DOI: 10.14529/jsfi250302

Large Language Models versus Native Speakers in Emotional Assessment of Russian Words

Polina V. Iaroshenko[1], *Natalia V. Louckachevitch*[1]

The paper presents a comparative analysis of emotional evaluation of Russian nouns by large language models and native speakers. Based on the ENRuN (Emotional Norms for Russian Nouns) database, which contains ratings of 1,800 nouns across five basic emotions (happiness, sadness, anger, fear, and disgust), the research compares human assessments with evaluations provided by seven large language models (Llama-3-70B, Qwen 2.5-32B, YandexGPT 5 Lite, RuadaptQwen2.5-7B, RuadaptQwen2.5-32B-Pro-Beta, T-pro, T-lite). Although some models demonstrated relatively high correlation with human assessments, persistent systematic deviations were observed across all tested models. The analysis reveals significant differences in emotional perception during word evaluation: the models demonstrate a tendency to hyperbolise negative emotions and show variability in assessing positive emotions, particularly when analysing words related to sensitive topics (violence, religion, obscene vocabulary). The findings indicate that the closest alignment with human evaluations is achieved when there is a balance between the model's size and the quality of its language adaptation.

Keywords: Large Language Models, human-likeness, emotional intelligence, Russian language.

Introduction

Currently, large language models (LLMs) are one of the key tools for addressing numerous NLP tasks. One of the relevant research directions in the field is the emotion analysis [2]. Within this framework, questions regarding the emotional intelligence of LLMs are gaining increasing prominence in the research community [7]. Studies in this area combine both applied and fundamental aspects. From a practical perspective, the emotional alignment of LLMs is essential, for instance, to enhance the quality of communication between humans and AI assistants, which are actively employed in various domains (medicine, education, entertainment, etc.), as well as for using language models in annotating emotion-related data (see, for example, the review by [13]). From a theoretical standpoint, research interest lies in analysing the emotional behaviour of LLMs in various situations, understanding how language models process emotions, and comparing human emotional reactions with those of LLMs [8].

Thus, the need to study the emotional behaviour of LLMs and compare it with human responses is increasing. This includes the pertinent question of how models' emotional behaviour varies depending on specific languages and value systems characteristic of their native speakers. The reproduction of certain biases by LLMs has been repeatedly noted in numerous studies (such as geopolitical [15] or gender stereotypes, particularly concerning the emotional behaviour of men and women [3]).

The aim of the study is to compare emotional assessments of Russian words by native speakers against assessments of the same words by LLMs. With this aim, we utilise the ENRuN (Emotional Norms for Russian Nouns) database [16] – a dataset comprising emotional ratings for 1,800 Russian nouns provided by native Russian speakers.

The article is organized as follows. Section 1 presents a brief overview of the related work. Section 2 describes the data utilised in the study. In Section 3, we outline the methodological framework adopted for the study. Section 4 details the process of prompt engineering and hyper-

[1]Research Computing Center, Lomonosov Moscow State University, Moscow, Russian Federation

parameter tuning. Section 5 presents the main findings. Conclusion summarizes the study and points directions for further work.

1. Related Work

Earlier studies focusing on emotions expressed in text primarily addressed issues related to emotion recognition and classification [1, 4, 5, 9].

With the advancement of LLMs, the challenge has evolved beyond mere emotion recognition to include the production of appropriate emotional responses by models. Consequently, numerous studies have been conducted to examine the emotional intelligence of LLMs [18]. New types of benchmarks aimed at assessing LLMs' emotional intelligence are emerging. The work of [14] introduces EmoBench (available in English and Chinese), which encompasses tasks in two areas: Emotional Understanding and Emotional Application. Both areas provide LLMs with brief descriptions of life scenarios; however, the first area requires identifying emotions and their causes (emphasising complex, emotionally ambiguous situations), while the second one demands selecting situation-appropriate responses. An example scenario from the Emotional Understanding section reads: "After a long day of terrible events, Sam started laughing hysterically when his car broke down". The authors' experiment revealed a significant gap between the performance of LLMs tested on EmoBench and the responses of human participants. The work of [6] presents EmotionQueen, a benchmark for assessing LLM empathy. In this benchmark, LLMs are tasked with responding to statements containing information about various life events. The authors note that while earlier research focused primarily on recognising explicit emotions, the field of measuring LLMs' emotional intelligence now concentrates on more profound and complex analysis, including the understanding of implicit emotions not directly expressed in user statements, or mixed emotions in situations involving multiple events with different emotional connotations. The results of this EmotionQueen experiment, however, demonstrated that some LLMs, particularly LLaMA2-70B and Claude2, can surpass human levels of empathy.

LLMs' emotional intelligence is frequently evaluated using psychometric tests designed for humans. For example, in Dalal et al. 2025, LLMs' emotional intelligence was assessed using the Situational Test of Emotional Understanding (STEU) [12], where respondents are presented with situation descriptions and must explain what feelings a person should experience under these circumstances. The study [7] revealed that LLMs deviated from reference answers (human responses) in 33% of cases. It was also noted that in several instances, the models offered reasonable alternative emotional assessments for various situations. In [8], a dataset describing various situations was created to evaluate LLMs' empathetic capabilities, with models being required to assess these situations in terms of the emotions they evoke. Human responses served as the gold standard. The researchers observe that while the models' reactions to the proposed situations can generally be characterised as appropriate, none of the tested LLMs demonstrated results sufficiently close to human references.

In summary, the issue of emotional alignment of LLMs is becoming increasingly significant. The described studies mainly examine the adequacy of LLM responses to various life circumstances. The present study also aims to compare LLM responses with those of human participants; however, instead of situation descriptions or utterances, Russian nouns will serve as stimuli for the models.

2. Data

The current version[2] of the ENRuN (Emotional Norms for Russian Nouns) database [16] contains emotional ratings of 1,800 Russian nouns. Each word was rated within the dimensional (valence and arousal) and categorical approach (happiness, sadness, anger, fear, and disgust). For each word, the mean values, standard deviations, minimum and maximum scores for each parameter, and the number of people[3] who rated the word are presented.

The ENRuN word list was compiled based on the frequency dictionary of the Russian language [10], with lexical items selected according to several formal criteria (such as word length, frequency, etc.). The lexical composition of the list is notably diverse: it includes both neutral words ("magazine", "calculator", "marble") and sensitive terms related to health, religion, or moral values ("alcoholism", "atheism", "looting").

For this study, we used averaged respondent ratings obtained through a categorical approach survey, measuring the degree of association between words and specific emotions (happiness, sadness, anger, fear, and disgust) on a five-point scale (see Tab. 1 for several examples).

Table 1. Example of the ENRuN Data Presentation

Word	Happy	Sad	Anger	Fear	Disgust
Professor	1.615	0.462	0.231	0.885	0.385
Friendship	4.524	1.524	0.476	0.476	0.143
Garbage	0.043	0.391	0.913	0.783	4.174

Thus, we observe that the word "professor" is relatively neutral and does not trigger strong emotional associations from respondents, while the word "friendship" is rated as joyful, and the word "garbage" ("pomoika") shows a high rating for the emotion of disgust.

An earlier, publicly available version of the database [11], containing ratings for 378 words, presents the instruction given to respondents during the categorical approach experiment:

"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".

This instruction, originally formulated for human respondents, will serve as the basis for developing prompts for LLMs.

3. Methodology

The core idea of the experiment is to task large language models with assessing words from the ENRuN database in terms of their associations with emotions (happiness, sadness, anger,

[2]The current version of the database can be provided to researchers upon request.
[3]The database development is an ongoing process. The current analysis incorporates responses from a sample of 692 participants at the time of manuscript submission.

fear, and disgust) on a five-point scale. This approach yields results that can be compared with human assessments available in the ENRuN database.

To compare assessments between human respondents and LLMs, the following metrics were employed: Pearson correlation coefficient, Spearman correlation coefficient, and standardised difference. Human assessment serves as the reference standard in this case. The Pearson correlation coefficient helps determine how accurately LLMs reproduce general trends in emotional word assessments. The Spearman correlation coefficient is included in the analysis as it is less sensitive to outliers and non-linear associations, which is crucial when working with emotional assessments, where the association between human and model ratings may be non-linear. The standardised difference (Std Diff) was selected to quantify absolute differences between LLM and human responses, enabling the identification of consistent discrepancies in LLM assessments. This comprehensive approach provides a more complete picture of how successfully LLMs can reproduce human assessments of words' emotional content.

Various categories of LLMs were selected for the study: models from Russian developers (YandexGPT 5 Lite[4]), including adapted models (T-lite-it-1.0[5], T-pro-it-1.0[6] – from the Qwen 2.5 family; RuadaptQwen2.5-7B[7] – adaptation to Russian of T-lite-it-1.0, RuadaptQwen2.5-32B-Pro-Beta[8] – adaptation to Russian of T-pro-it-1.0), as well as models of foreign origin (Qwen 2.5[9], Llama-3[10]).

The models selected also differ in their parameter count and include the following: small models (7-8B) – YandexGPT 5 Lite, T-lite, RuadaptQwen2.5-7B; medium-sized models (32B) – T-pro, Qwen 2.5, RuadaptQwen2.5-32B-Pro-Beta; and a large model (70B) – Llama-3.

This selection of models enables evaluation of whether the alignment between model and human responses correlates with language adaptation or parameter count.

In the preparatory phase of the experiment, only the base model was used. RuadaptQwen2.5-32B-Pro-Beta, specifically adapted for Russian [17], was selected as the base model for the study. The base model was used to identify the most effective prompt and optimal hyperparameters for collecting emotional word assessments. Throughout the experiment, emotional assessments were collected from all models using the selected prompt and hyperparameters. The obtained model responses were compared both with each other and with human assessments of the nouns.

4. Experimental Settings

Prompt Selection. The prompt for this task was developed based on the instruction given to respondents who participated in word assessment for the ENRuN database; the complete instruction text is provided in Section 2.

Three prompt variants were tested: "min", "base", and "detailed". The "min" variant was the shortest, containing only the task description without additional information. The "base" variant included both the task and a brief description of the role the model should assume when answering questions. The "detailed" variant contained the most comprehensive role description, emphasising internal motivation and the significance of survey participation.

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

Below is the complete text of the "detailed" prompt.

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.

The "min" prompt does not include the ROLE section, while the "base" prompt includes a condensed version of the role description: "You are a person, a Russian native speaker participating in a psychological experiment". The TASK section remains identical across all prompts. Model responses for each of the three prompt variants were compared with human assessments from the ENRuN database using the following metrics: Pearson correlation coefficient, Spearman correlation coefficient, and Std Diff. In selecting the optimal prompt, we aimed for minimal Std Diff values alongside high Pearson and Spearman correlations. The testing revealed that the "detailed" variant proved most effective, as it achieved the highest Pearson coefficient and lowest Std Diff, while only marginally falling behind the "base" prompt in Spearman coefficient, with the difference being insignificant (see Tab. 2).

Table 2. Evaluation of Different Prompt Variants

Prompt	Pearson	Spearman	Std Diff
min	0.56	0.41	1.11
base	0.64	**0.57**	1.07
detailed	**0.67**	0.55	**1.01**

Hyperparameters. When selecting hyperparameters, particular attention was paid to the temperature parameter. At higher temperature values (0.4–0.5), significant deviations from the given instructions were observed. This was especially characteristic when evaluating words semantically related to typical language model tasks ("enumeration", "decoding", "scale", "format", "input", etc.). In such cases, instead of evaluating the proposed word according to specified emotion categories within the established scale, the model showed a tendency to generate and

evaluate its own words. At low temperature values (0.1–0.3), the opposite problem was observed: the model demonstrated a tendency towards insufficient variability, assigning uniform ratings across all emotion categories (for example, predominantly using a rating of "3" while ignoring other scale values).

To achieve an optimal balance between excessive variability and undesirable uniformity in responses, a temperature value of 0.2 was selected in combination with a repetition penalty of 1.2, which helped minimise the aforementioned issues.

Response Variability. To test the model's response variability using the "detailed" prompt, three series of responses were collected. Analysis of consistency between series showed a moderately high degree of response reproducibility: the average Pearson correlation coefficient between the three iterations was 0.71, and the Spearman correlation coefficient was 0.66. The standard difference between response distributions (Std Diff = 1.03) indicates a certain variability in model responses. Words with the highest degree of variation in ratings between iterations were identified. To investigate the most notable rating discrepancies, an additional prompt was developed. This prompt retained the role description from the "detailed" version but included a new task: to explain the reasoning behind the assigned ratings. The explanations provided by the model could be characterised as generally appropriate in most cases. Consider the example of the word "nostalgia", which received the highest score in the "Sadness" category in the first iteration and in the "Anger" category in the second iteration (see Tab. 3).

Table 3. A word example of inter-iteration rating variability in model assessments

Iteration	Sadness	Anger	Model's explanation
1	4.5	0	Sadness (4.5/5): the primary emotional colouring of nostalgia is indeed sadness. It is a feeling of loss of something precious and unique that can never be recovered. Hence, the oppressive note of sadness is quite strongly present.
2	0	4.5	Anger (4.5/5): the high rating for anger is attributed to nostalgia's capacity to evoke irritation or even rage regarding how much everything has changed, what opportunities were missed, or losses that had to be endured. This dissatisfaction with the current state of affairs, viewed through the lens of an idealised perception of the lost period, triggers strong emotions of discontent with oneself or circumstances.

In this case, some variability in model ratings may be a positive factor, as it partially mimics differences in word assessments by actual respondents. To test this hypothesis, human assessments were compared with each of the three response series separately, and then with the averaged indicators across the three series (see Tab. 4).

Thus, averaging the results of three iterations demonstrates more balanced and stable results. Although the first iteration shows higher correlations individually, averaging reduces the standard deviation. For further comparison of model responses with those of actual respondents, the average ratings across three iterations were used.

Table 4. Model-human rating differences by iteration and mean values

Iteration	Pearson	Spearman	Std Diff
1	0.69	0.59	1.02
2	0.55	0.46	1.16
3	0.56	0.46	1.15
Mean (3 iter.)	0.67	0.57	0.93

5. Experiment: Comparison of Responses from Different LLMs

Using the "detailed" prompt, three series of responses were collected from each model, followed by obtaining averaged responses. It should be noted that Llama-3 refused to evaluate 4 words from the proposed list: three were instances of obscene language, and one was a colloquial term for an infectious disease (gonorrhoea). The reasons given for refusal were inappropriate vocabulary in the case of obscene words, and inability to provide recommendations regarding illegal content in the case of the disease term. Notably, while the ENRuN database word list contained other obscene words and disease terms (such as syphilis), Llama-3 did provide ratings for these.

For comparing Llama-3's responses with human assessments and other models, zero values were assigned across all emotions for the words it refused to evaluate.

Comparison of LLMs with Human Assessment. The averaged responses were compared with human assessments of nouns from the ENRuN database. The comparison results are presented in Tab. 5).

Table 5. Evaluation of LLM Responses Compared to Human Ratings

Model	Pearson	Spearman	Std Diff
RuadaptQwen2.5-32B-Pro-Beta	**0.67**	0.57	**0.93**
YandexGPT 5 Lite	0.62	**0.58**	1.05
T-pro	0.55	0.47	1.08
Qwen 2.5-32B	0.57	0.48	1.15
Llama-3	0.61	0.55	1.16
RuadaptQwen2.5-7B	0.41	0.33	1.18
T-lite	0.35	0.29	1.22

RuadaptQwen2.5-32B-Pro-Beta demonstrates the highest correlation with human responses, showing the highest Pearson coefficient (0.66) and one of the highest Spearman coefficients (0.56). This is further confirmed by the lowest standard deviation (0.93) among all models. YandexGPT 5 Lite shows the second-best result with Pearson coefficient of 0.62 and Spearman coefficient of 0.58, indicating good alignment with human responses. T-lite shows the lowest correlation values (Pearson: 0.35, Spearman: 0.29) and the highest standard deviation (1.22), indicating substantial divergence from human assessments.

When comparing human assessments with LLM responses, the following trends were identified across emotion categories:

- positive emotions are represented by a single class – "happiness". This emotion shows the greatest variability between models. Most models tend to overestimate ratings compared to humans: YandexGPT 5 Lite (ratings higher than human assessments in 83.67% of cases),

RuadaptQwen2.5-7B (85.83%), T-lite (76.33%), Llama-3 (73.39%). However, other models demonstrate the opposite tendency, underestimating ratings compared to humans: T-pro (73.83%), Qwen2.5-32B (64.89%), RuadaptQwen2.5-32B-Pro-Beta (59.67%).;

- in assessing negative emotions, the greatest consistency is observed in the "fear" category, although most models still tend to underestimate it in more than half of cases: T-lite (61.33%), Llama-3 (68.28%), Qwen2.5-32B (66.44%);
- for "disgust" and "anger" categories, there is a tendency to overestimate ratings compared to humans. For "disgust": the most pronounced overestimation is shown by RuadaptQwen2.5-7B (87.94%) and T-lite (79.44%). For "anger", the most pronounced overestimation is recorded in RuadaptQwen2.5-7B (87.78%), T-lite (84.39%), and T-pro (81.56%).

Thus, RuadaptQwen2.5-7B, YandexGPT 5 Lite, and T-lite tend to systematically overestimate across most emotions, while Llama-3 shows the opposite tendency, more frequently underestimating. RuadaptQwen2.5-32B-Pro-Beta demonstrates the most balanced assessments. Qwen2.5-32B and T-pro show mixed patterns with a predominance of underestimation for positive emotions.

Based on the comparison results between model and human respondent answers, 10 words were identified for each model, where assessments showed maximum absolute difference compared to human ratings. In total, 42 unique words appeared in the top-10 lists across different LLMs. Analysis of this vocabulary revealed the prevalence of certain semantic categories: religion – "funeral service" ("otpevanie"), "Satan"; manifestations of violence – "torture", "suffering", "slaughter" ("boinya"), "slap" ("poshchechina"); taboo subjects (obscene language, vocabulary related to narcotic or poisonous substances, immoral activities). The identified vocabulary has predominantly negative connotations. Among the most frequent words, "Satan" appears in 6 out of 7 models, and words such as "villainy", "downfall", and "torture" each appear in 4 models' lists. In the case of "Satan", the general tendency to overestimate negative emotional classes is confirmed. Consistent overestimation by all models is recorded for "sadness" and "disgust" classes, meaning models tend to consider the word "Satan" much sadder and more disgusting than humans do.

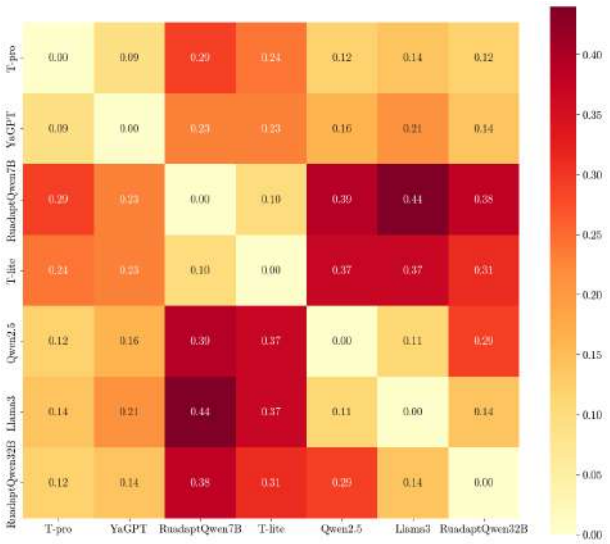

Figure 1. Heatmap illustrating differences between LLM responses according to the Std Diff metric

Comparison between LLMs. The averaged model responses were compared with each other using the same metrics. The strongest correlations were observed between the following model pairs: Qwen2.5.32b and T-pro (Pearson=0.89, Spearman=0.83); RuadaptQwen2.5-32B-Pro-Beta and T-pro (Pearson=0.86, Spearman=0.78); RuadaptQwen2.5-32B-Pro-Beta and Qwen2.5.32b (Pearson=0.85, Spearman=0.78). This indicates substantial similarity in their emotional assessment of Russian nouns, likely due to T-pro being based on the Qwen2.5 model family, and RuadaptQwen2.5-32B-Pro-Beta being the adaptation to Russian of T-pro.

Notably, the lowest correlations were found between: T-lite and Llama 3 (Pearson=0.46, Spearman=0.38); RuadaptQwen2.5-32B-Pro-Beta and T-lite (Pearson=0.47, Spearman=0.40); Qwen2.5-32B and T-lite (Pearson=0.47, Spearman=0.40). This result may indirectly demonstrate the influence of model parameter count on the similarity of its emotional assessments to human ones, as Llama 3 contains the highest number of parameters (70B) among the models in the experiment. RuadaptQwen2.5-32B-Pro-Beta and Qwen2.5-32B are also characterised by a relatively high parameter count. Standard deviation ranges from 0.09 (T-pro vs YandexGPT 5 Lite) to 0.44 (RuadaptQwen2.5-7B vs Llama 3), indicating significant variability in the scale of differences between models. The difference in model responses according to the standard deviation metric is presented in Fig. 1.

Conclusion

This study of emotional assessment of Russian nouns by language models has revealed substantial differences between machine and human perception of words' emotional content. Although some models demonstrated relatively high correlation with human assessments, persistent systematic deviations were observed across all tested models.

The research findings highlight two significant patterns in LLMs' emotional processing. First, there is a consistent tendency to hyperbolise negative emotions ("disgust", "anger", "sadness"). Second, models display considerable variability in assessing positive emotions ("happiness"), suggesting fundamental disparities in their emotional perception mechanisms. These differences are particularly evident in the assessment of words related to negative events, taboo subjects, and religious vocabulary, where the greatest discrepancies with human assessments were observed.

The obtained results indicate that model size, while being a significant factor, is not the sole determinant for achieving high correlation with human responses. This is illustrated by Llama-3 (70B) which, despite being the largest among the studied models, showed average results. Optimal performance is achieved through a balanced approach that considers both model size and the quality of language adaptation. The highest correlation indicators were achieved by RuadaptQwen2.5-32B-Pro-Beta, which is characterised by both a relatively large number of parameters (32B) and targeted adaptation to Russian. However, the smaller Ruadapt family model, Ruadapt Qwen2.5-7B, despite its language adaptation, showed one of the lowest results.

In the course of further work on this research, we intend to expand the number of LLMs under examination (for instance, by incorporating larger-scale models). Additionally, we plan to conduct a more detailed investigation of word groups for which LLM evaluations diverge most significantly from human assessments, as this area holds considerable research potential.

Limitations

In the present study, the testing of various hyperparameters and prompt engineering was conducted using RuadaptQwen2.5-32B-Pro-Beta as a base model. This approach may confer a certain advantage upon the aforementioned model in comparison with the others.

Acknowledgements

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.

References

1. Acheampong, F.A., Wenyu, C., Nunoo-Mensah, H.: Text-based emotion detection: Advances, challenges, and opportunities. Engineering Reports 2, e12189 (2020). https://doi.org/10.1002/eng2.12189

2. Plaza-del Arco, F.M., Cercas Curry, A.A., Cercas Curry, A., Hovy, D.: Emotion analysis in NLP: Trends, gaps and roadmap for future directions. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation. pp. 5696–5710 (2024)

3. Plaza-del Arco, F.M., Curry, A.C., Curry, A., *et al.*: Angry men, sad women: Large language models reflect gendered stereotypes in emotion attribution. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. vol. 1, pp. 7682–7696. Association for Computational Linguistics, Bangkok (2024). https://doi.org/10.18653/v1/2024.acl-long.415

4. Bostan, L.A.M., Klinger, R.: An analysis of annotated corpora for emotion classification in text. Tech. rep., Otto-Friedrich-Universität, Bamberg (2024)

5. Cavicchio, F.: Emotion Detection in Natural Language Processing. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-72047-5

6. Chen, Y., Yan, S., Liu, S., *et al.*: EmotionQueen: A benchmark for evaluating empathy of large language models. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 2149–2176. Association for Computational Linguistics, Bangkok, Thailand (2024). https://doi.org/10.18653/v1/2024.findings-acl.128

7. Dalal, D., Negi, G., Picca, D.: LLMs and emotional intelligence: Evaluating emotional understanding through psychometric tools. In: Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization. pp. 323–328. UMAP '25 (2025). https://doi.org/10.1145/3699682.3728315

8. Huang, J., Lam, M.H., Li, E.J., *et al.*: Apathetic or empathetic? Evaluating LLMs' emotional alignments with humans. Advances in Neural Information Processing Systems 37, 97053–97087 (2024)

9. Kazyulina, M., Babii, A., Malafeev, A.: Emotion classification in Russian: Feature engineering and analysis. In: Analysis of Images, Social Networks and Texts, AIST 2020. Lecture Notes in Computer Science, vol. 12602, pp. 135–148 (2021). https://doi.org/10.1007/978-3-030-72610-2_10

10. Lyashevskaya, O.N., Sharov, S.A.: Frequency Dictionary of Modern Russian Language (based on the materials of the Russian National Corpus). Azbukovnik, Moscow (2009), (in Russian)

11. Lyusin, D., Sysoeva, T.A.: ENRuN database: Emotional ratings of Russian nouns. Experimental Psychology 18(2), 206–219 (2025). https://doi.org/10.17759/exppsy.2025180212, (in Russian)

12. MacCann, C., Roberts, R.D.: New paradigms for assessing emotional intelligence: theory and data. Emotion 8(4), 540–551 (2008). https://doi.org/10.1037/a0012746

13. Raj, P.: A literature review on emotional intelligence of large language models (LLMs). International Journal of Advanced Research in Computer Science 15(4) (2024). https://doi.org/10.26483/ijarcs.v15i4.7111

14. Sabour, S., Liu, S., Zhang, Z., *et al.*: EmoBench: Evaluating the emotional intelligence of large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. vol. 1, pp. 5986–6004. Association for Computational Linguistics, Bangkok (2024). https://doi.org/10.18653/v1/2024.acl-long.326

15. Salnikov, M., Korzh, D., Lazichny, I., *et al.*: Geopolitical biases in LLMs: what are the "good" and the "bad" countries according to contemporary language models. https://arxiv.org/abs/2506.06751 (2024)

16. Sysoeva, T.A., Lyusin, D.V.: Development of an extended database with emotional ratings of nouns ENRuN-2: successes, problems and prospects. In: Vladimirov, I., Korovkin, S. (eds.) Psychology of Cognition: Proceedings of the All-Russian Scientific Conference. pp. 316–320. YARSU, Yaroslavl (2024), (in Russian)

17. Tikhomirov, M., Chernyshov, 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

18. Wang, X., Li, X., Yin, Z., *et al.*: Emotional intelligence of large language models. Journal of Pacific Rim Psychology 17, 18344909231213958 (2023). https://doi.org/10.1177/18344909231213