DOI: 10.15514/ISPRAS-2026-38(3)-53

Emotion Recognition Capabilities of Large Language Models: A Comparative Analysis

[1,2] E.S. Diatlinko, ORCID: 0009-0008-1514-285X <diatlinko@ispras.ru>
[1] M.D. Pavlov, ORCID: 0009-0002-6232-9875 <m.pavlov@ispras.ru>
[3] S.T. Tigranyan, ORCID: 0000-0003-1536-9954 <shtigranyan@ispras.ru>
[1] A.A. Avetisyan, ORCID: 0000-0002-7066-6954 <a.a.avetisyan@ispras.ru>

[1] Ivannikov Institute for System Programming of the Russian Academy of Sciences,
25, Alexander Solzhenitsyn st., Moscow, 109004, Russia.

[2] Lomonosov Moscow State University,
GSP-1, Leninskie Gory, Moscow, 119991, Russia.

[3] Russian-Armenian (Slavonic) University,
123, Hovsep Emin st., Yerevan, 0051, Armenia.

Abstract. Large language models (LLMs) are increasingly integrated into conversational systems, where understanding emotional cues is essential for maintaining coherent, engaging, and safe interactions. This study evaluates how effectively modern instruction-tuned large language models (LLMs) can recognize emotions from text only without task-specific fine-tuning. We benchmark multiple open-weight LLM families (<15B parameters) across four prompting strategies – Baseline, Context, Few-shot, and Context+Few-shot – on two English ERC benchmarks (IEMOCAP, MELD) and one Russian dataset (RESD). We find that the optimal prompting strategy is dataset-dependent: semantically redundant data such as IEMOCAP benefits most from few-shot demonstrations (best 73.3% weighted F1-score (WF1) with Context+Few-shot), whereas MELD gains primarily from incorporating dialogue history (best 60.3% WF1 with Context). Robustness experiments show that LLMs are largely insensitive to reordering few-shot examples, but performance degrades substantially when the label space is corrupted, indicating that coherent labels space matters more than order of examples or their ground truths. Cross-lingual evaluation reveals a notable drop on Russian RESD (best 45.8% WF1), highlighting a persistent gap between English and Russian affect understanding in current LLMs. Overall, non-finetuned LLMs serve as strong prompt-only baselines for ERC, yet remain clearly behind specialized supervised systems.

Keywords: large language models; emotion recognition; robustness; few-shot learning; emotion understanding.

For citation: Diatlinko E.S., Pavlov M.D., Tigranyan S.T., Avetisyan A.A. Emotion Recognition Capabilities of Large Language Models: A Comparative Analysis. Trudy ISP RAN/Proc. ISP RAS, vol. 38, issue 3, part 4, 2026, pp. 157-174. DOI: 10.15514/ISPRAS-2026-38(3)-53.

Возможности распознавания эмоций у больших языковых моделей: сравнительный анализ

[1,2] Е.С. Дятлинко, ORCID: 0009-0008-1514-285X <diatlinko@ispras.ru>
[1] М.Д. Павлов, ORCID: 0009-0002-6232-9875 <m.pavlov@ispras.ru >
[3] Ш.Т. Тигранян, ORCID: 0000-0003-1536-9954 <shtigranyan@ispras.ru>
[1] А.А. Аветисян, ORCID: 0000-0002-7066-6954 <a.a.avetisyan@ispras.ru>

[1] Институт системного программирования им. В.П. Иванникова РАН,
Россия, 109004, г. Москва, ул. А. Солженицына, д. 25.

[2] Московский государственный университет имени М.В. Ломоносова,
Россия, 119991, Москва, Ленинские горы, д. 1.

[3] Российско-Армянский (Славянский) университет,
123, ул. Овсепа Эмина, Ереван, 0051, Армения.

Аннотация. Большие языковые модели (LLMs) всё чаще интегрируются в диалоговые системы, где понимание эмоциональных сигналов является ключевым для поддержания связных, увлекательных и безопасных взаимодействий. В этом исследовании оценивается, насколько эффективно современные инструктивные LLM могут распознавать эмоции только по тексту без тонкой настройки под конкретную задачу. Мы тестируем несколько семейств LLM с открытым весом (количество параметров <15B) по четырем стратегиям запроса – базовая, контекстная, на примерах и комбинированная стратегия, объединяющая контекстные данные и примеры с разметкой – на основе двух английских наборов данных (IEMOCAP, MELD) и одного набора данных на русском языке (RESD). Результаты экспериментов показывают, что оптимальная стратегия поиска подсказок зависит от набора данных: семантически избыточные данные, такие как IEMOCAP, больше всего выигрывают от демонстрации нескольких кадров, в то время как MELD выигрывает в основном от включения истории диалога (лучшее значение 60.3% WF1 с контекстом). Эксперименты с надежностью показывают, что LLM-системы в значительной степени нечувствительны к изменению порядка нескольких примеров, но производительность существенно снижается при повреждении пространства меток, что указывает на то, что согласованное пространство меток имеет большее значение, чем порядок примеров или их правильных классов. Кросс-языковая оценка выявляет заметное снижение качества на русскоязычном RESD, что подчёркивает устойчивый разрыв между пониманием эмоций на английском и русском языках в современных LLM. Полученные выводы свидетельствуют о том, что LLM без дообучения выступают в качестве сильных базовых решений, основанных только на промптах, для распознавания эмоций, однако всё ещё заметно уступают специализированным системам, обученным с учителем.

Ключевые слова: большие языковые модели; распознавание эмоций; устойчивость; обучение с малым числом примеров; понимание эмоций.

Для цитирования: Дятлинко Е.С., Павлов М.Д., Тигранян Ш.Т., Аветисян А.А. Возможности распознавания эмоций у больших языковых моделей: сравнительный анализ. Труды ИСП РАН, том 38, вып. 3, часть 4, 2026 г., стр. 157–174 (на английском языке). DOI: 10.15514/ISPRAS–2026–38(3)–53.

1. Introduction

Large language models (LLMs) have become widely adopted across various domains, and their role in human-computer interaction continues to expand. As these systems increasingly participate in everyday communication, it is important to assess whether they are capable of interpreting emotional cues in an accurate and stable way. Emotion recognition from text provides a natural framework for evaluating this capability, as it requires models to infer affective states based not only on lexical content, but also on contextual and pragmatic cues.

In recent years, users have begun to engage with LLMs in ways that resemble interpersonal communication, often attributing social agency or personality-like qualities to these systems. This shift increases the importance of reliable emotion understanding, as misinterpretation of affective

signals can lead to misunderstandings, reduced user trust, or even emotional harm. Moreover, robust emotion recognition supports more adaptive, context-sensitive responses, enabling models to provide appropriate feedback in applications such as mental health support, education, and collaborative decision-making. Therefore, evaluating and improving the emotional awareness of LLMs is a crucial step toward developing safer and more human-aligned conversational systems.

In this study, we investigate how effectively contemporary LLMs can perform emotion recognition without any task-specific fine-tuning. We evaluate multiple instruction-tuned models on two widely used English-language benchmarks, MELD and IEMOCAP, which focus on utterance-level emotion classification in conversational settings. Additionally, we assess cross-lingual generalization by examining performance on the Russian-language dataset RESD, where resources and prior work remain considerably more limited.

Beyond measuring classification performance, we also examine the robustness of LLMs in few-shot prompting scenarios. Specifically, we analyze the effect of perturbing the structure and order of in-context examples to determine how sensitive LLM predictions are to prompt configuration.

The key questions addressed in this work are therefore:

(1) Can LLMs reliably recognize emotions from text without fine-tuning?

(2) Do LLMs perform this task consistently under changes to prompt structure?

(3) How well do LLMs generalize to emotion recognition in Russian compared to English?

2. Related works

In the field of emotion understanding, several related tasks have been extensively studied, including emotion recognition (ER), emotion recognition in conversation (ERC), sentiment analysis, personal trait recognition, video captioning, and open-vocabulary emotion recognition.

In recent years, researchers have expanded beyond the traditional text focus by incorporating multiple modalities, including video, facial images, speech/audio, physiological signals, gestures, and full-body motion [1-2]. In this paper, however, we focus exclusively on the text modality for ER and ERC. This choice is motivated by two considerations. First, many real-world applications provide only textual content (e.g., chat logs, online forums, customer support messages, or privacy-constrained settings), where audio-visual signals are unavailable or unreliable; thus, it is important to understand the upper bound of emotion recognition from language alone. Second, the core objective of this work is to evaluate prompted large language models, which are primarily designed and optimized for textual inference. By restricting the input to text, we isolate the contribution of linguistic cues and directly assess how well LLMs can infer emotional states without leveraging nonverbal information such as tone, prosody, or facial expressions.

Although state-of-the-art results in emotion recognition are still predominantly achieved by task-specific models designed and fine-tuned explicitly for emotion-related objectives, rather than by large language models (LLMs) trained primarily for text generation, recent studies have increasingly explored how effectively LLMs can perform emotion classification tasks [3].

Existing approaches for leveraging LLMs in emotion classification can be broadly categorized into three main directions. The first and most common involves alignment-based fine-tuning, typically using Low-Rank Adaptation (LoRA) [4] or similar parameter-efficient techniques. Notably, fine-tuning is not always performed directly on the emotion recognition (ER) task itself. For instance, in InstructERC[5], the model is fine-tuned on two auxiliary tasks: speaker identification and future emotion prediction, where the model predicts the emotional state of the speaker in subsequent utterances. However, fine-tuning large language models remains computationally expensive and resource-intensive. We evaluate exclusively non-fine-tuned (frozen) LLMs, treating them as black-box models where internal weights are hidden and only the resulting outputs are accessible for examination.

The second major direction focuses on prompt engineering. In these approaches, only textual information is used, but the prompt is augmented with additional contextual or descriptive cues derived from the dataset or external sources. For instance, LaERC-S [6] extracts knowledge about participants' mental states, behaviours, and personality traits via preliminary questions to the model; the resulting information is then concatenated with the main query to enhance emotion prediction. Similarly, BiosERC [7] incorporates speaker biography information directly into the prompt, allowing the model to leverage background knowledge about the speaker when interpreting emotional cues.

Many studies, such as AER-LLM [8], DialogueLLM [9], InstructERC [5], LaERC-S [6], and Beyond Silent Letter [10], adopt a strategy in which the historical dialogue context (i.e., preceding utterances in a conversation) is provided alongside the target utterance. This approach aligns with findings in emotion recognition in conversation (ERC), where contextual cues often play a critical role in accurately interpreting emotions.

A widely used technique for large language models, few-shot prompting [11], has also been successfully applied to emotion recognition. In this paradigm, the model is first presented with several examples (pairs of input utterances and their corresponding correct emotional labels) before being asked to classify a new instance. The few-shot setting has been explored in works such as AER-LLM [8] and InstructERC [5], showing that LLMs can benefit significantly from in-context examples even without task-specific fine-tuning.

The third direction involves chain-of-thought (CoT) prompting [12], where LLMs generate intermediate reasoning steps before the final answer, mimicking human problem-solving. This approach improves performance on tasks requiring multi-step reasoning. In OmniVox [13], the model first describes acoustic features, then analyzes the segment step by step, and only after that classifies the emotion. In [14], reasoning questions are generated by an LLM and then answered sequentially by the same model to reach the final prediction. These studies show that explicit reasoning guidance can enhance LLM performance on ER and ERC tasks.

The studies mentioned above evaluate their approaches on a limited set of models, which is a notable limitation. The experiments typically involve LLaMA [15] or Vicuna [16] families, while some works also include closed-source models such as GPT-3 [11].

Recent studies have shown that few-shot in-context learning can be highly unstable with respect to various factors, including example order, example selection, input–label alignment, and label-space configuration. Zhao et al. [17] showed that for GPT-3 and GPT-2, accuracy can vary from near-random to nearly perfect solely by changing the order of examples in the prompt. The authors also showed that instability depends on class distribution within the few-shot context. Y. Lu et al. [18] further confirmed that even minor reordering can cause dramatic performance swings, and that effective example sequences are not transferable between models.

Finally, S. Min et al. in [19] found that the input-label pairing in examples is often not crucial – replacing labels with random ones has little effect – while the label space, input distribution, and prompt format play a much larger role.

These findings indicate that few-shot prompting is inherently unstable, and its robustness cannot be assumed. In our work, we examine this instability across models of different sizes using the emotion classification task as a test case.

Most research on emotion recognition has been conducted primarily in English. However, other languages exhibit their own linguistic and cultural specificities, which may influence how emotions are expressed and interpreted. Despite this, comparatively few studies have examined emotion recognition in these languages. In this work, we consider Russian as an example of a language for which research in this area remains limited. Existing studies for Russian mostly rely on non-generative models such as RuBERT [20] and its variants (RuBERT-tiny2, RuBERT-large), which are based on the BERT architecture. Emotion2vec [21] applies self-supervised learning,

incorporating the audio modality, and BLSP-Emo [22] performs LLM alignment for this task. In [23] authors explore zero-shot emotion recognition, but only evaluate a single LLM.

To our knowledge, no prior work has conducted a comprehensive evaluation of multiple multilingual LLMs on Russian emotion recognition without fine-tuning. Our study addresses this gap.

3. Methods

3.1 Datasets

Three datasets were used in this study: two English-language corpora, IEMOCAP [24] and MELD [25], and one Russian-language corpus, RESD [26].

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) [24] dataset is a multimodal corpus developed at the SAIL Laboratory of the University of Southern California (USC). It contains approximately 12 hours of audiovisual recordings, including video, speech audio, facial motion capture data, and corresponding text transcriptions. The corpus comprises five dyadic sessions in which professional actors perform both scripted and improvised dialogues designed to express a variety of emotional states. Following the standard evaluation setup used in prior studies [1], our experiments were conducted on Session 5, which contains 2,071 utterances. Consistent with earlier works [27], we merge the “excited” and “happy” categories, resulting in four target emotion classes: angry, sad, neutral, and happy.

The Multimodal Emotion Lines Dataset (MELD) [25] contains audio, video, and text modalities. It consists of more than 1,400 dialogues and 13,000 utterances extracted from the TV series Friends, involving multiple speakers in each conversation. Every utterance is annotated with one of seven emotion categories: anger, disgust, sadness, joy, neutral, surprise, and fear. Our experiments were conducted on the test split, which comprises 2,610 utterances. We perform emotion classification using all seven emotion classes.

Both MELD and IEMOCAP are English-language, dialogue-based datasets in which classification is performed at the utterance level within a conversational context.

The Russian dataset of Emotional Speech Dialogues (RESD) [26] was compiled from approximately 3.5 hours of live speech recorded by actors who expressed predefined emotions in dialogues lasting about 3 minutes each. The test split includes 280 Russian-language utterances, each labeled with one of seven emotion categories: fear, anger, happy, enthusiasm, neutral, disgust, and sadness. Unlike MELD and IEMOCAP, RESD is not conversational, its utterances are independent and not contextually linked within dialogues.

In all datasets we use only text modality.

3.2 Prompt-techniques

In our study, we employ several prompt construction strategies. The baseline strategy uses a standard system prompt for emotion classification. The context strategy extends this baseline by additionally providing several preceding utterances to the model, thereby incorporating conversational history. The few-shot strategy augments the prompt with illustrative examples of emotion classification to guide the model’s reasoning. Finally, the context + few-shot strategy combines both techniques, integrating contextual dialogue information along with example-based guidance.

3.2.1 Baseline

The Baseline setting serves as the simplest prompting strategy, where the model receives a single utterance from the dataset as a user message, accompanied by a task-specific system prompt defining the target emotion categories.

Each dataset uses a tailored prompt specifying its label set and concise emotion definitions, instructing the model to output a single emotion label without additional commentary.

The following system prompt is used in the baseline setting for the MELD dataset: *“You are an emotion classification assistant. Your task is to analyze the input text and classify the underlying emotion as one of the following seven categories: – Neutral: Default for ambiguous/mixed tones, factual statements, no detectable emotional tone or mixed/ambiguous cases. – Joy: Expressions of happiness, excitement, amusement, affection, or positive feelings. – Surprise: Reactions of astonishment, unexpectedness, disbelief, or shock (positive or negative). – Anger: Frustration, hostility, irritation, sarcasm, raised voice, or displeasure. – Sadness: Grief, hopelessness, disappointment, sorrow, regret, or apologetic expressions. – Fear: Anxiety, nervousness, insecurity, worry, or expressions of being threatened. – Disgust: Aversion, contempt, dislike, revulsion, or disapproval. Respond with only one of the seven emotion labels: Neutral, Joy, Surprise, Anger, Sadness, Fear, or Disgust. Do not provide explanations or additional commentary”.* Classify the utterance of conversation. Prompts for the other datasets are constructed in an analogous manner.

3.2.2 Context

Both MELD and IEMOCAP are dialogue-based datasets in which the emotional interpretation of an utterance depends on its conversational context. Therefore, we propose a Context approach that extends the baseline by incorporating a fixed number of preceding utterances, controlled by the context_window hyperparameter.

The system prompt is constructed as follows:

```
{system_prompt}{history_prefix}{history}
```

where:

- system_prompt is identical to the baseline version, except that the final line is modified from *"Classify the utterance of conversation."* to "*Consider conversation, then classify the last utterance of conversation."*
- history_prefix is a fixed text: *"This is a conversation:"*
- history is the concatenation of up to context_window preceding utterances. If fewer than context_window utterances are available, all previous utterances are used.

This modification allows the LLM to utilize contextual information and assess how prior dialogue influences the emotional tone of the final utterance.

3.2.3 Few-shot

The Few-shot approach includes a small set of input-output examples in the prompt, allowing the model to infer task patterns and generalise to emotion recognition without additional fine-tuning. Each example follows the template:

```
Utterance <utterance> has emotion label <emotion>
```

Before presenting the examples, we prepend a fixed prefix *“There are some examples”.* Following the strategy of InstructERC [5], examples are selected only from the training split.

Sentence embeddings are computed using SBERT [28], and the most similar examples to the target utterance are identified via cosine similarity. A set of the top_k most similar labelled examples is prepended to the prompt, where top_k specifies the number of demonstrations used for few-shot learning.

3.2.4 Context + Few-shot

The Context + Few-shot approach combines the two previous strategies – context and few-shot prompting. Similar to the context setup, the model receives the previous context_window utterances from the dialogue, along with the target utterance. In addition, as in the few-shot setting, the prompt includes several demonstration examples.

Each example contains both the target utterance and its preceding context, ensuring that the examples and the final query share a consistent conversational format. This consistency helps the model better align its reasoning and output style with the examples provided.

The system prompt is constructed using the following template:

```
{system_prompt}{example_prefix}{example}{history_prefix}{history}
```

All components follow the same definitions as in the previous approaches.

4. Experiments

We evaluate models using Weighted F1-score (WF1), Unweighted Accuracy (UA), and Weighted Accuracy (WA). The WF1 accounts for class imbalance by balancing precision and recall. UA measures average performance across all classes equally, while WA reflects class frequency, providing a more realistic estimate of overall accuracy. Together, these metrics offer a comprehensive assessment of model performance.

4.1 Models

This study evaluates a set of open-weight large language models (LLMs) containing fewer than 15 billion parameters. The examined models belong to several representative families: Gemma 3 [29] (gemma-3-1b-it, gemma-3-4b-it, gemma-3-12b-it), LLaMA [15] (llama-3.1-8b-instruct, llama-3.2-1b, llama-3.2-3b, llama-3.2-11b-vision-instruct), Mistral [30] (mistral-7b-instruct-v0.2, mistral-13b-instruct-v0.2), Phi 3 [31] (phi-3.5-mini-instruct), Qwen 2.5 [32] (qwen2.5-1.5b-instruct, qwen2.5-3b-instruct, qwen2.5-7b-instruct, qwen2.5-14b-instruct), and Vicuna [16] (vicuna-7b-v1.5, vicuna-13b-v1.5).

Only instruction-tuned variants of these models were considered. Such models are specifically optimized for instruction-following and natural language understanding tasks, making them more suitable for zero-shot and few-shot classification-style prompting compared to their base counterparts.

All models were evaluated on the IEMOCAP and MELD datasets, whereas models supporting multiple languages (see Table 1 for details) were additionally tested on the RESD dataset to assess cross-lingual emotion recognition performance.

4.2 Experiments in English

The experiments were conducted in accordance with the prompting techniques described in the previous section. For both the MELD and IEMOCAP datasets, all four prompting configurations were evaluated: Baseline, Context, Few-shot, and Context + Few-shot.

For each model, the optimal values of context_window (the number of preceding utterances) and top_k (the number of few-shot examples) were determined empirically. The tables report the best performance metrics obtained for each model under each prompting strategy.

To retrieve semantically relevant examples for the few-shot setting, we employed the sentence-transformers [28] model all-MiniLM-L6-v2. This model maps sentences and paragraphs into a 384-dimensional dense vector space and is commonly used for semantic similarity, clustering, and information retrieval tasks.

4.2.1 Results on IEMOCAP

The baseline evaluation based on the IEMOCAP is presented in Table 2. LLaMA-3.1-8B achieves the highest performance among all models.

Table 1. Description of the utilized models.

Model	Size (B)	Model family	RU support
google/gemma-3-1b-it	1	gemma 3	✓
google/gemma-3-4b-it	4	gemma 3	✓
google/gemma-3-12b-it	12	gemma 3	✓
meta-llama/Llama-3.1-8B-Instruct	8	Llama	✓
meta-llama/Llama-3.2-1B-Instruct	1	Llama	✓
meta-llama/Llama-3.2-3B-Instruct	3	Llama	✓
meta-llama/Llama-3.2-11B-Vision-Instruct	11	Llama	×
mistralai/Mistral-7B-Instruct-v0.2	7	mistralai	×
microsoft/Phi-3.5-mini-instruct	4	Phi-3	✓
Qwen/Qwen2.5-1.5B-Instruct	1.5	Qwen2.5	✓
Qwen/Qwen2.5-3B-Instruct	3	Qwen2.5	✓
Qwen/Qwen2.5-7B-Instruct	7	Qwen2.5	✓
Qwen/Qwen2.5-14B-Instruct	14	Qwen2.5	✓
lmsys/vicuna-7b-v1.5	7	vicuna	×
lmsys/vicuna-13b-v1.5	13	vicuna	×

*Table 2. Baseline performance of various instruction-tuned language models on the IEMOCAP dataset. Results are reported in terms of metrics (%) weighted F1-score (WF1), unweighted accuracy (UA), and weighted accuracy (WA). The best metric values within each model family are highlighted in **bold**, for each metric, the best result across all models is underlined.*

Model	WF1	UA	WA
gemma-3-1b-it	41.4	42.5	46.1
gemma-3-4b-it	47.0	46.2	48.7
gemma-3-12b-it	**48.0**	**49.4**	**49.5**
Llama-3.1-8B-Instruct	**51.4**	**52.5**	**52.3**
Llama-3.2-1B-Instruct	15.0	21.0	26.3
Llama-3.2-3B-Instruct	44.1	46.7	48.4
Llama-3.2-11B-Instruct	50.2	51.3	51.3
Mistral-7B-Instruct-v0.2	44.0	46.8	45.7
Phi-3.5-mini-instruct	44.4	46.9	46.4
Qwen2.5-1.5B-Instruct	**50.6**	**51.4**	51.5
qwen-3b-instruct	36.9	43.0	41.0
qwen-7b-instruct	49.5	50.7	51.6
qwen-14b-instruct	49.9	50.9	**52.3**
vicuna-7b-v1.5	**50.0**	**50.6**	**47.9**
vicuna-13b	42.7	46.6	43.7

When comparing different prompting strategies on the IEMOCAP dataset (Table 3), the Context + Few-shot configuration yields the best overall results. At the same time, the Few-shot setup alone consistently surpasses the Context-only approach. This pattern likely stems from the nature of IEMOCAP, which contains many semantically similar utterances; consequently, few-shot demonstrations provide more effective emotional cues than extended dialogue context.

The best results are achieved with a WF1 of 0.733, a UA of 0.731, and a WA of 0.747 by Qwen2.5-14B-Instruct under the Context + Few-shot setup.

Within the LLaMA model family LLaMA-3.1-8B (medium-size variant) achieved the strongest results. In contrast, in the Qwen, Gemma and Vicuna families, the largest models (Qwen2.5-14B, Gemma-3-12B and Vicuna-13B) demonstrated the best performance.

*Table 3. Comparison of prompting strategies (Context, Few-shot, and Context + Few-shot) on the IEMOCAP dataset, reported in terms of metrics (%) WF1, UA, and WA. The best metric values within each model are highlighted in **bold**, for each metric, the best result across all models is underlined.*

Model	Method	WF1	UA	WA
gemma-3-1b-it	context	52.8	52.9	57.4
	few-shot	53.9	53.6	55.2
	context few-shot	**58.1**	**57.7**	**62.2**
gemma-3-4b-it	context	56.0	55.8	61.1
	few-shot	62.6	62.2	64.8
	context few-shot	**68.4**	**68.0**	**70.3**
gemma-3-12b-it	context	56.3	56.6	59.8
	few-shot	65.5	65.3	66.6
	context few-shot	**72.0**	**71.7**	**72.6**
Llama-3.1-8B-Instruct	context	57.2	57.0	60.3
	few-shot	66.2	66.0	67.5
	context few-shot	**67.4**	**67.1**	**68.8**
Llama-3.2-1B-Instruct	context	6.8	19.9	25.1
	few-shot	25.8	30.5	34.3
	context few-shot	**37.9**	**34.5**	**35.1**
Llama-3.2-3B-Instruct	context	44.4	44.2	50.0
	few-shot	50.0	48.8	53.1
	context few-shot	**54.0**	**54.5**	**59.4**
Llama-3.2-11B-Instruct	context	56.6	56.4	60.2
	few-shot	65.8	65.4	67.0
	context few-shot	**67.1**	**66.9**	**68.8**
Mistral-7B-Instruct-v0.3	context	50.6	52.1	53.6
	few-shot	61.1	61.1	63.5
	context few-shot	**63.9**	**62.5**	**65.8**
Phi-3.5-mini-instruct	context	48.0	49.2	51.6
	few-shot	56.9	57.2	55.0
	context few-shot	**63.4**	**62.7**	**65.4**
Qwen2.5-1.5B-Instruct	context	48.5	49.3	49.9
	few-shot	**56.3**	55.8	**59.3**
	context few-shot	**56.3**	**56.0**	58.4
qwen-3b-instruct	context	34.7	41.6	37.5
	few-shot	51.0	53.0	52.2
	context few-shot	**63.9**	**64.0**	**63.4**
qwen-7b-instruct	context	53.8	54.6	54.9
	few-shot	67.1	66.9	67.2
	context few-shot	**70.3**	**70.2**	**70.8**
qwen-14b-instruct	context	59.1	59.6	61.9
	few-shot	64.2	64.1	65.3
	context few-shot	**73.3**	**73.1**	**74.7**
vicuna-7b-v1.6	context	50.8	52.2	47.2
	few-shot	56.9	57.2	55.0
	context few-shot	**59.4**	**59.1**	**56.8**
vicuna-13b	context	48.2	50.4	48.1
	few-shot	60.9	60.8	59.9
	context few-shot	**62.2**	**61.3**	**60.0**

4.2.2 Results on MELD

Across models, the baseline results (Table 4) show clear capacity effects within model families (e.g., Gemma improves markedly from 1B to 12B), but the strongest baseline model depends on the metric: Gemma-3-12B achieves the best baseline WF1 (0.588), Vicuna-13B the best baseline UA (0.602), and Qwen-14B the best baseline WA (0.508).

*Table 4. Baseline performance of various instruction-tuned language models on the MELD dataset. Results are reported in terms of metrics (%) WF1, UA, and WA. The best metric values within each model family are highlighted in **bold**, for each metric, the best result across all models is underlined.*

Model	WF1	UA	WA
gemma-3-1b-it	47.0	44.4	29.0
gemma-3-4b-it	52.2	50.6	41.8
gemma-3-12b-it	**58.8**	**57.9**	**46.4**
Llama-3.1-8B-Instruct	**57.8**	**58.7**	**39.8**
Llama-3.2-1B-Instruct	30.5	30.3	19.7
Llama-3.2-3B-Instruct	51.9	55.9	29.7
Llama-3.2-11B-Instruct	57.4	58.3	39.4
Mistral-7B-Instruct-v0.2	56.3	57.3	39.8
Phi-3.5-mini-instruct	57.2	57.0	40.1
Qwen2.5-1.5B-Instruct	43.4	42.9	33.4
qwen-3b-instruct	56.3	**57.2**	39.3
qwen-7b-instruct	**58.0**	56.9	49.2
qwen-14b-instruct	56.6	55.1	**50.8**
vicuna-7b-v1.5	53.7	55.4	35.5
vicuna-13b	**57.0**	**60.2**	**36.6**

When evaluating prompting strategies (Table 5), the overall best configuration is obtained by Qwen-14B with Context, which yields the highest WF1 (0.603) and WA (0.534) across all settings, while the best UA is achieved by Vicuna-13B with Context (0.613). Importantly, the summary of per-model winners reveals a consistent metric-dependent trend: WF1 and UA are most often maximized by Baseline or Few-shot prompting, whereas WA strongly favors context augmentation. This suggests a trade-off, where incorporating dialogue context primarily improves weighted accuracy (WA), whereas few-shot demonstrations are especially beneficial for smaller models and more frequently enhance WF1 and UA.

Overall, MELD appears highly context-dependent, and the most reliable way to improve WA is to incorporate preceding turns, while the optimal choice for WF1/UA remains model-specific, frequently favoring shorter baseline or few-shot prompts.

Across both datasets, the optimal prompting strategy is dataset-dependent. On IEMOCAP, where many utterances are semantically similar, few-shot demonstrations provide particularly informative emotional cues, leading to the strongest overall performance under the Context + Few-shot setup. In contrast, MELD is more context-dependent, and incorporating preceding turns yields the strongest performance. Beyond dataset effects, scaling trends are broadly consistent: larger models typically perform better within families (e.g., Gemma), though the top baseline model can vary by metric. Overall, these findings suggest that effective emotion recognition with LLM prompting requires aligning the prompting strategy with the dataset's structure, favoring demonstrations for IEMOCAP-like data and dialogue context for MELD-like conversations.

To identify which emotion classes are most frequently confused, we analyze the row-normalized confusion matrices (see Fig. 1). For IEMOCAP, most errors are driven by false positives for the neutral class, with the strongest confusion occurring between happy and neutral. For MELD, misclassifications into neutral are also present but less pronounced; the most frequent confusion is disgust being incorrectly predicted as anger.

*Table 5. Comparison of prompting strategies (Context, Few-shot, and Context + Few-shot) on the MELD dataset, reported in terms of metrics (%) WF1, UA, and WA. The best metric values within each model are highlighted in **bold**, for each metric, the best result across all models is underlined.*

Model	Method	WF1	UA	WA
gemma-3-1b-it	context	44.9	42.8	34.6
	few-shot	**48.5**	**46.9**	36.1
	context few-shot	44.9	42.6	**40.2**
gemma-3-4b-it	context	**48.0**	**46.5**	**47.5**
	few-shot	43.2	42.0	43.4
	context few-shot	40.8	40.2	46.3
gemma-3-12b-it	context	**58.8**	**56.6**	**53.2**
	few-shot	54.8	53.0	49.3
	context few-shot	56.3	54.4	50.3
Llama-3.1-8B-Instruct	context	**51.9**	50.2	43.8
	few-shot	51.8	**50.6**	41.7
	context few-shot	45.6	43.7	**45.2**
Llama-3.2-1B-Instruct	context	2.3	4.4	16.6
	few-shot	**36.1**	**33.6**	**22.7**
	context few-shot	14.4	11.7	16.9
Llama-3.2-3B-Instruct	context	53.0	51.6	36.6
	few-shot	**53.7**	**55.3**	34.5
	context few-shot	51.0	48.7	**38.2**
Llama-3.2-11B-Instruct	context	50.3	48.5	44.0
	few-shot	**51.1**	**49.8**	41.3
	context few-shot	44.6	43.1	**46.2**
Mistral-7B-Instruct-v0.3	context	55.8	54.8	**45.9**
	few-shot	**57.3**	**57.3**	44.7
	context few-shot	55.6	53.5	44.6
Phi-3.5-mini-instruct	context	**55.6**	**53.7**	**43.5**
	few-shot	53.7	52.5	42.7
	context few-shot	53.0	50.7	41.9
Qwen2.5-1.5B-Instruct	context	**50.9**	**52.4**	35.9
	few-shot	40.1	39.8	38.4
	context few-shot	49.3	48.0	**40.5**
qwen-3b-instruct	context	51.7	55.8	34.4
	few-shot	**57.3**	**57.9**	**42.2**
	context few-shot	51.8	52.5	40.2
qwen-7b-instruct	context	**57.1**	**55.6**	**49.7**
	few-shot	49.8	48.6	47.4
	context few-shot	52.6	51.4	47.3
qwen-14b-instruct	context	**60.3**	**59.0**	**53.4**
	few-shot	53.5	52.0	49.9
	context few-shot	56.4	55.1	51.5
vicuna-7b-v1.6	context	**52.5**	**51.8**	38.0
	few-shot	49.7	50.6	31.6
	context few-shot	50.8	49.3	**41.2**
vicuna-13b	context	**58.8**	**61.3**	**39.2**
	few-shot	56.2	60.1	34.7
	context few-shot	54.4	55.8	31.1

Table 6 indicates that LLMs constitute a viable ERC solution primarily as a strong prompt-only baseline, but their performance remains clearly inferior to purpose-built models. Our prompt-based approach achieves 60.3 WF1 on MELD and 73.3 WF1 on IEMOCAP, demonstrating that instruction-tuned LLMs capture meaningful affective cues without task-specific training. However, the gap to specialized ERC systems is substantial: on MELD, prior methods reach roughly 69–72 WF1, and on IEMOCAP the best models achieve 81–85 WF1, leaving our approach about 8–12 points behind on both datasets. Overall, these results suggest that prompt-only LLMs mainly reflect raw general-purpose capabilities and should be viewed as competitive baselines rather than replacements for state-of-the-art ERC architectures, which better exploit dialogue structure and supervised adaptation.

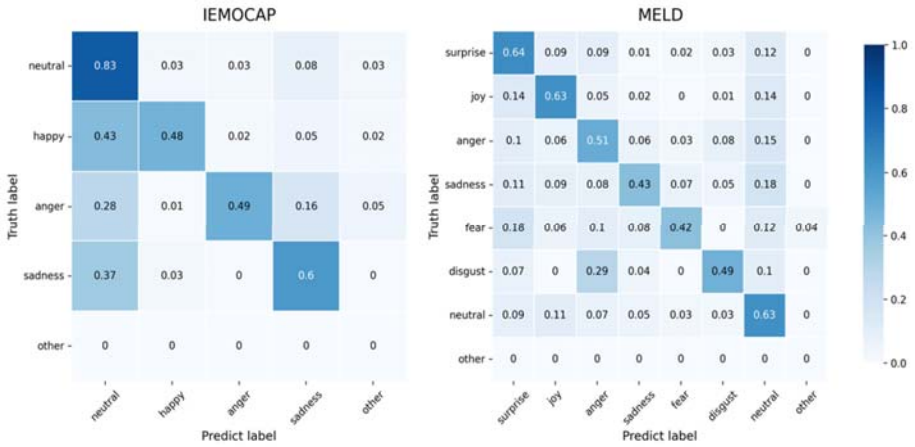

Fig. 1. Row-normalized confusion matrices for emotion classification on IEMOCAP (left) and MELD (right). Each cell shows the proportion of predictions for a given ground-truth class (rows) assigned to each predicted label (columns). The "other" class denotes cases where the model produced an output outside the predefined set of emotion labels.

Table 6. Comparison with prior methods on MELD and IEMOCAP in terms of WF1 (%).

Method	MELD	IEMOCAP	Type
DialogXL [33]	62.41	73.02	transformer-based
ELR-GNN [34]	68.70	-	graph-based
DAG-ERC [35]	63.65	78.08	
DialogueGCN [36]	58.10	71.58	
COGMEN [37]	-	81.55	
CORRECT [38]	-	84.64	
OmniVox [13]	62.80	-	finetuned llm-based
DialogueLLM [9]	71.90	-	
InstructERC [5]	69.15	-	
Laerc-s [6]	69.27	-	
Bioserc [7]	69.83	-	
aer-llm [8]	-	57.90	
Ours	60.30	73.30	llm-based

4.3 Experiments in Russian

Experiments on Russian language emotion recognition were conducted using the RESD dataset. As this corpus contains isolated utterances rather than dialogues, contextual information could not be incorporated into the prompts, only the Baseline and Few-shot configurations were evaluated.

To ensure linguistic consistency, the experiments were limited to models explicitly supporting the Russian language. Few-shot examples were selected using FRIDA [39], a Russian-adapted sentence-transformer developed by AI Forever, which facilitated more accurate semantic matching within the language.

The Few-shot strategy substantially outperformed the baseline, improving all evaluation metrics by more than 10% (see Table 7 for details). The best performance was achieved by Qwen2.5-7B-Instruct, reaching a weighted F1 of 0.458, a UA of 0.450, and a WA of 0.445. Within model families, the medium-sized variants Qwen2.5-7B and LLaMA-3.1-8B showed the strongest performance, while for Gemma, the largest model (Gemma-3-12B) achieved the best results. However, the best results on this dataset were reported in [26], the authors achieved an average accuracy ranging from 72% to 81%.

*Table 7. Comparison of Baseline and Few-shot on RESD, reported in terms of metrics (%) WF1, UA, and WA. The best metric values within each model are highlighted in **bold**, for each metric, the best result across all models is underlined.*

Model	Method	WF1	UA	WA
gemma-3-1b-it	baseline	19.3	21.1	21.1
	few shot	**26.1**	**27.5**	**28.1**
gemma-3-4b-it	baseline	27.6	31.4	31.3
	few shot	**38.0**	**38.9**	**38.5**
gemma-3-12b-it	baseline	27.9	30.4	30.4
	few shot	**38.5**	**39.3**	**39.4**
Llama-3.1-8B-Instruct	baseline	23.4	25.7	25.3
	few shot	**42.0**	**42.1**	**42.4**
Llama-3.2-1B-Instruct	baseline	7.6	11.1	12.9
	few shot	**18.6**	**20.4**	**21.9**
Llama-3.2-3B-Instruct	baseline	11.7	17.9	18.5
	few shot	**25.2**	**27.5**	**26.8**
Phi-3.5-mini-instruct	baseline	19.7	22.9	22.7
	few shot	**35.7**	**35.4**	**35.5**
Qwen2.5-1.5B-Instruct	baseline	18.1	23.6	23.2
	few shot	**27.6**	**32.9**	**31.3**
qwen-3b-instruct	baseline	18.7	22.9	22.7
	few shot	**35.0**	**36.4**	**36.5**
qwen-7b-instruct	baseline	31.0	31.8	31.5
	few shot	**45.8**	**45.0**	**44.5**
qwen-14b-instruct	baseline	29.9	31.1	30.3
	few shot	**43.1**	**43.2**	**42.9**

Overall, the obtained scores remain significantly lower than those on English datasets, reflecting the current performance gap between English and Russian LLMs. This disparity highlights that emotion recognition in Russian remains an open challenge, as most LLM research and optimization still focus predominantly on English-language data.

4.4 Robustness to few-shot prompt changes

To assess robustness to changes in few-shot prompt structure, we evaluate four perturbations of the in-context examples: (1) shuffling example order, (2) reversing the order, (3) shuffling the provided ground-truth labels among examples, and (4) assigning fully random labels. Fig. 2 reports the change in weighted F1 (ΔWF1) relative to the original prompt. For interpretability, models are grouped by scale into small (1-3B), medium (4-8B), and large (>10B) parameter ranges.

Across both datasets, purely structural changes to the prompt are comparatively benign. Shuffling the example order produces only negligible performance shifts (generally within ~1%), and reversing the order results in slightly larger but still moderate drops (up to ~2.5%). This indicates that, for these instruction-tuned models, in-context classification is largely insensitive to the sequence of examples, suggesting that models mostly extract a task template rather than relying heavily on a specific positional arrangement.

More informative is the contrast between label-level perturbations. Shuffling labels among examples consistently degrades performance by a small-to-moderate margin (~1.5-2.3%), but the effect is still far weaker than fully random class assignments. When labels are randomized, performance collapses sharply (approximately -15% to -21% on IEMOCAP and -7% to -9% on MELD across model sizes). The key takeaway is that the decisive factor is not minor prompt formatting or even imperfect example – label pairing, but whether the prompt preserves a coherent label space. In other words, the models appear to rely primarily on which labels are available and how the task is framed, while the exact correspondence between a particular example and its label contributes comparatively less – an interpretation supported by the large gap between the mild impact of label shuffling and the severe failure under random labels.

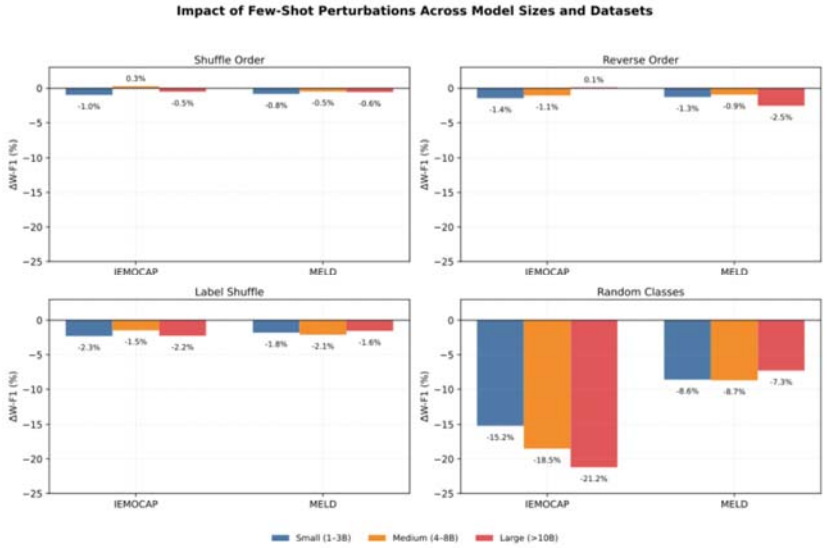

Fig. 2. Effect of four types of few-shot prompt perturbations on model performance (ΔWF1, %), grouped by model size and dataset.

Overall, the results suggest that modern LLMs are robust to few-shot prompt reformatting (reordering/shuffling), and only moderately sensitive to breaking example – label consistency, but they strongly depend on having a meaningful, non-adversarial label space. This implies that ensuring a stable and well-defined set of target classes is more critical for reliable few-shot classification than preserving the exact structure or ordering of the few-shot demonstrations.

5. Conclusion

This study presents a comprehensive evaluation of modern large language models (LLMs) in recognizing emotional states from text without task-specific fine-tuning. The analysis spans multiple model families across different parameter scales, various prompting strategies, and three datasets: two English-language (IEMOCAP, MELD) and one Russian-language (RESD).

Our experiments show that the optimal prompting strategy depends on the properties of the benchmark. In practice, strong performance is achieved when the prompt is matched to the dataset's structure – few-shot demonstrations are particularly helpful for IEMOCAP-style data, whereas MELD-like conversations benefit more from explicitly incorporating dialogue history.

Robustness analysis indicates that modern LLMs are largely insensitive to few-shot prompt reformatting (e.g., reordering or shuffling demonstrations) and only moderately sensitive to breaking example-label consistency. However, they strongly depend on a meaningful, non-adversarial label space. This suggests that maintaining a stable and well-defined set of target classes is more important for reliable few-shot classification than or ordering of the demonstrations or their labels.

For the Russian RESD benchmark, performance remains markedly below the English results, pointing to a persistent gap in multilingual affect understanding. This suggests that, despite nominal multilingual capabilities, robust emotion recognition in Russian is still underdeveloped – largely reflecting the English-centric focus of current model development and training data.

Taken together, these findings position non-finetuned LLMs as a practical option mainly in the role of a prompt-only baseline. While they exhibit meaningful zero-/few-shot capabilities, they still lag behind specialized ERC systems that leverage supervised adaptation and richer modeling of conversational structure, and therefore should be treated as competitive reference points rather than substitutes for state-of-the-art architectures.

References

[1]. Wu C., Cai Y., Liu Y., Zhu P., Xue Y., Gong Z., et al. Multimodal emotion recognition in conversations: A survey of methods, trends, challenges and prospects. arXiv preprint arXiv:2505.20511, 2025.

[2]. Kalateh S., Estrada-Jimenez L. A., et al. A systematic review on multimodal emotion recognition: building blocks, current state, applications, and challenges. IEEE Access, 2024, 12, 103976-104019.

[3]. Peng L., Zhang Z., Pang T., Han J., Zhao H., Chen H., Schuller B.W. Customising general large language models for specialised emotion recognition tasks. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP, April, 2024, IEEE, pp. 11326-11330.

[4]. Hu E.J., Shen Y., Wallis P., Allen-Zhu Z., Li Y., et al. Lora: low-rank adaptation of large language models. ICLR, 2022, 1(2):3.

[5]. Lei S., Dong G., Wang X., Wang K., Qiao R., Wang S. Instructerc: reforming emotion recognition in conversation with multi-task retrieval-augmented large language models. arXiv preprint arXiv:2309.11911, 2023.

[6]. Fu Y., Wu J., Wang Z., Zhang M., Shan L., Wu Y., Li B. Laerc-s: improving llm-based emotion recognition in conversation with speaker characteristics. arXiv preprint arXiv:2403.07260, 2024.

[7]. Xue J., Nguyen M.-P., Matheny B., Nguyen L.-M. Bioserc: integrating biography speakers supported by llms for erc tasks. In International Conference on Artificial Neural Networks, Springer, 2024, pp. 277-292.

[8]. Hong X., Gong Y., Sethu V., Dang T. Aer-llm: ambiguity-aware emotion recognition leveraging large language models. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2025, pp. 1-5.

[9]. Zhang Y., Wang M., Wu Y., Tiwari P., Li Q., Wang B., Qin J. Dialoguellm: context and emotion knowledge-tuned large language models for emotion recognition in conversations. arXiv preprint arXiv:2310.11374, 2023.

[10]. Z. Wu, Z. Gong, L. Ai, P. Shi, K. Donbekci, and J. Hirschberg. Beyond silent letters: amplifying llms in emotion recognition with vocal nuances. arXiv preprint arXiv:2407.21315, 2024.

[11]. T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.

[12]. J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837, 2022.

[13]. J. Murzaku and O. Rambow. Omnivox: zero-shot emotion recognition with omni-llms. arXiv preprint arXiv:2503.21480, 2025.

[14]. K. Hama, A. Otsuka, and R. Ishii. Emotion recognition in conversation with multi-step prompting using large language model. In International Conference on Human-Computer Interaction, pages 338–346. Springer, 2024.

[15]. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

[16]. W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, et al. Vicuna: an open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See https://vicuna. lmsys. org (accessed 14 April 2023), 2(3):6, 2023.

[17]. Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh. Calibrate before use: improving few-shot performance of language models. In International conference on machine learning, pages 12697–12706. PMLR, 2021.

[18]. Y. Lu, M. Bartolo, A. Moore, S. Riedel, and P. Stenetorp. Fantastically ordered prompts and where to find them: overcoming few-shot prompt order sensitivity. arXiv preprint arXiv:2104.08786, 2021.

[19]. S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer. Rethinking the role of demonstrations: what makes in-context learning work? arXiv preprint arXiv:2202.12837, 2022.

[20]. D. Zmitrovich, A. Abramov, A. Kalmykov, M. Tikhonova, E. Taktasheva, D. Astafurov, M. Baushenko, A. Snegirev, T. Shavrina, S. Markov, V. Mikhailov, and A. Fenogenova. A family of pretrained transformer language models for russian, 2023. arXiv: 2309.10931 [cs.CL].

[21]. Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen. Emotion2vec: self-supervised pre-training for speech emotion representation. arXiv preprint arXiv:2312.15185, 2023.

[22]. C. Wang, M. Liao, Z. Huang, J. Wu, C. Zong, and J. Zhang. Blsp-emo: towards empathetic large speech-language models. arXiv preprint arXiv:2406.03872, 2024.

[23]. H. Zou, F. Lv, D. Zheng, E. S. Chng, and D. Rajan. Large language models meet contrastive learning: zero-shot emotion recognition across languages. arXiv preprint arXiv:2503.21806, 2025.

[24]. C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. Iemocap: interactive emotional dyadic motion capture database. Language resources and evaluation, 42(4):335–359, 2008.

[25]. S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea. Meld: a multimodal multi-party dataset for emotion recognition in conversations. arXiv preprint arXiv:1810.02508, 2018.

[26]. Н. Д. Артем Аментес Илья Лубенец. Открытая библиотека искусственного интеллекта для анализа и выявления эмоциональных оттенков речи человека. https://huggingface.com/aniemore/Aniemore, 2022.

[27]. Mai, J., Xing, X., Li, Y., & Xu, X. (2025). Chain-of-Thought Distillation with Fine-Grained Acoustic Cues for Speech Emotion Recognition. In Proc. Interspeech 2025 (pp. 5438-5442).

[28]. N. Reimers and I. Gurevych. Sentence-bert: sentence embeddings using siamese bert-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Nov. 2019. URL: https://arxiv.org/abs/1908.10084

[29]. G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, et al. Gemma 3 technical report. arXiv preprint arXiv:2503.19786, 2025.

[30]. A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b, 2023. arXiv: 2310.06825 [cs.CL]. URL: https://arxiv.org/abs/2310.06825.

[31]. Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., ... & Zhang, Y. Phi-3 technical report: a highly capable language model locally on your phone, 2024. arXiv: 2404.14219 [cs.CL]. URL: https://arxiv.org/abs/2404.14219.

[32]. Qwen, : A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu. Qwen2.5 technical report, 2025. eprint: 2412.15115.

[33]. W. Shen, J. Chen, X. Quan, and Z. Xie. Dialogxl: all-in-one xlnet for multi-party conversation emotion recognition. In Proceedings of the AAAI conference on artificial intelligence, volume 35 of number 15, pages 13789–13797, 2021.
[34]. Y. Shou, W. Ai, J. Du, T. Meng, H. Liu, and N. Yin. Efficient long-distance latent relation-aware graph neural network for multi-modal emotion recognition in conversations. arXiv preprint arXiv:2407.00119, 2024.
[35]. W. Shen, S. Wu, Y. Yang, and X. Quan. Directed acyclic graph network for conversational emotion recognition. arXiv preprint arXiv:2105.12907, 2021.
[36]. D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh. Dialoguegcn: a graph convolutional neural network for emotion recognition in conversation. arXiv preprint arXiv:1908.11540, 2019.
[37]. A. Joshi, A. Bhat, A. Jain, A. Singh, and A. Modi. Cogmen: contextualized gnn based multimodal emotion recognition. In Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies, pages 4148–4164, 2022.
[38]. C.-V. T. Nguyen, A.-T. Mai, T.-S. Le, H.-D. Kieu, and D.-T. Le. Conversation understanding using relational temporal graph neural networks with auxiliary cross-modality interaction. arXiv preprint arXiv:2311.04507, 2023.
[39]. AI-Forever. FRIDA. Hugging Face. Available at: https://huggingface.co/ai-forever/FRIDA (accessed April 21, 2026).

Информация об авторах / Information about authors

Екатерина Сергеевна ДЯТЛИНКО – магистрант ВМК МГУ, старший лаборант Института системного программирования. Сфера научных интересов: обработка естественного языка (NLP), большие языковые модели (LLM), мультимодальное обучение, компьютерное зрение.

Ekaterina Sergeevna DIATLINKO – master student at the CMC faculty of Lomonosov Moscow State University, senior research assistant at the Institute of System Programming of the RAS. Research interests: natural language processing (NLP), Large Language Models (LLM), multimodal learning, computer vision.

Матвей Дмитриевич ПАВЛОВ – лаборант Института системного программирования с 2025 года. Сфера научных интересов: машинное обучение и глубокое обучение, обработка естественного языка, мультимодальное обучение, компьютерное зрение, системное программирование и компиляторы.

Matvey Dmitrievich PAVLOV – laboratory assistant at the Institute for System Programming since 2025. Research interests: machine learning and deep learning (ML/DL), natural language processing (NLP), systems programming, and compilers.

Шагане Тиграновна ТИГРАНЯН – аспирант Российско-Армянского университета. Сфера научных интересов: распознавание эмоций, мультимодальное обучение, анализ сигналов.

Shahane Tigranovna TIGRANYAN is a postgraduate student at the Russian-Armenian University. Research interests: emotion recognition, multimodal learning, and signal processing.

Арам Арутюнович АВЕТИСЯН работает в Институте системного программирования. Сфера научных интересов: применение нейронных сетей для анализа медицинских данных, определения эмоций, федеративное обучение.

Aram Arutyunovich AVETISYAN is working in the Institute for System Programming of RAS. His research interests include deep learning in medical applications, emotion recognition, and federated learning.