Тестирование больших языковых моделей на минимальных эмоциональных предложениях

Д. А. Мурсалимов*, П. В. Ярошенко‡, Н. В. Лукашевич§

На минимальных русскоязычных предложениях оценивались точность классификации эмоций восемью моделями и влияние языковой формы на их ответы. Точность Alice AI LLM и Qwen 3.5-397B-A17B при округлении составила 0,751. Наибольший macro-F1 наблюдался у Alice AI LLM (0,742). В утвердительных предложениях прошедшего времени замена точки вопросительным знаком чаще меняла метку у каждой модели, чем у аннотаторов.

Ключевые слова: языковые модели, классификация эмоций, отрицание, пунктуация, русский язык.

1. Задача и данные

Распознавание эмоций в тексте важно для диалоговых систем и анализа сообщений [1]. Проверки на контролируемых изменениях текста выявляют ошибки моделей, скрытые за общими метриками [2]. Минимальные предложения позволяют оценить реакцию на отдельные языковые признаки при одном и том же глаголе.

Для 473 глаголов словаря русских эмоциональных слов RusEmoLex [3] составлено по 12 предложений: «Она (не) + глагол в прошедшем времени женского рода» или «(Не) + императив единственного числа» с точкой, вопросительным или восклицательным знаком. Все 5 676 предложений размечены пятью носителями русского языка. Аннотаторы размечали предложения по следующим категориям: радость, грусть, злость,

* *Мурсалимов Данил Азатович* — аспирант факультета ВМК МГУ имени М. В. Ломоносова; науч. руководитель: Кузина Лариса Николаевна, e-mail: murdanil7@yandex.ru, ORCID: 0009-0001-8005-7757.

Mursalimov Danil — PhD student, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University; supervisor: Larisa Nikolaevna Kuzina.

‡ *Ярошенко Полина Владимировна* — канд. филол. наук, науч. сотр. НИВЦ МГУ, e-mail: polina.iaroshenko@yandex.ru, ORCID: 0000-0002-9425-5842.

Iaroshenko Polina V. — PhD in Philology, Researcher, Research Computing Center, Lomonosov Moscow State University.

§ *Лукашевич Наталья Валентиновна* — д-р техн. наук, вед. науч. сотр. НИВЦ МГУ, e-mail: louk_nat@mail.ru, ORCID: 0000-0002-1883-4121.

Loukachevitch Natalia V. — DSc in Engineering, Leading Researcher, Research Computing Center, Lomonosov Moscow State University.

Опубликовано на условиях лицензии CC-BY 4.0
Published under CC-BY 4.0

страх, удивление, нейтральный класс и неграмматично. После отбора по совпадению не менее трёх оценок и наличию не более одной пометки «неграмматично» осталось 4 127 предложений для 470 глаголов.

2. Эксперимент и оценка языковых моделей

Восьми моделям без дообучения подавалось по одному предложению с общей русскоязычной инструкцией. При температуре 0 требовалось выбрать одну из шести меток по смыслу высказывания, предпочитая эмоциональную при сомнении, и вернуть её в JSON.

Таблица 1: Качество классификации. Точки показывают оценки метрик, отрезки — 95%-е доверительные интервалы.

Метод	Accuracy	0,65 0,70 0,75	Macro-F1	0,65 0,70 0,75
Alice AI LLM	**0,751**		**0,742**	
Qwen 3.5-397B-A17B	**0,751**		0,733	
YandexGPT Pro 5.1	0,732		0,734	
YandexGPT Lite 8B	0,722		0,716	
Qwen 3.6-27B	0,722		0,719	
GPT-OSS 120B	0,713		0,708	
DeepSeek V4.1 Flash	0,674		0,675	
Qwen3.5-9B	0,670		0,662	
RusEmoLex	0,584		0,509	
«Грусть» (частый класс)	0,224		—	
Равномерный выбор	0,167		—	

Метка глагола из RusEmoLex присваивалась всему предложению.

По точности лидировали Alice AI LLM и Qwen 3.5-397B-A17B. Интервалы и парные тесты рассчитывались бутстрэпом по глаголам. С поправкой Холма (14 тестов на лидера) Alice значимо опередила 6 моделей по точности и 5 по macro-F1, Qwen — 5 и 2 соответственно ($p < 0{,}05$). Между лидерами не выявлено значимых различий ни по одной метрике.

3. Влияние языковой формы на ответы моделей

Метки всех моделей и аннотаторов сравнивались на 3 539 парах. Исходная форма — утверждение в прошедшем времени с точкой (рис. 1).

В логистической GEE с кластеризацией по глаголам тест Вальда выявил различия в реакции на преобразования ($\chi^2(80) = 1021{,}55$, $p < 0{,}001$). Совместно проверялись 80 коэффициентов взаимодействия (8 моделей по 10 сравнений с опорным преобразованием относительно разметки).

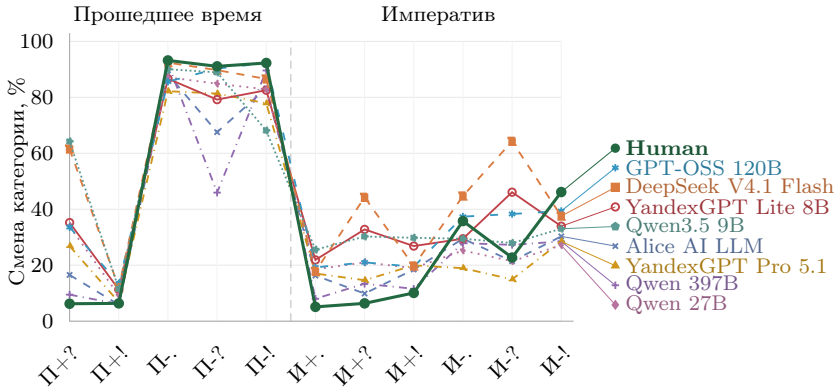


Рис. 1: Смена категории на общих парах. П — прошедшее время, И — императив, + — без отрицания, − — с отрицанием. Human — метка большинства аннотаторов.

4. Пробинг внутренних представлений

Линейный классификатор с L_2-регуляризацией обучался предсказывать метку аннотаторов по активациям YandexGPT Lite 8B до генерации ответа. Для обучения и валидации использовались случаи согласия модели с аннотаторами (2 132 и 428). Проверялось, переносилась ли связь между активациями и меткой с примеров согласия на примеры расхождения. Глаголы теста не встречались при обучении и валидации.

Сравнивались среднее, максимум и минимум по токенам предложения, векторы последнего токена предложения и служебного символа между вводом и ответом модели. Наибольший macro-F1 на валидации дал вектор служебного символа на выходе 20-го слоя. На полном тесте (589 примеров) точность классификатора составила 0,783, ответов модели — 0,713. Классификатор предсказал метку аннотаторов в 58 из 169 случаев, где модель ошиблась.

5. Заключение

В работе оценивались точность моделей на минимальных эмоциональных предложениях и влияние языковой формы на их ответы. Модели сходно с аннотаторами реагировали на отрицание, а замена точки вопросительным знаком чаще меняла метку у моделей. На части примеров, где YandexGPT Lite 8B ошиблась, линейный классификатор правильно предсказал метку аннотаторов по активациям внутренних слоёв модели.

При подготовке текста и кода анализа применялись большие языковые модели.

Список литературы

[1] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, S. Ravi, "GoEmotions: A Dataset of Fine-Grained Emotions", *Proc. 58th ACL*, 2020, 4040–4054. DOI: 10.18653/v1/2020.acl-main.372.

[2] M. T. Ribeiro, T. Wu, C. Guestrin, S. Singh, "Beyond Accuracy: Behavioral Testing of NLP Models with CheckList", *Proc. 58th ACL*, 2020, 4902–4912. DOI: 10.18653/v1/2020.acl-main.442.

[3] P. V. Iaroshenko, N. V. Loukachevitch, "RusEmoLex: Russian Emotion Lexicon", *Russian Journal of Linguistics*, **29**:3 (2025), 659–687. DOI: 10.22363/2687-0088-44439.

Статья поступила 6 ноября 2026 г.

Testing Large Language Models on Minimal Emotional Sentences

D. Mursalimov, P. V. Iaroshenko, N. V. Loukachevitch

Eight language models classified emotions in minimal Russian sentences. Alice AI LLM and Qwen 3.5-397B-A17B tied in rounded accuracy (0.751). Alice AI LLM had the highest macro-F1 (0.742). Replacing a period with a question mark in affirmative past-tense sentences changed labels more often for each model than for annotators.

Keywords: language models, emotion classification, negation, punctuation, Russian.

References

[1] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, S. Ravi, "GoEmotions: A Dataset of Fine-Grained Emotions", *Proc. 58th ACL*, 2020, 4040–4054. DOI: 10.18653/v1/2020.acl-main.372.

[2] M. T. Ribeiro, T. Wu, C. Guestrin, S. Singh, "Beyond Accuracy: Behavioral Testing of NLP Models with CheckList", *Proc. 58th ACL*, 2020, 4902–4912. DOI: 10.18653/v1/2020.acl-main.442.

[3] P. V. Iaroshenko, N. V. Loukachevitch, "RusEmoLex: Russian Emotion Lexicon", *Russian Journal of Linguistics*, **29**:3 (2025), 659–687. DOI: 10.22363/2687-0088-44439.

Received on November 6, 2026