DOI: 10.14529/jsfi250303

LLM for Semantic Role Labeling of Emotion Predicates in Russian

Ivan V. Smirnov[1], *Daniil S. Larionov*[1], *Elena N. Nikitina*[1], *Grigory A. Kazachonok*[2]

Semantic role labeling (SRL) for morphologically rich languages, such as Russian, faces significant challenges due to complex case marking systems, free word order, and limited annotated resources. These challenges are particularly acute for emotion predicates, which require specialized linguistic expertise to capture distinctions between roles denoting those who feel, causes and objects of feelings. We propose a novel approach that leverages large language models to address SRL for Russian emotion predicates through few-shot in-context learning combined with predicate-specific instructions. Our method was evaluated on a manually annotated dataset of 169 sentences containing six emotion predicate groups extracted from Russian social media texts. We compared three state-of-the-art LLMs (Claude 3.7 Sonnet, GPT-5 Mini, and DeepSeek V3) against a RuELECTRA-based trained sequence labelling baseline using both exact and partial matching criteria. Claude 3.7 achieved the highest performance with 74.85% F1 score on partial matching, substantially outperforming the baseline (22.67%). For general predicates on FrameBank, our adapted method with GPT-5 Mini reached 85.0% F1 compared to the previous state-of-the-art of 80.1%. The LLM-based approach successfully handles complex linguistic phenomena, including syntactic zeros and multi-word arguments, while requiring minimal manually annotated training data. We demonstrate that LLM-based methods can significantly advance SRL for Russian by reducing dependency on large-scale annotated corpora while achieving competitive performance.

Keywords: semantic role labeling, llm, russian language, deep learning, neural networks.

Introduction

Semantic Role Labeling (SRL) is a fundamental task in natural language processing that aims to identify the semantic relationships between predicates and their arguments in sentences [11]. Traditional approaches to SRL have largely relied on supervised learning methods trained on carefully annotated corpora such as FrameBank [14]. However, for morphologically rich languages like Russian, the task presents significant challenges due to complex case marking systems, relatively free word order, and limited availability of annotated resources. The annotation scarcity problem is a significant bottleneck, particularly for complex semantic classes such as emotion predicates. Emotion predicates, which include verbs of fear and emotional attitude, and psychological states (such as “пугать-пугаться” (to frighten – to be frightened), “бояться” (to fear), “нравиться” (to like), “любить” (to love), etc.), present unique challenges due to their complex argument structures [17] and the subjective nature of emotional experiences, which specifically appears in syntactic zeros of experiencer (the argument denoting those who undergo emotions): *Пугает неопределенность* (Uncertainty frightens ∅); *вид устрашает...как в Чернобыле..* (The view frightens ∅ like in Chernobyl); *Рост тарифов ЖКХ не страшил бы так, если бы в городе создавались новые высокотехнологичные рабочие места* (The grow of utilities rates would not have frightened ∅ so much if the new high tech jobs had been created)[3] [16].

The annotation of emotion predicates requires specialized linguistic expertise to capture the subtle distinctions between different types of emotional roles, such as experiencers, stimuli, and

[1] Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, Moscow, Russian Federation

[2] Moscow Institute of Physics and Technology, Dolgoprudny, Russia

[3] Here and everywhere else examples are provided with original authors grammar.

targets of emotions [19], as well as their superficial expressions. See, for example, split of Causator role into two arguments: **Мэр** *удивил горожан* **гранитными бордюрами**, clause and infinitive arguments: *Владимир, а вам нравится,* **когда вас с кем сравнивают?***; боится* **лишние секунды потерять**. Traditional approaches rely heavily on manually curated training data, which is both expensive to produce and is limited in coverage. This creates a particular challenge for languages like Russian, where comprehensive emotional semantic resources are scarce compared to English.

Recent advances in Large Language Models (LLMs) have demonstrated remarkable capabilities across various natural language processing tasks through in-context learning and prompting-based approaches [12]. These models show particular promise for tasks that benefit from semantic understanding and structured reasoning, making them natural candidates for semantic role labeling applications.

We propose a novel approach that leverages the linguistic capabilities of LLMs to address the challenges of semantic role labeling in Russian, particularly for the domain of emotion predicates. Our method combines two key techniques: few-shot learning and instruction retrieval. Few-shot learning enables the model to generalize from a small number of examples, while instruction retrieval allows the system to access relevant linguistic knowledge and annotation guidelines dynamically.

Our contributions advance the state of semantic role labeling for Russian by demonstrating that LLM-based approaches can achieve competitive performance while reducing the dependency on large-scale manually annotated corpora. The proposed framework provides flexibility for different application scenarios: the expert-curated instructions and examples for high-precision and low-data cases in specialized domains, and the generic retrieval approach for broader coverage where more annotated data is available. We release the source code of our approach at `https://github.com/ru-nlp/llm-for-srl`.

The article is organized as follows. Section 1 reviews related work on LLM-based approaches to semantic role labeling, the FrameBank resource for Russian, and previous research on emotion predicates. Section 2 describes our methodology, including the construction of the Russian emotion predicates dataset, the configuration of three evaluated LLMs (Claude 3.7 Sonnet, GPT-5 Mini, and DeepSeek V3) and the RuELECTRA-SRL baseline, our few-shot prompting strategy with predicate-specific instructions, and the evaluation protocol with exact and partial matching criteria. We also describe the adaptation of our approach to general predicates on FrameBank. Section 3 presents experimental results for both emotion predicates and general semantic role labeling, with detailed analysis of model performance across different semantic roles and matching criteria. Section 4 discusses the computational and linguistic implications of our findings, examining how LLMs handle complex phenomena such as syntactic zeros, anaphoric references, multi-word arguments, and irregular constructions in social media texts. The Conclusion summarizes our results and outlines directions for future research.

1. Related Work

1.1. LLM for SRL

[8] contains a comprehensive survey on various methods and applications of semantic role labeling. Large language models have become the dominant paradigm for solving NLP tasks, and naturally, almost all of the recent approaches to SRL employ LLMs in some way or form.

Current approaches integrate LLMs into the SRL pipeline in several innovative ways. A common technique involves using the embeddings generated by an LLM as rich feature inputs for a downstream classifier, often enhanced with syntactic information such as dependency parses. This hybrid approach, exemplified by [11], combines deep semantic understanding with structural grammatical cues.

The current state-of-the-art results for both English and Chinese, as demonstrated by [12], are achieved through a more integrated method. Their model employs prompts to a fine-tuned LLM, equipped with a self-correction mechanism and a searchable database to improve accuracy and consistency.

The prompting paradigm itself is a substantial area of study. The research by [9] explores a few-shot prompt-based approach, analyzing the inherent capability of LLMs to understand semantic structure without extensive task-specific training. Similarly, [22] investigates a zero-shot technique for SRL and sentiment analysis, further probing the model's cross-lingual abilities by testing if semantic roles are preserved when translating sentences from English to Arabic.

In [9], a few-shot prompt-based approach is introduced, with a discussion of LLMs' capabilities for understanding semantics. [22] features a zero-shot prompting technique for semantic role labeling and sentiment analysis in English. The authors also investigate whether LLMs can correctly translate English sentences into Arabic, preserving the semantic role labels.

The application of these techniques also extends to specialized domains. [4] evaluate prompt-based SRL within the legal domain, comparing a general-purpose LLM to one fine-tuned specifically on legal corpora. Their findings indicate that domain-specific adaptation yields significant gains in performance and efficiency for processing complex legal texts. Finally, some research moves beyond pure prompting architectures. For instance, [27] introduces a novel framework that combines prompts to LLMs with Graph Neural Networks, aiming to capture both semantic and relational dependencies between arguments.

1.2. FrameBank

FrameBank [14] is a semantically annotated database of Russian sentences primarily based on the Berkeley FrameNet project [3]. It serves as a valuable resource for automating SRL and has been widely used in Russian NLP research.

FrameBank has been instrumental in training various SRL systems for Russian. [10] utilized FrameBank to train a semantic role classifier based on a Support Vector Machine (SVM) model, demonstrating the FrameBank's utility for traditional machine learning approaches. In [23], the authors experimented with training a neural network on this dataset. In a more recent work, [11] employed FrameBank to train a neural network encoder specifically designed to identify arguments and assign them their correct semantic roles.

1.3. Emotion Predicates

In [18], different approaches to emotion identification are compared. In short, there are three main approaches: the first categorizes emotions across entire texts, the second labels separate emotionally charged words, and the third classifies emotions within clauses. [25], similar to the third approach and our own, studies emotions within semantic frames: a frame represents an event that triggers an emotional response.

Authors in [7] present SRL4E, a unified evaluation framework that consolidates six heterogeneous emotion datasets under a standard annotation scheme based on Plutchik's emotions, enabling consistent training and evaluation of systems that identify not only emotions, but also their semantic constituents (experiencer, target, and stimulus) within text.

2. Methodology

2.1. Russian Language Dataset of Emotion Predicates

We constructed a specialized evaluation dataset[4] for Russian semantic role labeling, focusing on psychological predicates. The dataset comprises 169 manually annotated sentences extracted from Russian social media and informal text sources. Our annotation targets six predicate groups representing emotional and psychological states. They are verbs of fear: *пугать* (frighten), *ужасать* (horrify), *бояться* (fear), *опасаться* (be apprehensive), *страшить* (intimidate), and verbs of emotional attitude: *нравиться/любить* (like/love).

Each sentence was annotated by an expert linguist following a predefined semantic role taxonomy comprising five primary roles:

- **Experiencer**: The entity experiencing the psychological state (**Его** *страшит неопределенность;* **Он** *страшится будущего;* **Он** *любит девушку;* **Ему** *нравится девушка*);
- **Causator**: The entity or event that triggers the psychological response (*Его страшит* **неопределенность***; Он страшится* **будущего**);
- **Instrument**: The means or medium through which the Causator induces the response (*Будущее пугает* **неопределенностью**);
- **Deliberative**: The entity about whose welfare the Experiencer is concerned (typically marked by the Russian preposition *за*) (*Он боится* **за сына**);
- **Object**: The entity or event towards which the Experiencer feels the attitude (*Он любит* **девушку***; Ему нравится* **девушка**).

2.2. Model Configuration

We have evaluated several available LLMs, including both closed and open-weight ones:

- Anthropic Claude Sonnet 3.7 [1] – a proprietary State-of-the-Art (SOTA) LLM, developed by Anthropic. Excels at instruction following and is a particularly powerful tool for non-English language processing. The model is a hybrid reasoner; however, we have specifically disabled reasoning in our experiments;
- OpenAI GPT-5 Mini[5] – a mini variant of SOTA LLM from OpenAI. The model is reasoning-based, which means that the reasoning component cannot be disabled in it. Thus, we set it to minimal reasoning effort;
- DeepSeek V3 [13] – an open-weight non-reasoning LLM. The model contains 671 billion parameters and requires substantial hardware resources for running: up to 16 H100/A100 GPUs. Despite the size, this model is particularly interesting due to its open availability, which allows practitioners to utilize their existing HPC resources.

[4] `https://huggingface.co/datasets/dl-ru/srl-emotion-predicates`

[5] `https://openai.com/index/gpt-5-system-card/`

We established a baseline using RuELECTRA-SRL [2], a transformer-based model specifically fine-tuned for Russian semantic role labeling through token classification using a dataset from the previous work[6]. This approach mimics the functionality of NER models with BIO-style annotation to capture multi-word arguments.

2.3. Prompting Strategy

Our LLM-based approach is focused on a few-shot in-context learning approach with predicate-specific demonstrations. The prompting template for LLM consists of four main components:

1. **System Prompt**: Role specification as a native Russian linguist with explicit instructions for null-role handling ("No-Roles#No-Roles");
2. **Rule Specification**: JSON-formatted semantic role definitions tailored to each predicate group;
3. **Few-shot Examples**: All available training instances from the target predicate group, formatted as input-output pairs;
4. **Target Query**: The sentence requiring semantic role analysis.

The output format specification requires the model to generate role annotations as "- argument#role" pairs, facilitating parsing and evaluation. See example of the prompt in Fig. 1.

2.4. Evaluation Protocol

We evaluated model performance using both exact and partial matching criteria. Partial matching is necessary to capture additional aspects of model performance in multi-word arguments, particularly in clauses or phrases. It is not always possible to perfectly align LLMs with expert annotators on what should be deemed an argument in a multi-word case. The evaluation protocol is implemented as follows:

Exact Matching: A predicted argument-role pair (a_p, r_p) is considered correct if and only if there exists a ground truth pair (a_t, r_t) such that $\text{normalize}(a_p) = \text{normalize}(a_t)$ and $r_p = r_t$, where $\text{normalize}(\cdot)$ applies lowercase conversion and whitespace normalization.

Partial Matching: A prediction is considered correct if $r_p = r_t$ and either $\text{normalize}(a_p) \subseteq \text{normalize}(a_t)$ or $\text{normalize}(a_t) \subseteq \text{normalize}(a_p)$ (substring containment).

For each matching criterion, we computed:

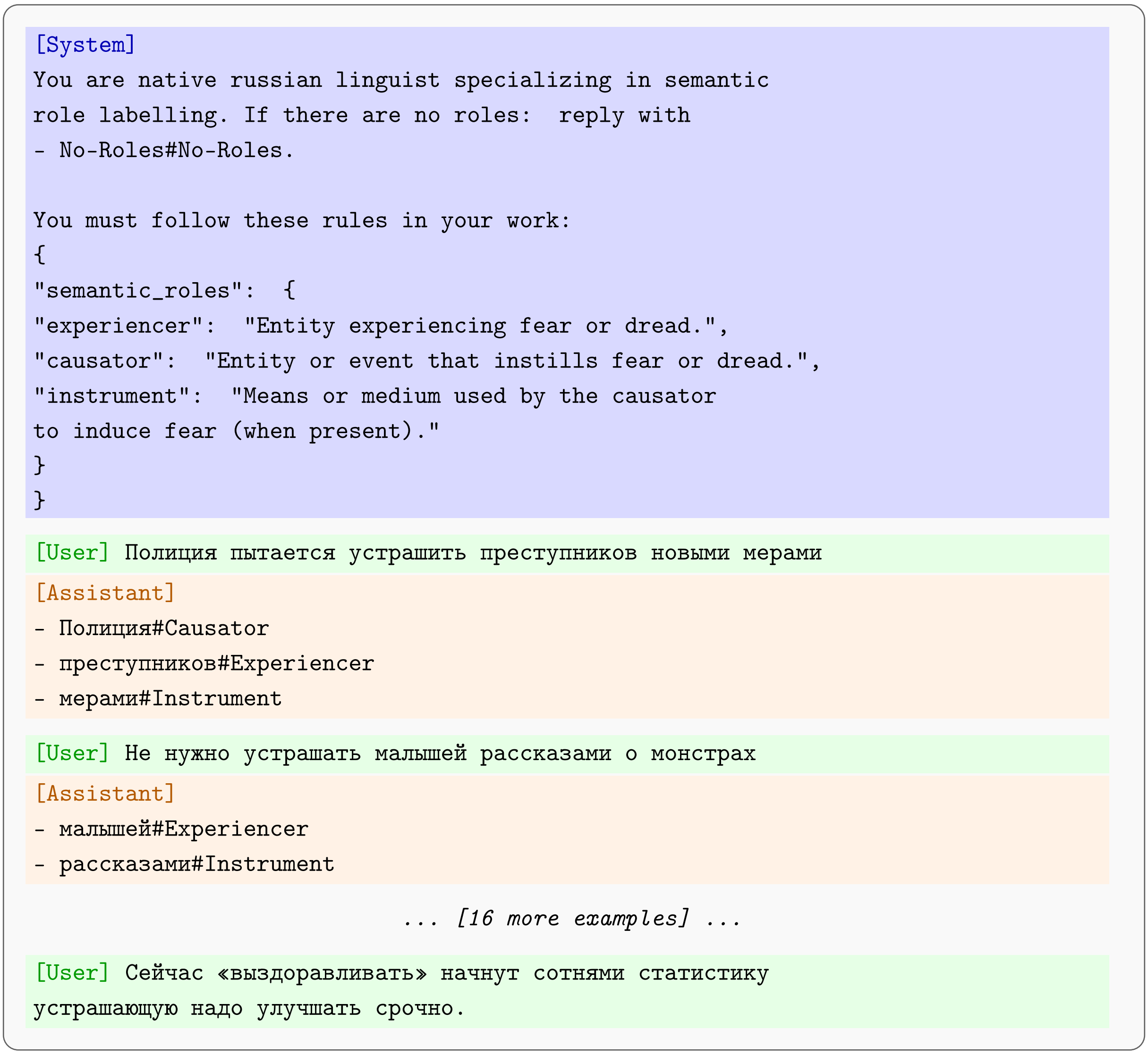
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{2}$$

$$\text{F}_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3}$$

We report both macro-averaged metrics across all instances and per-role micro-averaged metrics to assess performance across semantic roles.

[6] `https://huggingface.co/datasets/dl-ru/srl-data-ru`



```
[System]
You are native russian linguist specializing in semantic
role labelling. If there are no roles: reply with
- No-Roles#No-Roles.

You must follow these rules in your work:
{
"semantic_roles": {
"experiencer": "Entity experiencing fear or dread.",
"causator": "Entity or event that instills fear or dread.",
"instrument": "Means or medium used by the causator
to induce fear (when present)."
}
}
[User] Полиция пытается устрашить преступников новыми мерами
[Assistant]
- Полиция#Causator
- преступников#Experiencer
- мерами#Instrument
[User] Не нужно устрашать малышей рассказами о монстрах
[Assistant]
- малышей#Experiencer
- рассказами#Instrument
                    ... [16 more examples] ...
[User] Сейчас «выздоравливать» начнут сотнями статистику
устрашающую надо улучшать срочно.
```

Figure 1. Few-shot prompt for semantic role labeling of Russian Emotion predicates. The system message defines three semantic roles (Experiencer, Causator, Instrument), followed by 18 labeled examples with user inputs and assistant responses demonstrating various linguistic patterns and colloquial usage

2.5. General Role Labeling

We have additionally tried the few-shot approach described above for labeling any predicates, not only the emotion ones. To do this, we have used the semantically annotated FrameBank dataset [14]. First, the dataset is filtered so that every sentence contains exactly one annotated predicate and every predicate appears at least 10 times in the dataset. Then, a portion of the sentences is removed to be later used as examples. The rule is that for every predicate, there must be at least 5 examples, and for every semantic role that its arguments can take on, there must be at least one example.

We have employed Gemini 2.5 Flash, GPT-5 Mini, and DeepSeek-V3 for this problem. The prompt template contained additional instructions concerning the model's behavior: in what form it should answer, how it should identify the arguments if multiple words fit, etc. It also contained the target sentence, the predicate, and examples of semantic roles for that predicate.

Since RuELECTRA-SRL only deals with emotion predicates, a different baseline had to be chosen. We chose the approach from [11]. This approach is based on a pre-trained language model, fine-tuned on FrameBank, and it has the high SRL score on FrameBank corpora. For evaluation, we gave the same 10000 randomly selected sentences to every model.

3. Results

3.1. Semantic Role Labeling for Emotion Predicates

Table 1. Semantic Role Labeling Evaluation Results. For per-role results we present F1 score. For overall results we present macro-averaged F1 score, precision and recall. **Bold numbers** indicates best scores in each category across models with exact matching. Underlined numbers indicate best score with partial matching

Role/Metric	**Claude 3.7**		**GPT-5 Mini**		**DeepSeek-V3**		**Baseline**	
	Exact	**Exact+Partial**	**Exact**	**Exact+Partial**	**Exact**	**Exact+Partial**	**Exact**	**Exact+Partial**
Causator	**0.4625**	0.7000	0.4100	0.6400	0.4146	0.6463	0.0235	0.0235
Cause	0.0000	0.6667	**0.1818**	0.1818	0.0000	0.4000	0.0000	0.4000
Deliberative	0.0000	0.8571	**0.8000**	0.8000	0.3333	0.5000	0.0000	1.0000
Experiencer	**0.6636**	0.7465	0.6204	0.7130	0.5381	0.5685	0.3543	0.3657
Instrument	0.4444	0.4444	0.4211	0.4211	**0.7500**	0.7500	0.2500	0.2500
Object	**0.6889**	0.8667	**0.6882**	0.8817	0.4595	0.7027	0.1026	0.1197
Overall F1	**0.5808**	0.7485	0.5404	0.6949	0.4739	0.6174	0.1965	0.2267
Overall Precision	**0.6068**	0.7821	0.5087	0.6540	0.5317	0.6927	0.2746	0.3169
Overall Recall	0.5569	0.7176	**0.5765**	0.7412	0.4275	0.5569	0.1529	0.1765

Table 1 presents the evaluation results for semantic role labeling of Russian emotion predicates across four models. Claude 3.7 Sonnet achieved the highest overall performance with a 0.7485 F1 score on partial matching, followed by GPT-5 Mini (0.6949) and DeepSeek-V3 (0.6174), while the RuELECTRA-SRL baseline showed substantially lower performance (0.2267). The performance gap between exact and partial matching metrics shows that identifying precise argument boundaries is challenging for LLMs. Notably, all LLM-based approaches struggled with the rare Deliberative role (marked by the preposition *за*), though GPT-5 Mini achieved 0.8000 F1 on exact matching for this category. Per-role analysis shows considerable variance across semantic categories, with more frequent Experiencer and Object roles generally yielding higher scores across all models, while Instrument and Cause roles presented consistently lower performance. The results demonstrate that while LLMs substantially outperform traditional token classification approaches, significant room for improvement remains, particularly in handling less frequent semantic roles and accurately identifying argument boundaries.

3.2. General Semantic Role Labeling

Three LLMs were compared to the baseline model scores. Overall, they performed significantly better than the baseline model. Out of the three models we used, the one that showed the best results was GPT-5 Mini.

However, as one can see from Tab. 2, our few-shot approach performs substantially worse on some less common roles. Out of the 10000 test sentences, Gemini got only 61.1% completely

Table 2. Performance of different models on semantic role labeling for general predicates. For specific roles, the F1 scores are provided. The best results in each category are in **bold**

Role / Metric	DeepSeek-V3	GPT-5 Mini	Gemini 2.5 Flash	Baseline
agent (11.7%)	82.2	**87.3**	82.9	79.5
patient (10.2%)	86.6	**88.4**	85.7	86.9
theme (6.9%)	83.6	86.5	**86.9**	77.6
sbj of psychol. state (6.2%)	89.8	**92.4**	90.4	85.2
goer (5.7%)	87.9	**91.4**	89.0	85.9
cause (4.7%)	86.1	85.1	**88.3**	87.4
speaker (4.5%)	86.2	**89.7**	86.7	75.8
location (4.1%)	82.8	82.4	82.1	**84.9**
content of action (3.6%)	83.0	85.3	82.4	**86.3**
content of thought (3.4%)	86.8	**88.1**	84.8	77.0
content of speech (3.4%)	80.8	**82.1**	79.3	72.6
final destination (3.4%)	87.0	**87.1**	86.1	59.8
result (2.8%)	87.0	85.6	**91.1**	58.4
patient of motion (2.6%)	83.5	**86.6**	83.1	84.4
stimulus (2.4%)	**87.2**	86.1	85.3	78.1
cognizer (2.3%)	81.9	**85.7**	83.2	80.8
addressee (1.8%)	86.0	**86.5**	85.9	77.4
perceiver (1.7%)	88.4	**93.5**	91.7	84.3
counteragent (1.6%)	87.0	**87.8**	86.7	60.9
effector (1.4%)	66.1	65.8	67.6	**78.9**
subject of social attitude (1.1%)	77.9	81.0	**82.4**	80.8
initial point (1.1%)	88.3	84.5	**88.4**	78.1
topic of speech (1.0%)	**82.5**	81.7	78.3	68.0
manner (1.0%)	64.0	54.5	65.7	**76.0**
recipient (1.0%)	81.0	**86.2**	79.8	74.5
goal (0.9%)	**76.0**	75.5	75.8	73.3
field (0.7%)	78.8	80.0	74.5	**91.3**
attribute (0.7%)	73.1	70.3	67.0	**82.5**
source of sound (0.7%)	86.4	**92.2**	87.2	71.6
behaver (0.6%)	78.2	**86.5**	77.4	84.6
situation in focus (0.6%)	75.1	73.4	65.6	**88.2**
counteragent of social attitude (0.6%)	**84.6**	80.2	83.9	65.5
sbj of physiol. reaction (0.6%)	89.0	**91.4**	88.3	80.4
topic of thought (0.6%)	70.8	67.1	70.5	**92.2**
potential patient (0.5%)	88.3	89.8	89.3	**90.1**
status (0.5%)	**87.8**	86.1	78.1	83.3
patient of social attitude (0.5%)	58.4	58.2	60.7	**80.8**
standard (0.5%)	86.4	**88.0**	85.6	82.7
term (0.5%)	91.2	**86.6**	90.4	**86.6**
attribute of action (0.5%)	79.5	74.5	74.8	**80.4**
causer (0.4%)	54.3	**69.3**	68.6	68.7
initial possessor (0.4%)	76.1	**81.0**	63.3	78.3
potential threat (0.4%)	84.1	**87.0**	78.6	77.9
path (2.3%)	65.7	67.5	61.8	**84.9**
Argument Extraction F1	87.6	**88.4**	86.4	79.4
Argument Extraction Precision	**87.7**	84.2	84.9	74.5
Argument Extraction Recall	87.5	**93.2**	86.4	85.1
Role Labeling F1	83.3	**85.0**	83.1	80.1

right, and both DeepSeek and GPT-5 got 60.5%. Curiously, there is a substantial overlap of examples that the models got wrong. All three LLMs gave wrong answers on the same 20% of the data, indicating that some sentences may be inherently "counterintuitive" for LLMs (see also [9], where mistakes made by LLMs were compared to mistakes by human non-experts).

3.3. Analysis of Exact vs. Partial Matching Disparities

The performance patterns in Tab. 1 reveal a notable discrepancy across models in their ability to identify precise argument boundaries. Claude 3.7 Sonnet, despite achieving the highest overall F1 score (0.7485) on partial matching, recorded zero scores for exact matching in two semantic categories (Cause and Deliberative), while maintaining substantial partial matching scores for these same roles (0.6667 and 0.8571 respectively). In contrast, GPT-5 Mini demonstrated non-zero exact matching performance across all categories, including a particularly strong 0.8000 F1 score for the Deliberative role. DeepSeek V3 exhibited zero exact matching only for the Cause role.

This pattern correlates with our multiword argument analysis, where Claude achieved the highest success rate (70%) in correctly extracting and labeling multiword arguments, compared to GPT-5 Mini (65%) and DeepSeek V3 (63.33%). The data suggest that Claude adopts a more expansive approach to argument boundary identification, consistently capturing the semantic core of arguments while frequently deviating from expert-annotated boundaries. This is particularly evident in rare semantic roles: the Deliberative role appears only in 7 instances (4.1% of arguments), and Cause in 5 instances (3.0%). For such infrequent categories, Claude tendency to extract semantically appropriate but boundary-imprecise arguments results in complete exact matching failure, yet high partial matching success.

GPT-5 Mini consistent non-zero exact matching across all roles reflects a more conservative extraction strategy that better aligns with annotator boundary conventions, albeit at the cost of missing some semantically relevant material. The model reasoning component (which does not operate in our experiments) does not contribute to boundary-aware predictions. DeepSeek V3 intermediate behavior, with exact matching failure only for Cause, suggests it falls between these two strategies, as does reasoning, but lacks in depth understanding.

4. Discussion

The LLM-based approach allows practitioners to trade increased computational resources for reduced data annotation costs while also substantially improving the method robustness. In some instances, such as with SRL for the Russian language, this tradeoff is particularly effective, which stems from the natural complexity of semantic linguistic annotation. Indeed, the LLM-based approach, in our case, requires only minimal human annotation to briefly cover the predicate groups we are focusing on. Training a regular deep learning model, such as those we demonstrate as baselines in the section above, would require 30-40x more annotated examples to achieve the desired level of quality, not accounting for complex cases, such as multi-word arguments. This computational intensity makes this approach well-suited for HPC environments, where accelerated processing capabilities can efficiently handle the increased resource demands of LLM inference at scale.

Linguistically, the well-known difficulties of text annotation and semantic roles prediction in Russian, as a morphologically rich language, mentioned in the Introduction, should be extended by a number of additional problems. Generally speaking, they concern the problems of grammar of constructions, on one hand, and text production and genre, on the other hand.

Firstly, in the Russian grammar, there exist constructions with the personal and impersonal subjects of the sentence (Doer, Experiencer, etc.) omitted, which are generally referred to as "syntactic zeros" [15] of complete sentences and "zero pronouns" [6]. The syntactic zeros (∅) can take such meanings as "others (somebody excluding the speaker)" – *Уводили тебя на рассвете* (Akhmatova, ∅ took you away at the sunrise), "everybody including the speaker" – *Как потопаешь, так и полопаешь* (Proverb, As ∅ sow so shall ∅ mow) [5]. There are also 1st and 2nd person forms of verbs that definitely indicate the subject of a sentence: the Speaker (I – *Люблю грозу в начале мая* (Tyutchev, ∅ love the thunderstorm in the beginning of May)) and Listener (you – *Послушайте! Ведь, если звезды зажигают...* (Mayakovsky, ∅ Listen! In fact, if somebody lights the stars...)).

Secondly, text production supposes that the subject of the sentence can be replaced by a pronoun or omitted for the reasons of economy and coherence. Such modified subjects refer to the anaphora and can be represented by 3d person pronouns or syntactic zero. For both, the referent can be typically found in the left context of the text. See also summarizing presentation in [20].

Thirdly, we also encountered some unexpected difficulties with genre as we analyzed social media discussion content. This type of text belongs to written speech genres and consists of dialogue and, to some extent, monologue fragments. They are characterised by spoken and spontaneous features, and contain interrupted elements within them, as well as irregular usages. The latter, for example, is expressed in irregular cases of Causator: *ужаснулся* **над дырявостью наших законов**. Some syntactic fragments are brief and incomplete, and they frequently appear in dialogues. Some fragments are extensive, for instance in monologues. Structurally, this can result in a distant position between the emotion predicate and its Experiencer argument: **Водитель** *не видел что в другом ряду Жигули остановились, явно пропуская пешехода, куда все несутся, время экономит,* **боится** *лишние секунды потерять, хорошо не задел её.*

The monologue parts structurally and semantically are similar to egocentric 'I'('Я')-texts (such as narrative memoirs, diaries) and therefore contain 'I' (1st person pronoun) omitted constructions. Likely for all actual (present tense) emotion expressions, syntactic zeros of 'I' Experiencer are very specific for emotion verb constructions (compare: *Жаль; Обидно; Грустно; Пугает, что...*). Likely in 'I'-texts, constructions of emotion verbs denoting emotions of 'I' Experiencer often include corresponding syntactic zero of 1st person pronoun even in the past tense: *Проезжал мимо и просто ужаснулся,что же за гений архитектор это нарисовал!!!*(∅ was passing by and purely got frightened what a genius architect painted all this <...>). An additional argument for 'I' Experiencer reconstruction in this sentence is that the position of the Causator is filled with a direct speech clause (the problem of direct/indirect speech in complement clauses is discussed in [21, 26]).

Evidently, we could have never expected to extract arguments represented by such syntactic phenomena as noun phrases, clauses, and syntactic zeros, even with the help of LLM. Meanwhile, the expert analysis of what LLM recognizes as arguments of verbs of emotions convinced us that its process of thinking can be very productive and cover complicated and unsolvable cases for other methods, and carefully identify their semantic roles.

See examples of LLM prediction of Causator and Object arguments expressed by a single noun, noun phrase, and clause given below (Tab. 3). Our method is able to analyse and identify constructions with more than one pretender to be the argument of a verb, i.e. independent nouns and noun phrases connected by coordinating conjunction: *Толкучки и общественные места любите?* At the same time it in an arbitrary way can incorrectly shorten a long noun phrase: *Мне не понравилось [отсутствие реакции]#Object руководства больницы на мое предложение поставить там кофейный автомат*; *Особенно понравилось [исполнение]#Object второй песни* – instead of *Мне не понравилось [отсутствие реакции руководства больницы на мое предложение поставить там кофейный автомат]#Object*; *Особенно понравилось [исполнение второй песни]#Object.*

Additionally, our method recognizes not only Causators based on dependent clauses (followed by conjunction), but also independent clauses (no conjunction) ones: *Боюсь, что [в 24 году не до паркунов будет] - Если можно было бы взять его 4-ым, а бы взяла, но боюсь [тогда меня из дома выгонят вместе с ним].* It generally takes place in contexts of internal state verbs in the Present tense with 'I' (speaker) subjects. The meaning of 'I'-subject here becomes closer to Thinker than Experiencer, which reflects in the replacement of indirect (followed by a conjunction) by direct (no conjunction) speech, as we noticed above.

See Tab. 4 for examples of LLM predictions for syntactic zeros, both anaphoric (the meaning derived from the text) and paradigmatic ones (the meaning derived from the form of the verb or construction). It is interesting that in contexts of V3pl construction with a Nominative noun omitted and a present locative component, the latter is predicted as the subject of the sentence, i.e., the Experiencer, which entirely aligns with the common idea of grammatical interpretation.

Table 3. Examples of LLM Method recognizing Causator and Object arguments expressed by a single noun, noun phrase, clause, etc.

Syntactic status of argument	**Text**	**Fragment in focus translated**
Single noun or pronoun	**Женщин** люблю; **никого** не боялся	(I) like **women**; (I was) afraid of **nobody**
Noun phrase	Ольга, я с вами согласна, я тоже люблю **больших собак !**	I like **big dogs**
Independent elements	**Толкучки** и **общественные места** любите? ; Люблю **больших** И **коренастых**!	(Do you) like **crowds** and **public places**; I like **big** and **stocky** (ones)
Infinive or infinitive phrase	Татьяна, я живу на 5 этаже и у меня за окном такие сосульки висят, но самостоятельно, я лично, **сбивать** боюсь.; Наталья, а Вы не думаете, что персонал боится **заразиться через детей и их родителей**???	I am afraid **to break down icicles**; <...> the staff are afraid **to get infected through children and their parents**
Dependent clause	Не боитесь что **поклоники Высоцкого вас побьют**?	Aren't you afraid **to be beaten up by (his) fans**?
Pronoun followed by dependent clause	Дмитрий, а тех **кто летает по дорогам, как в <...> жаленый, не глядя на пешеходные переходы,** ты любишь?	<...> and do you like **those who rush along the roads** <...>

The gerund constructions are also known as those that carry a syntactic zero of the subject of the action (coreferent to the subject of the main predicate) [24], and LLM can establish such a zero subject in our texts.

We have to consider particular cases that LLM performs incorrectly. It concerns 1st and 2nd person pronouns and syntactic zeros, which are rather deictic than anaphoric, i.e., they do not need to be replaced by nouns from the text, as their referents are the Speaker and the Listener, marked by corresponding personal pronouns. See LLM failures in Tab. 5.

On the other hand, LLM does not operate with classic anaphoric 3rd person pronouns, they are left without their context referents: *Катерина, ну да, мы* **его** *просто очень любим* (we love **him** very much); *Но мне* **это** *не нравится* (But I don't like **it**); *Натали,* **они** *как огня боятся областную жил инспекцию* (**they** are afraid of the local housing inspectorate). Probably, it happens for the reason that their referents are cut away in the process of text syntactic segmentation and therefore stay behind the frame of the sentence analysed. In general, as far as grammatically complicated cases are concerned, the advantages of the great performance of our method are complemented by the disadvantages of irregularity and instability of performance. Someone can still not entirely rely on the decisions of LLM. At the same time, in comparison to all other existing methods for automatic SRL for the Russian Language at scale, LLMs demonstrate the highest accuracy and robustness to linguistic phenomena.

Table 4. LLM prediction of Experiencer for syntactic zeros

∅	**Predicted from**	**Meaning**	**Text**	**Translation**
Syntactic zero paradigmatic	Form of the verb, V_{2s}	Personal pronoun 2 listener Вы Вы Ты	Ksu, ну хорошо что **любите** **Не страшитесь**, мне тоже нечупно и душевно и сопли от дыма уже пошли. Татьяна, **ужасайся** дальше	It is good ∅ **love** ∅ **Don't frighten** <...> ∅ **be horrified** again (= don't stop being horrified)
Syntactic zero paradigmatic	Form of the verb, V_{1s}	Personal pronoun 1 speaker Я	Александр, только что сказали, что в Москве закрыты аэропорты больничный без посещения поли-ки, а мне позвонили и сказали, что надо идти в регистратуру и забирать его, вот тоже этого **страшусь** уже.	∅ also **am afraid** already
		Я	Валера, уже напротяжении 16 лет **люблю** творчество граффити, это позерство	∅ have been **loving** graffiti art for 16 years <...>
Syntactic zero anaphoric	V_{ps} chain of predicates	Personal pronoun 1 speaker Я	Приезжала в город в январские каникулы, **ужаснулась**	I went to the town for January vacation, ∅ **got frightened**
Syntactic zero anaphoric	Pronoun, V_{ppl} chain of predicates	Все	А так, все увидели и **ужаснулись**.	Everybody saw and ∅ **got frightened**
Syntactic zero paradigmatic	V_{3pl} construction	У нас	Или опять пешеход виноват, как у нас **любят** говорить?	<...> as ∅ **like** saying at our's (=People say that..., It is generally said that)
Syntactic zero anaphoric	Gerund	-	И хватит уже оглядываться на всяких либералов, **опасаясь** задеть их	∅ Stop taking into account liberals ∅ **being afraid** to hurt them
Syntactic zero anaphoric	V_{ps} chain of predicates, *для него* pronoun	Он	Решил развернуться, затем для него внезапно появилась машина (которая его почти догнала), в итоге **испугался**, потерял контроль над управлением.	∅ Decided to turn around, then suddenly for him a car appeared, finally ∅ **got afraid**
Syntactic zero anaphoric	V_{3s} chain of predicates	Марс	Марс) самый милый пухляш на земле, супер ласковый и **любит** сидеть на ручках	Mars the nicest little fat kid on the Earth, very sweet and ∅ **loves** sitting in the arms

Table 5. LLM predictions for 1st and 2nd person pronouns

Predicted pronoun	**Predicted from**	**Functioning**	**Text**	**Translation**
2nd person pronoun *вам*	Владимир	Wrong	Владимир, а вам **нравится**, когда вас с кем сравнивают?	Vladimir, and do you **like** <...>
1st person pronoun *мне*	Александр	Wrong	Александр, соты мне **понравились**	Alexander, I **like** honeycombs
1st person pronoun *мы*	мы	Right	Сынок, мы тебя очень **любим**!!!	Sonny, we **love** you very much
Syntactic zero (*Я*)	Валера	Wrong	Валера, уже напротяжении 16 лет **люблю** творчество граффити, а это позерство	Valera, ∅ have been **loving** graffiti art for 16 years

Conclusion

We presented a novel approach to semantic role labeling for Russian emotion predicates using large language models with few-shot learning. Our method demonstrates that LLMs can achieve significantly better performance than traditional supervised approaches, with Claude 3.7 achieving 74.85% F1 score on partial matching compared to 22.67% for the RuELECTRA baseline. For general predicates on FrameBank, GPT-5 Mini reached 85.0% F1, substantially outperforming the previous state-of-the-art of 80.1%.

The method successfully handles complex linguistic phenomena, both specific to Russian and natural to various languages in general, including syntactic zeros, anaphoric references (less than others), multi-word arguments, and clausal structures. LLMs demonstrate remarkable capability in identifying emotion arguments even in challenging social media texts with interrupted elements and irregular constructions. However, the approach shows limitations with certain anaphoric 3rd person pronouns and occasionally produces arbitrary segmentation of long noun phrases.

Future work should address the stability of predictions across different predicate types and explore hybrid approaches combining LLM reasoning with structured linguistic knowledge. The development of larger annotated corpora for Russian emotion predicates remains critical for advancing the field. Additionally, the study of cross-lingual transfer learning could leverage resources from morphologically similar languages to improve coverage of rare semantic roles.

Acknowledgements

This research was supported by the Ministry of Science and Higher Education of the Russian Federation, project No. 075-15-2024-544.

The research was carried out using the infrastructure of the shared research facilities “High Performance Computing and Big Data” of FRC CSC RAS (CKP “Informatics”).

References

1. Claude 3.7 Sonnet System Card. https://api.semanticscholar.org/CorpusID:276612236

2. Arkhipov, M., Trofimova, M., Kuratov, Y., Sorokin, A.: Tuning Multilingual Transformers for Language-Specific Named Entity Recognition. In: Erjavec, T., Marcińczuk, M., Nakov, P., *et al.* (eds.) Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing. pp. 89–93. Association for Computational Linguistics, Florence, Italy (Aug 2019). https://doi.org/10.18653/v1/W19-3712

3. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The Berkeley FrameNet project. In: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1. pp. 86–90. Association for Computational Linguistics, Montreal, Quebec, Canada (Aug 1998). https://doi.org/10.3115/980845.980860

4. Bakker, R., Schoevers, A., van Drie, R., *et al.*: Semantic role extraction in law texts: a comparative analysis of language models for legal information extraction. Artificial Intelligence and Law. P. 1–35 (03 2025). https://doi.org/10.1007/s10506-025-09437-x

5. Bulygina, T.V.: Ya (I), ty (you) and others in the Russian grammar. Res philologica. Philological researches pp. 111–126 (1990)

6. Bulygina, T.V., Shmelev, A.D.: Language conceptualization of the world (based on the material of Russian grammar). Shkola “Jazyki russkoj kul’tury”, Moscow (1997)

7. Campagnano, C., Conia, S., Navigli, R.: SRL4E – Semantic Role Labeling for Emotions: A unified evaluation framework. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 4586–4601. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.acl-long.314

8. Chen, H., Zhang, M., Li, J., *et al.*: Semantic role labeling: A systematical survey (2025). https://doi.org/10.48550/arXiv.2502.08660

9. Cheng, N., Yan, Z., Wang, Z., *et al.*: Potential and Limitations of LLMs in Capturing Structured Semantics: A Case Study on SRL. In: Advanced Intelligent Computing Technology and Applications: 20th International Conference, ICIC 2024, Tianjin, China, August 5–8, 2024, Proceedings, Part I. pp. 50–61. Springer-Verlag (2024). https://doi.org/10.1007/978-981-97-5663-6_5

10. Kuznetsov, I.: Semantic Role Labeling for Russian Language Based on Russian FrameBank. In: Khachay, M.Y., Konstantinova, N., Panchenko, A., *et al.* (eds.) Analysis of Images, Social Networks and Texts. pp. 333–338. Springer International Publishing, Cham (2015). https://doi.org/10.1007/978-3-319-26123-2_32

11. Larionov, D., Shelmanov, A., Chistova, E., Smirnov, I.: Semantic Role Labeling with Pretrained Language Models for Known and Unknown Predicates. In: Mitkov, R., Angelova, G. (eds.) Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). pp. 619–628. INCOMA Ltd., Varna, Bulgaria (Sep 2019). https://doi.org/10.26615/978-954-452-056-4_073

12. Li, X., Chen, H., Liu, C., *et al.*: LLMs Can Also Do Well! Breaking Barriers in Semantic Role Labeling via Large Language Models. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Findings of the Association for Computational Linguistics: ACL 2025. pp. 23162–23180. Association for Computational Linguistics, Vienna, Austria (Jul 2025). https://doi.org/10.18653/v1/2025.findings-acl.1189

13. Liu, A., Feng, B., Xue, B., *et al.*: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024). https://doi.org/10.48550/arXiv.2412.19437

14. Lyashevskaya, O., Kashkin, E.: FrameBank: a database of Russian lexical constructions. In: International Conference on Analysis of Images, Social Networks and Texts. Communications in Computer and Information Science, vol. 542, pp. 350–360. Springer (2015). https://doi.org/10.1007/978-3-319-26123-2_34

15. Mel'chuk, I.A.: About the syntactic zero. Typology of passive constructions, diatheses and voices pp. 343–360 (1974)

16. Nikitina, E.N., Onipenko, N.K.: Semantics and pragmatics of statements with psych verbs. Siberian Journal of Philology (2), 271–285 (2022). https://doi.org/10.17223/18137083/79/19

17. Nikitina, E.N., Smirnov, I.V.: Predicate-argument structure for intelligent text analysis of social media content. Speech Technology (1-2) (2020), https://api.semanticscholar.org/CorpusID:256224132

18. Oberländer, L.A.M., Klinger, R.: Token sequence labeling vs. clause classification for English emotion stimulus detection. In: Gurevych, I., Apidianaki, M., Faruqui, M. (eds.) Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics. pp. 58–70. Association for Computational Linguistics, Barcelona, Spain (Online) (Dec 2020), https://aclanthology.org/2020.starsem-1.7/

19. Oberländer, L.A.M., Reich, K., Klinger, R.: Experiencers, stimuli, or targets: Which semantic roles enable machine learning to infer the emotions? In: Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media. pp. 119–128 (2020)

20. Onipenko, N.K.: The model of the perspective of subjects (persons) and the problem of classification of egocentric elements. In: The problems of functional grammar: The principle of natural classification, pp. 92–121. Jazyki slavyanskoy kul'tury, Moscow (2013)

21. Paducheva, E.V.: Egocentric units of language and the modes of interpretation. In: Computational linguistics and intellectual technologies. Papers from the Annual International Conference "Dialogue". vol. 1, pp. 486–503. RGGU, Moscow (2013)

22. Senator, F., Lakhfif, A., Zenbout, I., *et al.*: Leveraging ChatGPT for Enhancing Arabic NLP: Application for Semantic Role Labeling and Cross-Lingual Annotation Projection. IEEE Access 13, 3707–3725 (2025). `https://doi.org/10.1109/ACCESS.2025.3525493`

23. Shelmanov, A., Devyatkin, D.: Semantic role labeling with neural networks for texts in Russian. In: Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue". vol. 1, pp. 245–256. RGGU, Moscow (2017)

24. Testelec, Ja. G.: Introduction to General Syntax. RGGU, Moscow (2001)

25. Troiano, E., Klinger, R., Padó, S.: On the relationship between frames and emotionality in text. Northern European Journal of Language Technology 9(1) (2023). `https://doi.org/10.3384/nejlt.2000-1533.2023.4361`

26. Voloshinov, V.N.: Marxism and the philosophy of language. Priboy, Leningrad (1930)

27. Zhou, Y., Fan, J., Zhang, Q., *et al.*: Modeling semantic-aware prompt-based argument extractor in documents. Applied Sciences 15(10) (2025). `https://doi.org/10.3390/app15105279`