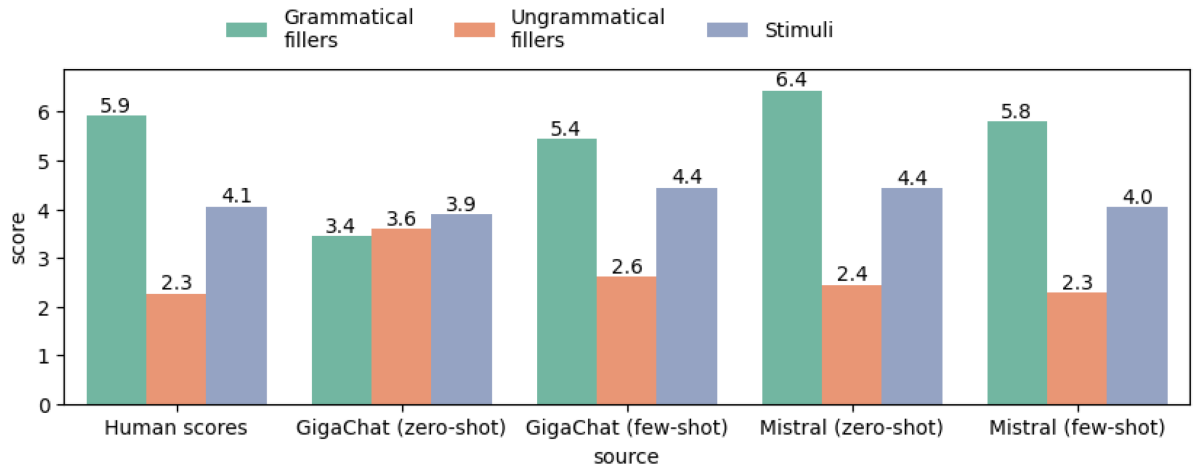

Figure 2. Average acceptability scores for human and LLMs

For a more detailed analysis of stimuli scores, we calculated the mean absolute error (MAE). The results are shown in Figure 3. The highest error for grammatical and ungrammatical fillers is obtained by GigaChat in zero-shot mode. Mistral's few-shot mode exhibits the lowest error for filler sentences. The error on stimuli for GigaChat in few-shot is slightly greater than for Mistral.

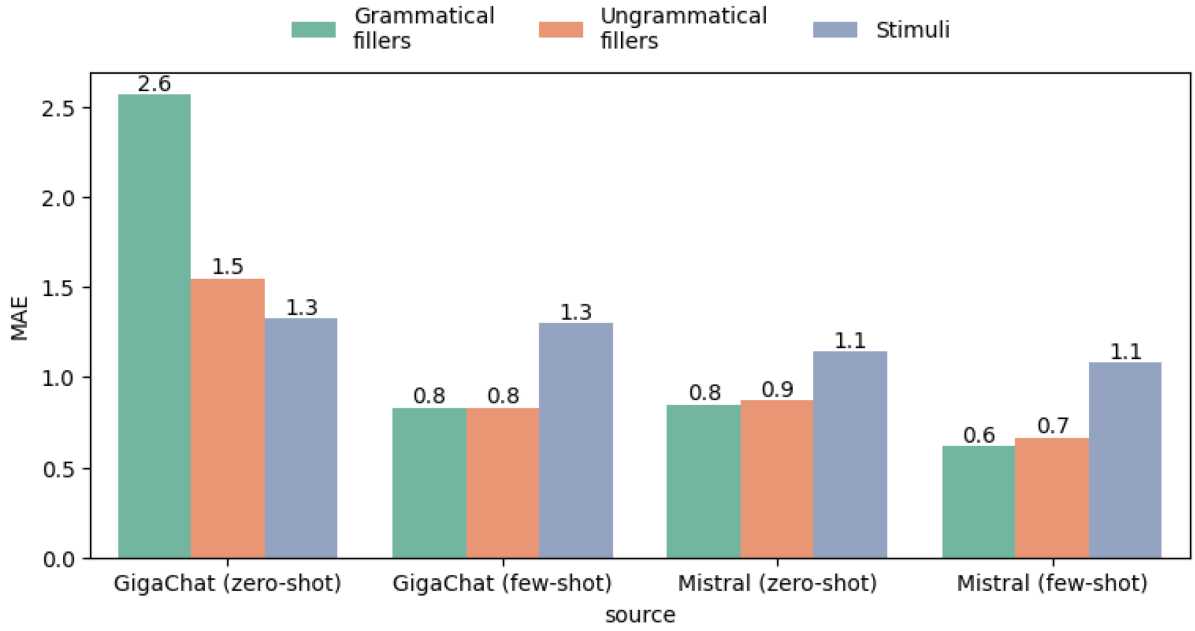

Figure 3. Mean absolute error (MAE)

MAE provides a quantitative estimate of the discrepancy between LLMs' predictions and human scores. However, it fails to indicate whether LLMs can capture relevant contrasts between experimental conditions as effectively as humans. To estimate this, we need to calculate if the conditions that showed significant differences in human experiments also exhibit significant differences in LLM scores, and conversely, if those insignificant for humans remain insignificant for LLMs. The ultimate metric is the percentage of matching pairwise comparisons between human scores and LLMs' responses.

As an illustration, we evaluated an experiment on constructions with a postpositive relative clause conducted by A. Golovnina (2022). The relevant experimental factors are agreement pattern (singular / plural verb in the embedded clause) and presence of the noun within the head (with noun / no noun) which gives four experimental conditions, see examples (2), (6) above. Figure 4 demonstrates mean values for conditions and grammatical and ungrammatical fillers. We employed the Mann-Whitney

U test and the Student's t-test to compute the significance tests[6]. The percentage of matches between human scores and LLM is presented in Table 3. The results show that most significant contrasts are captured by Mistral. The zero-shot mode performs even better than the few-shot one. On the contrary, GigaChat captures only half of the relevant contrasts, though the few-shot mode improves performance. Future research should involve conducting similar evaluations for other experiments in the KVaS.

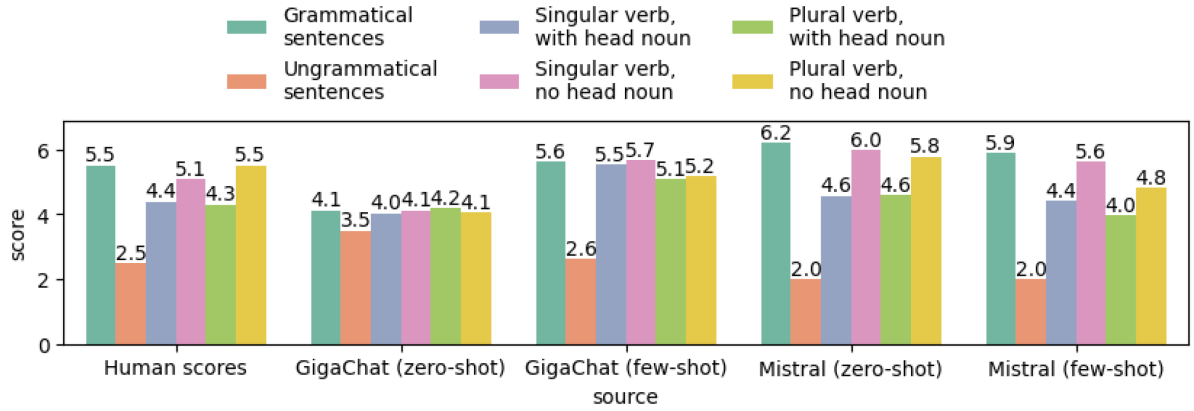

Figure 4. Mean values by condition from A. Golovnina's experiment (2022)

	Percentage of matching significant contrasts calculated with	
	Mann-Whitney U test	Student's t-test
GigaChat (zero-shot)	40%	40%
GigaChat (few-shot)	53.3%	53.3%
Mistral (zero-shot)	86.7%	86.7%
Mistral (few-shot)	73.3%	80%

Table 3. Percentage of matching significant contrasts between human and LLMs' scores

To conclude, GigaChat's performance varies depending on the selected mode, with the few-shot mode yielding superior results. Mistral's results are less dependent on the selected mode. Contrary to our hypothesis, the quantity of Russian data used during training did not affect the model performance. Multilingual Mistral shows higher proficiency than predominantly Russian-trained GigaChat. This suggests that LLMs' linguistic competence may correlate with their common-sense knowledge and rankings based on the tasks in the MERA benchmark (Fenogenova et al. 2024) or that the training data lacked a sufficient amount of variable grammatical phenomena.

4 Conclusion

This paper presents the first benchmark for evaluating LLMs on gradual acceptability judgments. Unlike binary acceptability classification, this task enables a fine-grained assessment of LLMs linguistic competence. The presented dataset comprises the results of syntactic acceptability judgment experiments on agreement variation in Russian. The stimulus sentences featuring agreement variation occupy the middle part of the acceptability scale, making them perfect material for testing gradual acceptability. The fillers presenting standard agreement serve as a baseline, occupying either high or low positions on the scale.

LLMs testing using the benchmark involved prompting. Two models – GigaChat Pro and Mistral Large – have been tested in two modes: zero-shot (instructions only) and few-shot (instructions plus training sentences). The mean absolute error (MAE) was calculated to compare the model predictions against human responses. The results revealed that GigaChat trained mainly on Russian data yielded lower quality compared to multilingual Mistral. GigaChat proved relatively unstable, with performance varying significantly between zero-shot and few-shot modes. Mistral demonstrated lower MAE value and maintained results regardless of the chosen mode. Importantly, Mistral preserved the majority of

[6] We used both parametric and non-parametric since it is unclear whether the data from LLMs is interval or ordinal.

human contrasts in the single experiment we examined, although these are preliminary conclusions as it is necessary to test all the experiments in the corpus. Our study indicates that the level of LLM linguistic competence as measured by grammatical phenomena of variation does not correlate with the amount of Russian data during training.

Acknowledgements

The work of Ekaterina A. Lyutikova and Anastasia A. Gerasimova, who collected the corpus and developed the metrics, was carried out with the support of MSU Program of Development, Project No 23-SCH02-10 “Linguistic Competence of Natural Language Speakers and Neural Network Models”. The work of Kseniia A. Studenikina, who conducted models’ evaluation, was supported by Non-commercial Foundation for Support of Science and Education “INTELLECT”.

References

[1] Chomsky N. Aspects of the theory of syntax. — Cambridge, MA : MIT Press, 1965.
[2] Davidjuk T. Agreement with disjoined subjects in Russian. — Argumentum, 2024. — Vol. 20. — P. 322–335.
[3] Fenogenova A., Chervyakov A., Martynov N., Kozlova A., Tikhonova M., Akhmetgareeva A., Emelyanov A., Shevelev D., Lebedev P., Sinev L., Isaeva U., Kolomeytseva K., Moskovskiy D., Goncharova E., Savushkin N., Mikhailova P., Minaeva A., Dimitrov D., Panchenko A., Markov S. MERA: A Comprehensive LLM Evaluation in Russian // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. — Bangkok, Thailand, 2024. — P. 9920–9948.
[4] Gerasimova A.A. (2021), The training materials of the workshop on experimental syntax. Selection of respondents. [Uchebnye materialy praktikuma po eksperimental'nomu sintaksisu. Otbor respondentov]. URL: https://agerasimova.com/wp-content/uploads/Gerasimova_Practice_Outliers.pdf
[5] Gerasimova A.A. (2023), Quantitative Methods of Investigating Grammar (A Case Study of Agreement Variation in Russian). PhD Thesis. Lomonosov Moscow State University. Moscow, 2023.
[6] Grashchenkov P.V., Pasko L.I., Studenikina K.A., Tikhomirov M.M. (2024). Russian parametric corpus RuParam [Parametricheskij korpus russkogo yazyka RuParam], Scientific and Technical Journal of Information Technologies, Mechanics and Optics [Nauchno-tekhnicheskij vestnik informacionnyh tekhnologij, mekhaniki i optiki.], Vol. 24, № 6, pp. 991–998.
[7] Golovnina A. S. (2022). Predikativnoje soglasovanije v otnositel’nyh klauzah s sojuznym slovom kto i tip mestoimennoj veršiny [Predicate agreement in relative clauses with the relative pronoun kto and the pronominal head type]. The student conference “Experimental studies of language”, Moscow, Russia.
[8] Hendrycks D., Burns C., Basart S., Zou A., Mazeika M., Song D., Steinhardt J. Measuring Massive Multitask Language Understanding // Proceedings of the International Conference on Learning Representations (ICLR), — Virtual Event, Austria, 2020.
[9] Lau J.H., Clark A., Lappin S. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge // Cognitive Science. — 2017. — Vol. 41. — No. 5. — P. 1201–1241.
[10] Likert R. A technique for the measurement of attitudes. — Archives of Psychology, 1932. — Vol. 22, № 140). — P. 5–55.
[11] Schütze C., Sprouse J. Judgment data // Research methods in linguistics / editors : D. Sharma, R. Podesva. – Cambridge : Cambridge University Press, 2014. – P. 27–50.
[12] Sprouse J. Continuous acceptability, categorical grammaticality, and experimental syntax // Biolinguistics. – 2007. – Vol. 1. – P. 123–134.
[13] Sprouse J., Schütze C. T., Almeida D. A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010. — Lingua, 2013. — Vol. 134. — P. 219–248.
[14] Wang A., Pruksachatkun Y., Nangia N., Singh A., Michael J., Hill F., Levy O., Bowman S. R. SuperGLUE: a stickier benchmark for general-purpose language understanding systems // Advances in Neural Information Processing Systems (NeurIPS). — 2019.
[15] Warstadt A., Singh A., Bowman S. R. Neural Network Acceptability Judgments. — Transactions of the Association for Computational Linguistics, 2019. — Vol. 7. — P. 625–641.
[16] Warstadt A., Parrish A., Liu H., Mohananey A., Peng W., Wang S. F., Bowman S. R. BLiMP: The benchmark of linguistic minimal pairs for English — Transactions of the Association for Computational Linguistics, 2020. — Vol. 8. — P. 377–392.