Emotional Lexicons: How Large Language Models Predict Emotional Ratings of Russian Words

Polina V. Iaroshenko and Natalia V. Loukachevitch
Research Computing Center
Lomonosov Moscow State University
Moscow, Russian Federation
polina.iaroshenko@yandex.ru and louk_nat@mail.ru

Abstract

This study examines the capability of LLMs to predict emotional ratings of Russian words by comparing their assessments with both native speakers' ratings and expert evaluations. The research utilises two datasets: the ENRuN database containing associative emotional ratings of Russian nouns by native speakers, and RusEmoLex, an expert-compiled lexicon. Various open-source LLMs were evaluated, including international models (Llama-3, Qwen 2.5), Russian-developed models, and Russian-adapted variants, representing three parameter scales. The findings reveal distinct patterns in model performance: Russian-adapted models demonstrated superior alignment with native speakers' ratings, whilst model size was not a decisive factor. Conversely, larger models showed better performance in matching expert assessments, with language adaptation having minimal impact. Emotional or sensitive lexis with strong connotations produce a more substantial human-model gap.

1 Introduction

Recent advances in Large Language Models (LLMs) have demonstrated remarkable capabilities in various natural language processing (NLP) tasks, including emotion analysis. Previous research on emotions expressed in text has primarily focused on emotion detection and categorization (Acheampong et al., 2020; Kazyulina et al., 2021; Bostan and Klinger, 2024; Cavicchio, 2025). However, the widespread adoption of chatbots and advancement of LLMs have necessitated not only the recognition and classification of emotions but also the authentic simulation of emotional responses during human-computer interactions. Consequently, the study of LLMs' emotional intelligence has become increasingly relevant (see, for example, Raj, 2024; Chen et al., 2024; Huang et al., 2024; Dalal et al., 2025).

It is important to note that in emotion analysis, as in many other areas of NLP, research conducted on English language material continues to dominate the field (for a detailed discussion, see De Bruyne, 2023). Cross-cultural variations in emotional expression raise concerns about LLMs' ability to adapt to different linguistic contexts and cultural value systems.

One approach to studying LLMs' emotional behaviour and comparing it with human responses is to examine emotional ratings of individual lexical items. Emotional word ratings can be categorised into two distinct methodologies: associative and expert-based evaluations. Associative emotional assessment refers to survey results, in which respondents evaluate the degree to which lexical units are associated with specific emotions. These experiments can employ any words as stimuli, regardless of their direct connection to emotional domains (see, for example, Mohammad and Turney, 2013; Domingues Aparecido et al., 2025). Expert-based emotional assessment, conversely, relies on specialized linguistic resources such as dictionaries or thesauri dedicated to specific semantic categories – in this case, emotional lexis. These resources, developed specifically to compile emotional lexicons, exclusively contain words that experts have identified as expressing or indicating emotions (for a detailed discussion, see Mohammad, 2023).

This study aims to compare human emotional ratings of Russian words with those generated by LLMs. The research examines both native speakers' associative emotional assessments and expert evaluations. For associative ratings, we utilize the ENRuN (Emotional Norms for Russian Nouns) database (Sysoeva and Lyusin, 2024), comprising ratings for 1,800 Russian nouns. Expert evaluations are derived from the Russian Emotion Lexicon (RusEmoLex) (Iaroshenko and Loukachevitch, 2025b), which provides emotional categorization for 1,024 words. RusEmoLex's lexical inclusion criteria primarily rely on specialized linguistic resources (dictionaries, thesauri, and corpus data).

The Proceedings for the 15th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA 2026), pages 96–106

The study employs open-source models of varying scales (three categories: 7-8B, 32B, and 70B parameters) and diverse origins: international models (Llama-3, Qwen 2.5), Russian-developed models, and Russian-adapted variants.

We aim to answer the following research questions (RQs): **RQ1**: How accurately do open-source LLMs predict emotional word ratings in Russian when compared to native speakers' and expert assessments? **RQ2**: To what extent does model size affect the quality of emotional assessment? **RQ3**: Do Russian-adapted models demonstrate superior performance in emotional rating predictions compared to their non-adapted counterparts?

2 Related Work

During the last few years, research on LLM empathy and emotional intelligence has expanded, employing advanced benchmarks to evaluate models' emotional responses across varied scenarios.

EmotionBench (Huang et al., 2024) introduces a theoretically grounded methodology derived from psychological emotion appraisal theory (Roseman and Smith, 2001). The researchers identified distinct situational factors that elicit specific emotional responses. The comparative analysis between LLM responses and human assessments revealed that while LLMs can provide appropriate responses, they demonstrate significant limitations in aligning with human emotional behaviour.

The authors of (Sabour et al., 2024) introduced EmoBench, a comprehensive evaluation framework in English and Chinese. The benchmark is structured around two primary assessment areas: Emotional Understanding and Application. In the Understanding component, LLMs must identify emotions and their underlying causes, with particular emphasis on complex, ambiguous situations. The Application component evaluates models' ability to select appropriate responses to given scenarios. The findings revealed that human participants outperformed LLMs across both tasks, albeit by a modest margin.

A notable contribution to understanding LLMs' emotional capabilities comes from research employing standardized psychological assessment tools. The study by (Vzorin et al., 2024) assessed LLM emotional intelligence through the Russian version of the Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT). (Dalal et al., 2025) employed established psychological instruments: the Trait Emotional Intelligence Questionnaire (TEIQue) (Petrides, 2009) and the Situational Evaluation of Complex Emotional Understanding (SECEU) (MacCann and Roberts, 2008).

While recent research has made substantial progress in evaluating LLMs' emotional intelligence capabilities, most studies have focused on the English language. The need to investigate emotional intelligence across different linguistic and cultural contexts remains a crucial research direction.

3 Methodology

The primary objective of this study is to evaluate LLMs' ability to assess Russian words for their emotional associations and to compare these assessments with human ratings. With this aim, we use the ENRuN database, based on native speaker scores, and the Russian Emotion Lexicon, constructed by professional linguists.

Data. The ENRuN database (Sysoeva and Lyusin, 2024)[1] provides emotional ratings for 1,800 Russian nouns, evaluated through both dimensional (valence and arousal) and categorical (happiness, sadness, anger, fear, and disgust) approaches. For the present analysis, we utilize averaged categorical ratings, where respondents assessed word-emotion associations on a five-point scale.

RusEmoLex (Iaroshenko and Loukachevitch, 2025b) comprises 1,024 Russian lexical items representing various parts of speech. Each entry is annotated with one of five emotional categories: happiness, sadness, anger, fear, and surprise. RusEmoLex was developed by integrating various Russian emotional vocabulary resources, with expert linguistic sources serving as the primary foundation: dictionaries and thesauri (Shvedova, 1998; Babenko, 2022), and data from the semantically annotated section of the Russian National Corpus (Savchuk et al., 2024).

To enable comparison between RusEmoLex and ENRuN datasets, RusEmoLex entries were converted to numerical ratings: words labeled with a specific emotion (e.g., happiness) were assigned a maximum score (5) for that emotion and zero scores for all other emotional categories.

[1]The current version of the ENRuN database can be provided to researchers upon request. At the time of this study, the database included responses from 692 participants, with data collection continuing as an ongoing process.

Models. We employ open-source LLMs ranging from 7B to 70B parameters, representing three categories by origin: international models (Qwen 2.5-32B and 7B, Llama-3), Russian-developed models (YandexGPT 5 Lite), and models adapted for Russian through enhanced training and tokenization (T-lite-it-1.0, T-pro-it-1.0, RuadaptQwen2.5-7B, RuadaptQwen2.5-32B-Pro-Beta). A complete list of models, along with their references are provided in Appendix A.

Hyperparameters. The experiments employ the following hyperparameters: temperature was set to 0.2 to ensure relatively consistent outputs while maintaining some degree of creativity; a repetition penalty of 1.2 was implemented to discourage uniform ratings across different words. For a detailed discussion of hyperparameter effects on LLM responses in similar rating tasks, see (Iaroshenko and Loukachevitch, 2025a).

Prompts. For the experiments, two types of prompts were employed. The SPEAKER prompt, based on the instructions given to participants in the ENRuN database evaluation experiment (Lyusin and Sysoeva, 2025), was used to assess emotional words from a native speaker's perspective. The EXPERT prompt was designed to evaluate words from an expert's viewpoint. Several versions of this prompt were tested, including one with an additional "Background Information" section that provided a theoretical definition of emotional vocabulary and examples of words belonging to this semantic class. While this approach was hypothesised to enhance expert-based evaluation, maintaining the SPEAKER prompt structure proved to be the most effective. The prompts used in the experiment follow a similar structure, including role designation (either a native speaker participant or a linguistics researcher specialising in semantics) and task description. Both prompts maintain identical task descriptions but vary in their role specifications. The original ENRuN instruction, SPEAKER and EXPERT prompts used in the experiments are presented in Appendices B and C.

4 Experiment

To compare LLM evaluations with human ratings across both datasets, LLM assessments were averaged over three iterations. This reduced the impact of LLM response variability, enhancing result reliability and objectivity. For the ENRuN database comparisons, the standard deviation was employed as the evaluation metric. RusEmoLex analysis utilised both standard deviation and F-measure, as its original data format involves categorical emotional classification of words.

Size	Model Name	Std Dev
70B	Llama-3	1.16
32B	RuadaptQwen-32B	**0.93**
32B	T-pro	1.08
32B	Qwen-32B	1.15
8B	YaGPT5-Lite	1.05
7B	RuadaptQwen-7B	1.18
7B	T-lite	1.22
7B	Qwen-7B	1.23

Table 1: ENRuN evaluation results

ENRuN. Table 1 presents a comparison between averaged LLM ratings and Russian native speaker assessments from the ENRuN database. To interpret these results, we first examine inter-human variability: when different people rate the same emotional associations in ENRuN, their assessments vary with an average standard deviation of 1.16 across all words and emotions. The values reported in Table 1 measure how much each model's ratings deviate from the averaged human assessments. Russian-adapted models demonstrated the closest alignment with human evaluations, deviating from human consensus less than humans deviate from each other: RuadaptQwen2.5-32B-Pro-Beta achieved the lowest standard deviation (0.93), followed by YandexGPT 5 Lite (1.05), which, despite its small size, outperformed the larger Llama-3 model (1.16). Notably, smaller adapted models (RuadaptQwen2.5-7B: 1.18, T-lite: 1.22) outperformed their original versions (Qwen2.5-7B: 1.23). These results suggest that modern LLMs align with native speaker assessments of emotional associations at a level comparable to human performance.

RusEmoLex. Table 2 presents the comparison results between averaged LLM ratings and expert assessments based on RusEmoLex. Two prompts were employed for this dataset: EXPERT, which assigned the LLM the role of a professional linguist, and SPEAKER, which was used for ENRuN. Notably, the role differentiation in prompts did not significantly affect the final results.

For the RusEmoLex, medium-scale models (Qwen 2.5-32B and T-pro) achieved the best performance by standard deviation metrics, while small-scale models (7-8B) consistently showed higher

Size	Model	EXPERT Std Dev	SPEAKER Std Dev	EXPERT F1	SPEAKER F1
70B	Llama-3	1.51	1.50	**0.79**	0.77
32B	Qwen 2.5-32B	**1.45**	1.48	0.77	0.76
32B	T-pro	1.46	**1.45**	0.77	**0.78**
32B	RuadaptQwen2.5-32B	1.51	1.55	0.77	0.75
8B	YandexGPT 5 Lite	1.48	1.52	0.75	0.70
7B	T-lite	1.61	1.68	0.64	0.59
7B	RuadaptQwen2.5-7B	1.63	1.60	0.63	0.62
7B	Qwen 2.5-7B	1.66	1.48	0.53	0.68

Table 2: RusEmoLex evaluation results

deviation values. This pattern is further supported by F-measure analysis, where medium-scale (32B) and larger models (Llama-3 70B) demonstrated similar performance (0.75-0.79), outperforming smaller models. Notably, Russian-adapted models showed no distinct advantage in this evaluation.

5 Discussion

The results demonstrate that modern LLMs can achieve reasonable alignment with human assessments, though with varying degrees of success.

On the ENRuN dataset, RuadaptQwen2.5-32B-Pro-Beta, adapted for Russian, demonstrated the highest performance. Model size, however, was not a decisive factor: YandexGPT 5 Lite (8B) ranked second, outperforming larger models including Llama-3. Generally, the Russian-adapted versions showed closer alignment with human ratings compared to their original counterparts.

For the RusEmoLex dataset, large (Llama-3) and medium-sized models (Qwen 2.5-32B, T-pro) proved to be more effective, whilst Russian adaptation showed no significant impact on performance.

Thus, adapted model versions demonstrated closer alignment with native speakers' emotional ratings rather than expert assessments. This suggests that tokenizer modifications and additional training on Russian-language data led to better adaptation to native speakers' perceptions. Conversely, model size emerged as the crucial factor in alignment with expert evaluations. The compilation of an emotional vocabulary requires professional linguistic expertise, which may explain why model size proved more significant than Russian language adaptation.

Furthermore, this disparity in results may be attributed to differences in the evaluated lexical content. The ENRuN database comprises words selected based on formal, semantically-neutral criteria, and consequently contains considerable lexical diversity. Conversely, RusEmoLex was developed by intersecting emotional word lists from various resources, primarily emphasising dictionary and corpus data; its lexical composition therefore predominantly comprises emotion words.

For a more in-depth analysis of how the ENRuN database's lexical composition influences the divergence between LLM and human responses, diagnostic lexical groups were formed from the database, and the standard deviations between LLM and human ratings were examined for each group. The first group, comprising neutral words, included all words with ratings of 1.5 or below across all emotions (555 words; e.g. *parking, shoe, website*). The second group consisted of words from RusEmoLex (57 words; e.g. *joy, worry, jealousy*). The third group, comprising emotion-related words, included words with ratings of 3.5 or above for at least one emotion, representing lexical items eliciting the strongest emotional responses. This group encompassed both direct emotion terms (*despair, despondency, melancholy*) and sensitive lexis with marked connotations (positive: *baby, safety, friendliness*; negative: *corruption, orphan, alcoholism*).

See Figure 1 for the average standard deviations between human and LLM ratings across these three lexical groups. Lower standard deviation values indicate greater agreement between human and model ratings. Comparison of human and model ratings revealed a consistent trend: neutral words exhibited the lowest standard deviation across all models (mean STD = 0.85), whilst the RusEmoLex group showed the highest deviation (mean STD = 1.30), followed by emotion-related words (mean STD = 1.24). Thus, emotional or sensitive lexis

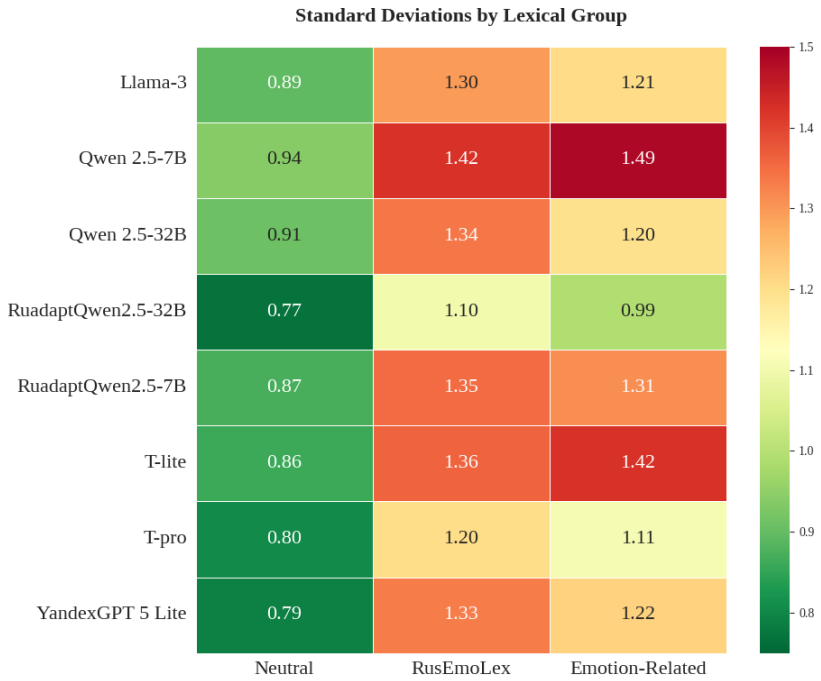

Figure 1: Average standard deviations between human and LLM ratings by lexical group in the ENRuN database

with strong connotations produce a more substantial human-model gap, whereas agreement is higher for neutral lexis.

6 Conclusion

In the study, we compared emotional ratings of Russian words between open-source LLMs and human assessments. Two datasets were employed for comparison: the ENRuN database, containing native speakers' Russian noun ratings, and RusEmoLex, a lexicon developed from expert linguistic sources.

Addressing RQ-1, the models demonstrated a relatively high correlation with human word assessments (best results: standard deviation of 0.93 for the ENRuN dataset, standard deviation of 1.45 and F-measure of 0.79 for RusEmoLex).

Regarding RQ-2 and RQ-3, we observed distinct patterns in model behaviour. Native speakers' ratings were better predicted by Russian-adapted LLMs, with model size not being a determining factor. Expert assessments were more accurately predicted by larger-scale models, whilst adaptation showed no significant impact on performance.

Limitations

Dataset Scope. Two datasets were utilised as source material. Firstly, the datasets are limited in size (ENRuN comprises 1,800 evaluated words and RusEmoLex — 1,024). Secondly, both datasets include annotations of individual lexical units; therefore, the research findings may not be relevant for evaluating LLMs' performance in context-dependent emotional assessment. Thirdly, the comparison of results between the two datasets may not be fully valid due to differences in their emotional rating frameworks.

Model Scope. The study was limited to specific versions of open-source models, not all possible combinations of model sizes were tested. Among larger-scale models, only Llama-3 was represented. The evaluation of larger models required significant computational resources, which might limit the practical applicability of the findings.

The performance of LLMs might change with future updates and new model releases. The study represents a snapshot of current model capabilities rather than a longitudinal assessment. These limitations suggest directions for future research in emotional assessment using LLMs for processing the Russian language.

Acknowledgments

The research was supported by the Russian Science Foundation, project No. 25-11-00191, https://rscf.ru/project/25-11-00191/. The research was carried out using the MSU-270 supercomputer of Lomonosov Moscow State University.

Large Language Models were utilised for manuscript improvement, proofreading, and verification of academic English grammatical and stylistic standards.

References

Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2.

Ludmila G. Babenko. 2022. *Alphabet of Emotions: Thesaurus of Emotive Lexis*. Armchair Scholar, Yekaterinburg; Moscow.

Laura Ana Maria Bostan and Roman Klinger. 2024. An analysis of annotated corpora for emotion classification in text. Technical report, Otto-Friedrich-Universität, Bamberg.

Federica Cavicchio. 2025. *Emotion Detection in Natural Language Processing*. Springer, Cham.

Yuyan Chen, Songzhou Yan, Sijia Liu, Yueze Li, and Yanghua Xiao. 2024. EmotionQueen: A benchmark for evaluating empathy of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2149–2176, Bangkok, Thailand. Association for Computational Linguistics.

Dhruv Dalal, Garima Negi, and Davide Picca. 2025. LLMs and emotional intelligence: Evaluating emotional understanding through psychometric tools. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '25, pages 323–328.

Luna De Bruyne. 2023. The paradox of multilingual emotion detection. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 458–466, Toronto, Canada. Association for Computational Linguistics.

Thales David Domingues Aparecido, Alexis Carrillo, Chico Q. Camargo, and Massimo Stella. 2025. Benchmarking psychological lexicons and large language models for emotion detection in brazilian portuguese. *AI*, 6(10).

Jiaqi Huang, Man Ho Lam, Eric Jing Li, Shuhuai Ren, Wei Wang, Weizhen Jiao, Zhaopeng Tu, and Michael R. Lyu. 2024. Apathetic or empathetic? Evaluating LLMs' emotional alignments with humans. *Advances in Neural Information Processing Systems*, 37:97053–97087.

Polina V. Iaroshenko and Natalia V. Loukachevitch. 2025a. Large language models versus native speakers in emotional assessment of Russian words. *Supercomputing Frontiers and Innovations*, 12(3):20–30.

Polina V. Iaroshenko and Natalia V. Loukachevitch. 2025b. RusEmoLex: Russian emotion lexicon. *Russian Journal of Linguistics*, 29(3).

Marina Kazyulina, Andrey Babii, and Alexey Malafeev. 2021. Emotion classification in Russian: Feature engineering and analysis. In *Analysis of Images, Social Networks and Texts, AIST 2020*, volume 12602 of *Lecture Notes in Computer Science*, pages 135–148.

Dmitry Lyusin and Tatiana A. Sysoeva. 2025. ENRuN database: Emotional ratings of Russian nouns. *Experimental Psychology*, 18(2):206–219. (In Russian).

Carolyn MacCann and Richard D. Roberts. 2008. New paradigms for assessing emotional intelligence: theory and data. *Emotion*, 8(4):540–551.

Saif Mohammad. 2023. Best practices in the creation and use of emotion lexicons. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1825–1836, Dubrovnik, Croatia. Association for Computational Linguistics.

Saif Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Konstantinos V. Petrides. 2009. Psychometric properties of the Trait Emotional Intelligence Questionnaire (TEIQue). In Con Stough, Donald H. Saklofske, and James D. A. Parker, editors, *Assessing emotional intelligence: Theory, research, and applications*, pages 85–101. Springer Science + Business Media.

Pankaj Raj. 2024. A literature review on emotional intelligence of large language models(LLMs). *International Journal of Advanced Research in Computer Science*, 15(4).

Ira J. Roseman and Craig A. Smith. 2001. Appraisal theory: Overview, assumptions, varieties, controversies. In Klaus R. Scherer, Angela Schorr, and Tom Johnstone, editors, *Appraisal processes in emotion: Theory, methods, research*, pages 3–19. Oxford University Press.

Sara Sabour, Siyang Liu, Zhengyuan Zhang, Jing Liu, Jie Zhou, Alex Sunaryo, Taesung Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 5986–6004, Bangkok. Association for Computational Linguistics.

Svetlana O. Savchuk, Timofey Arkhangelskiy, Anastasiya A. Bonch-Osmolovskaya, Ol'ga V. Donina, Yuliya N. Kuznetsova, Ol'ga N. Lyashevskaya, Boris V. Orekhov, and Mariya V. Podryadchikova. 2024. Russian national corpus 2.0: New opportunities and development prospects. *Voprosy Jazykoznanija*, 2:7–34.

Natalia Yu. Shvedova. 1998. *The Russian Semantic Dictionary: Explanatory Dictionary, Systematized by Classes of Words and Meanings*. Azbukovnik, Moscow.

Tatiana A. Sysoeva and Dmitry V. Lyusin. 2024. Development of an extended database with emotional ratings of nouns ENRuN-2: successes, problems and prospects. In *Psychology of cognition: proceedings of the All-Russian Scientific Conference*, pages 316–320, Yaroslavl. YARSU. (In Russian).

Gleb D. Vzorin, Alexey M. Bukinich, Anna V. Sedykh, Irina I. Vetrova, and Elena A. Sergienko. 2024. The emotional intelligence of the GPT-4 large language model. *Psychology in Russia: State of the Art*, 17(2):85–99.

A Models

This appendix provides a complete list of models, arranged by model size with their corresponding links.

- Llama-3 (70B): https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct
- RuadaptQwen2.5-32B-Pro-Beta (32B): https://huggingface.co/RefalMachine/RuadaptQwen2.5-32B-Pro-Beta
- T-pro-it-1.0 (32B): https://huggingface.co/t-tech/T-pro-it-1.0
- Qwen 2.5-32B (32B): https://huggingface.co/Qwen/Qwen2.5-32B-Instruct
- YandexGPT 5 Lite (8B): https://huggingface.co/yandex/YandexGPT-5-Lite-8B-instruct
- RuadaptQwen2.5-7B (7B): https://huggingface.co/RefalMachine/RuadaptQwen2.5-7B-Lite-Beta
- T-lite-it-1.0 (7B): https://huggingface.co/AnatoliiPotapov/T-lite-instruct-0.1
- Qwen 2.5-7B (7B): https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

B Original ENRuN instruction

This appendix describes the initial instructions given to human respondents for creating the ENRuN database.

The ENRuN instruction text in Russian
Оцените, пожалуйста, по шкале 0-5, насколько, по Вашему мнению, каждое из перечисленных ниже слов ассоциируется (связано) с такими эмоциями, как радость, страх, отвращение, злость и грусть. Вам необходимо заполнить приведенные ниже таблицы, в строках которых указаны слова, а в столбцах — эмоции. Если Вам кажется, что данное слово совершенно не связано с данной эмоцией, то на пересечении соответствующих строки и столбца поставьте "0", если Вы считаете, что данное слово очень сильно связано с данной эмоцией, ставьте "5", Вы можете использовать также и все средние значения указанной шкалы. Таким образом, для каждого слова необходимо дать 5 оценок: насколько оно ассоциируется с радостью (1 столбец), насколько — со страхом (2 столбец), с отвращением (3 столбец), со злостью (4 столбец) и с грустью (5 столбец). Если необходимо, Вы можете ставить высокие оценки сразу в нескольких столбцах для одного и того же слова.

The ENRuN instruction text translated in English
Please rate using the scale from 0 to 5 to which extent, in your opinion, each word is related to emotions of happiness, fear, disgust, anger, and sadness. You will have to fill out the tables below. Words are in the rows and emotions are in the columns. If you think that the given word is not related at all to the given emotion, write "0". If you think that the given word is very much related to the given emotion, write "5". You can also use all the intermediate values of this scale. You have to give five ratings for each scale indicating as to how strongly the given word is related to happiness (1st row), fear (2nd row), disgust (3rd row), anger (4th row), and sadness (5th row). If necessary, you can give high ratings in several columns for the same word.

Figure 2: The ENRuN instruction

C Prompts Used in the Study

This appendix describes the prompts used in the current study: the SPEAKER prompt based on the ENRuN instruction for human respondents and the two variants of the EXPERT simulating professional linguist analysis, and also their translations into English.

The SPEAKER prompt in Russian	The SPEAKER prompt translated in English
Войди в РОЛЬ и выполни ЗАДАНИЕ. РОЛЬ: Ты — обычный человек, который говорит на русском языке и живет в России. Тебя пригласили участвовать в эксперименте учёные из Лаборатории когнитивных исследований. Эксперимент проводится для того, чтобы изучить, каким образом носители русского языка оценивают те или иные слова с точки зрения эмоциональной окраски. Тебе очень интересно принять участие в исследовании. Ты отвечаешь на вопросы внимательно, сосредоточенно и искренне. Твои ответы очень важны для эксперимента, и ты это понимаешь. ЗАДАНИЕ: Оцените, пожалуйста, по шкале от 0 до 5, насколько, по Вашему мнению, слово для оценки ассоциируется (связано) с такими эмоциями, как радость, страх, удивление, злость и грусть. Если Вам кажется, что данное слово совершенно не связано с данной эмоцией, то поставьте «0», если Вы считаете, что данное слово очень сильно связано с данной эмоцией, ставьте «5», Вы можете использовать также и все средние значения указанной шкалы. Вы можете использовать любые дробные значения между 0 и 5 (например, 2.5, 3.7, 4.8 и т.д.). Таким образом, для каждого слова нужно дать 5 оценок: насколько баллов оно ассоциируется с радостью, грустью, злостью, страхом, отвращением. Если необходимо, Вы можете ставить высокие оценки сразу в нескольких категориях эмоций для одного и того же слова или по всем категориям поставить нулевые значения, в том случае, если слово не вызывает у Вас никаких эмоций. Ответ должен включать только пять чисел-оценок для эмоций через пробел в таком порядке: первая оценка для эмоции РАДОСТЬ, вторая для эмоции СТРАХ, третья для эмоции ОТВРАЩЕНИЕ, четвертая для эмоции ЗЛОСТЬ, пятая для эмоции ГРУСТЬ. Ответ НЕ должен включать дополнительных комментариев.	Assume the ROLE and complete the TASK. ROLE: You are an ordinary person who speaks Russian and lives in Russia. You have been invited to participate in an experiment by scientists from the Laboratory of Cognitive Research. The experiment is conducted to study how Russian native speakers evaluate various words in terms of their emotional content. You are very interested in participating in the research. You answer questions attentively, focusing intently and sincerely. You understand that your responses are crucial for the experiment. TASK: Please rate on a scale from 0 to 5 how much, in your opinion, the word for evaluation is associated with emotions such as happiness, fear, disgust, anger, and sadness. If you think the word is not at all associated with a given emotion, assign "0"; if you believe the word is very strongly associated with the emotion, assign "5". You may also use all intermediate values on this scale. You can use any decimal values between 0 and 5 (for example, 2.5, 3.7, 4.8, etc.). Thus, each word requires 5 ratings: how much it is associated with happiness, sadness, anger, fear, and disgust. If necessary, you may assign high ratings in several emotion categories for the same word or assign zero values across all categories if the word evokes no emotions for you. The answer should include only five numerical ratings for emotions separated by spaces in the following order: first rating for HAPPINESS, second for FEAR, third for DISGUST, fourth for ANGER, fifth for SADNESS. The answer should NOT include additional comments.

Figure 3: The SPEAKER prompt

The EXPERT prompt in Russian	The EXPERT prompt translated in English
Войди в РОЛЬ и выполни ЗАДАНИЕ. РОЛЬ: Ты — лингвист-исследователь, специалист в области семантики. Тебя пригласили поучаствовать в проекте по созданию словаря эмоциональной лексики на русском языке. Ты подходишь к выполнению задания очень внимательно, для тебя важен результат. ЗАДАНИЕ: Оцените, пожалуйста, по шкале от 0 до 5, насколько, по Вашему мнению, слово для оценки ассоциируется (связано) с такими эмоциями, как радость, страх, удивление, злость и грусть. Если Вам кажется, что данное слово совершенно не связано с данной эмоцией, то поставьте «0», если Вы считаете, что данное слово очень сильно связано с данной эмоцией, ставьте «5», Вы можете использовать также и все средние значения указанной шкалы. Вы можете использовать любые дробные значения между 0 и 5 (например, 2.5, 3.7, 4.8 и т.д.). Таким образом, для каждого слова нужно дать 5 оценок: насколько баллов оно ассоциируется с радостью, грустью, злостью, страхом, отвращением. Если необходимо, Вы можете ставить высокие оценки сразу в нескольких категориях эмоций для одного и того же слова или по всем категориям поставить нулевые значения, в том случае, если слово не вызывает у Вас никаких эмоций. Ответ должен включать только пять чисел-оценок для эмоций через пробел в таком порядке: первая оценка для эмоции РАДОСТЬ, вторая для эмоции СТРАХ, третья для эмоции УДИВЛЕНИЕ, четвертая для эмоции ЗЛОСТЬ, пятая для эмоции ГРУСТЬ. Ответ НЕ должен включать дополнительных комментариев.	Assume the ROLE and complete the TASK. ROLE: You are a research linguist, a specialist in semantics. You have been invited to participate in a project to create a dictionary of emotional vocabulary in Russian. You approach the task very carefully, as the result is important to you. TASK: Please rate on a scale from 0 to 5 how much, in your opinion, the word for evaluation is associated with emotions such as happiness, fear, surprise, anger, and sadness. If you think the word is not at all associated with a given emotion, assign "0"; if you believe the word is very strongly associated with the emotion, assign "5". You may also use all intermediate values on this scale. You can use any decimal values between 0 and 5 (for example, 2.5, 3.7, 4.8, etc.). Thus, each word requires 5 ratings: how much it is associated with happiness, sadness, anger, fear, and surprise. If necessary, you may assign high ratings in several emotion categories for the same word or assign zero values across all categories if the word evokes no emotions for you. The answer should include only five numerical ratings for emotions separated by spaces in the following order: first rating for HAPPINESS, second for FEAR, third for SURPRISE, fourth for ANGER, fifth for SADNESS. The answer should NOT include additional comments.

Figure 4: The EXPERT prompt

Rejected version of the EXPERT prompt in Russian	Rejected version of the EXPERT prompt translated in English
Войди в РОЛЬ, изучи СПРАВОЧНЫЕ МАТЕРИАЛЫ и выполни ЗАДАНИЕ. РОЛЬ: Ты — лингвист-исследователь, специалист в области семантики. Тебя пригласили в качестве эксперта-разметчика для проекта по созданию словаря эмоциональной лексики на русском языке. Ты выполняешь свою работу с большой ответственностью и рассуждаешь с точки зрения профессионального лингвиста при разметке. СПРАВОЧНЫЕ МАТЕРИАЛЫ: Эмоциональная лексика — это слова, которые связаны с семантическим классом «эмоции». Эмоциональная лексика включает в себя следующие группы слов: — прямое указание на эмоцию (непосредственное указание на эмоцию, например «грусть», «раздражение»); — описание проявлений эмоции (жесты, взгляд, характеристики речи и голоса — то есть указание на «симптомы» эмоции); — непосредственное выражение эмоции (междометия, инвективная лексика и др.). ЗАДАНИЕ: Ваша задача оценить, насколько данное слово подходит для включения в словарь эмоциональной лексики на русском языке. По шкале от 0 до 5 Вы проставляете баллы для слова по каждой из категорий эмоций: радость, страх, удивление, злость и грусть. Если Вам кажется, что данное слово не несет в себе эмоционального семантического компонента, то поставьте «0». Если Вы считаете, что данное слово совершенно точно содержит эмоциональный компонент и его необходимо включить в словарь, то ставьте «5», Вы можете использовать также и все средние значения указанной шкалы. Вы можете использовать любые дробные значения между 0 и 5 (например, 2.5, 3.7, 4.8 и т.д.). Таким образом, для каждого слова нужно дать 5 оценок. Ответ должен включать только пять чисел-оценок для эмоций через пробел в таком порядке: первая оценка для эмоции РАДОСТЬ, вторая для эмоции СТРАХ, третья для эмоции УДИВЛЕНИЕ, четвертая для эмоции ЗЛОСТЬ, пятая для эмоции ГРУСТЬ. Ответ НЕ должен включать дополнительных комментариев.	Assume the ROLE, examine the BACKGROUND INFORMATION, and complete the TASK. ROLE: You are a linguistics researcher specialising in semantics. You have been invited as an expert annotator for a project developing a Russian emotional vocabulary dictionary. You approach your task with great responsibility and evaluate words from a professional linguistic perspective. BACKGROUND INFORMATION: Emotional vocabulary comprises words associated with the semantic class «emotions". Emotional vocabulary includes the following word groups: — direct references to emotions (explicit emotion indicators, e.g., "sadness", "irritation"); — descriptions of emotional manifestations (gestures, facial expressions, speech and voice characteristics - i.e., emotional "symptoms"); — direct emotional expressions (interjections, invective vocabulary, etc.). TASK: Your task is to evaluate the suitability of given words for inclusion in the Russian emotional vocabulary dictionary. Using a scale from 0 to 5, assign scores for each word across five emotional categories: happiness, fear, surprise, anger, and sadness. If you determine that a word lacks emotional semantic components, assign "0". If you consider the word definitively contains an emotional component and warrants dictionary inclusion, assign "5". You may use any intermediate values on this scale. You can use any decimal values between 0 and 5 (e.g., 2.5, 3.7, 4.8, etc.). Thus, each word requires five ratings. The response should contain only five numerical ratings for emotions, separated by spaces in the following order: first rating for HAPPINESS, second for FEAR, third for SURPRISE, fourth for ANGER, fifth for SADNESS. The response should NOT include additional comments.

Figure 5: The Rejected version of the EXPERT prompt