PROGRAM SYSTEMS: THEORY AND APPLICATIONS ISSN 2079-3316

Research Article ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

UDC 004.912:81'322.2

10.25209/2079-3316-2026-17-5-3-42

Training Large Language Models to Handle Russian Verb–Noun Collocations

Vera Pavlovna **Krasnobaeva**✉

ITMO University, St. Petersburg, Russia

✉*krasnobaevavera2022@gmail.com*

Abstract. Non-compositional lexis — collocations, idioms and verb–noun constructions — remains a weak spot of large language models (LLMs), especially in morphologically rich languages such as Russian, where the boundary between compositional and non-compositional readings is graded. Russian verb–noun collocations were used as a controlled testbed for this boundary. New methodologically balanced dataset of 3,264 expressions is based on 479 verbs, with an equal number of collocations and compositional constructions and class balance for every verb, and define three subtasks over the same expressions: binary classification, single-verb paraphrasing, and gap filling. Seven multilingual and Russian-specialized LLMs are evaluated in zero-shot, one-shot and few-shot regimes, and two mid-size models are fine-tuned with LoRA in a multitask setting. Base models classify near chance and are unstable in generation; scale and language specialization help but do not close the gap. Fine-tuning raises classification accuracy from about 53% to about 80%, and on the generative subtasks a fine-tuned 8B model matches or surpasses much larger models. The paper extends a talk at the Dialogue-2026 conference; the dataset and code are publicly available. (*Linked article texts in English and in Russian*).

Key words and phrases: verb–noun collocations, non-compositionality, idiomaticity detection, large language models, parameter-efficient fine-tuning, LoRA, paraphrasing, gap filling, Russian NLP

2020 *Mathematics Subject Classification*: 68T50; 68T07, 91F20

For citation: Vera P. Krasnobaeva. *Training Large Language Models to Handle Russian Verb–Noun Collocations*. Program Systems: Theory and Applications, 2026, **17**:5(74), pp. 3–42. (*In English, in Russian*). https://psta.psiras.ru/read/psta2026_5_3-42.pdf

© Krasnobaeva V.P. 2026

Introduction

The automated processing of idiomatic and non-compositional language remains one of the most persistent challenges in natural language processing. Although modern large language models exhibit strong general linguistic competence, their ability to reliably detect and interpret non-compositional expressions is insufficiently understood. Russian verb–noun collocations are a particularly demanding case: they combine lexical, semantic and grammatical constraints that require a model to move beyond surface-level compositional interpretation. A collocation such as *oderzhat' pobedu* ('achieve victory') cannot be read as the sum of its parts, yet it is not a frozen idiom either: it inflects, admits limited modification, and paraphrases cleanly to a single verb (*pobedit'*, 'win').

Research on idiomaticity detection has been carried out primarily on English data and largely in classification settings, and no large, methodologically balanced corpus of Russian verb–noun collocations exists to date. Most work targets fully idiomatic expressions, leaving the intermediate zone underexplored. Few studies integrate classification, generation and reconstruction in a single framework, so recognition and generation abilities are usually measured on different data and cannot be compared directly.

The goal of this work is to build a balanced dataset of Russian verb–noun collocations, to define three interconnected subtasks over the same expressions, to systematically evaluate multilingual and Russian-specialized LLMs on them in zero-shot, one-shot and few-shot regimes, and to investigate to what extent parameter-efficient domain adaptation with LoRA closes the gap left by model scale and language specialization. The dataset and code are released at `https://github.com/Vera-bahval/russian-verb-noun-collocations`.

Differences from the conference version. This paper extends our article "Automatic Processing of Russian Verb-Noun Collocations Using Large Language Models", published in the proceedings of the Dialogue-2026 conference [1], which introduced the dataset, the three-task formulation, the evaluation of seven LLMs, and the fine-tuning results.

The dataset, task definitions, set of models, fine-tuning hyperparameters and headline metrics are unchanged, so the two publications report mutually consistent numbers. The new material comprises:

(1) a detailed account of the dataset methodology: the operational definition of a collocation, the role of the seed dictionary, the design of the Russian National Corpus queries, and the augmentation and balancing steps (Section 2);

(2) a per-task analysis of the characteristic errors of general-purpose models, with worked examples (Section 4);

(3) a side-by-side comparison of base and fine-tuned outputs of both models on the generative subtasks (Section 5);

(4) a consolidated cross-model, cross-task comparison and a discussion of the interplay between model scale, language specialization and targeted adaptation (Section 6).

1. Related work

A foundational step in idiomaticity research was the creation of large annotated corpora. MAGPIE [2] combined dictionary-based extraction, automated mining and multi-level crowdsourced annotation of English expressions; it contains over 50,000 annotated instances and established dictionaries plus large corpora as a viable data pipeline.

AStitchInLanguageModels [3] introduced multi-level classification of multiword expressions and used paraphrasing as a probe of semantic sensitivity in zero-shot, one-shot and few-shot regimes; it also showed that special tokenization of the target expression and the training regime affect measured quality.

FLUTE [4] framed figurative-language understanding as natural language inference with explanation generation; candidate sentences were generated with GPT-3 and then validated by experts.

SemEval-2022 Task 2 [5] set a multilingual benchmark for idiomaticity detection in English, Portuguese and Galician. Adapted encoder architectures led, and the winning one-shot system relied on ensembling, adversarial and contrastive training, and explicit marking of the multiword expression.

Phelps et al. [6] showed that general-purpose LLMs without adaptation underperform smaller encoder models fine-tuned for the task; at the same time, their ability to separate idiomatic from literal usage without dedicated training signals a general latent competence, and quality grows with parameter count.

Zhou et al. [7] integrated idiom definition generation, identification and reconstruction in a unified framework and adapted the model with parameter-efficient modules rather than full fine-tuning; this three-subtask decomposition directly inspired the design of our three tasks.

Two lessons emerge from this line of work and shape our design:

First, dictionaries and large reference corpora are the natural seed for building non-compositional resources, and LLM-based augmentation scales the process with moderate manual editing.

Second, non-compositionality is best probed by combining recognition and generation and by testing several models under several prompting regimes.

What is still missing is a large, methodologically verified corpus of Russian verb–noun collocations; the intermediate zone between compositional and idiomatic expressions remains underexplored; and few studies combine classification, paraphrasing and reconstruction over the same items. Our work addresses these gaps.

2. Dataset

2.1. Operational definition

In this work, a verb–noun collocation is a verb plus a noun (a direct or indirect object, optionally separated by a single preposition) whose meaning is non-compositional. The diagnostic criterion is the possibility of paraphrasing the whole construction with a single synonymous verb: *oderzhat' pobedu* ('achieve victory') qualifies, as it reduces to *pobedit'* ('win'), and *brosit' vzglyad* ('cast a glance') reduces to *posmotret'* ('look'), whereas *issledovat' vopros* ('investigate a question') is compositional.

The criterion is chosen deliberately, because it targets precisely the intermediate zone between fully compositional and fully idiomatic expressions.

2.2. Sources and construction

The initial source of positive examples was the dictionary *Stable Verbal-Nominal Collocations of the Russian Language* [8], which contains over 6,000 verb–noun expressions together with government (case-marking)

information. Its targeted scope and explicit grammatical information made it a convenient starting list, but it served only as a seed: all candidates were checked against contemporary corpus evidence and against the paraphrase criterion, so that items absent or marginal in present-day usage, or not reducible to a single verb (e.g., *brat' taksi*, 'take a taxi'), were excluded.

Automatic preprocessing (regular expressions and string operations) retained the verb, the noun and any single preposition; where both aspectual forms of the verb were listed, the imperfective was kept.

The *Russian National Corpus* (RNC) was used both as a contemporary frequency filter and as a source of new positive and negative candidates. Grammatical features in the corpus queries reflected each verb's syntactic profile: case matched the verb's government; a single intervening preposition was permitted only for prepositional complements; number was left unconstrained to capture morphological variability.

For each verb, the 100 most frequent constructions were screened manually: items satisfying the single-verb paraphrase criterion were labeled as collocations (label 1), the others as compositional examples (label 0). For instance, for the verb *vodit'* the construction *vodit' za nos* ('lead by the nose' → 'deceive') was added as a positive example on grounds of frequency and a clean single-verb paraphrase, while *vodit' kist'yu* ('move a brush') was added as a negative example, being both less frequent and compositional.

Because this procedure produced a surplus of positives, we balanced the set with data augmentation using GPT-4o [9]: the model was given an explicit definition and examples of both classes and was prompted to generate verb–noun expressions in batches of 100 (the prompt is given in the Appendix).

All outputs were manually verified for compositionality, grammaticality and non-duplication; generations that were themselves collocations (*vodit' druzhbu*, 'keep up a friendship') or semantically odd were discarded. This added 983 verified negative examples.

2.3. Balancing and final composition

A residual imbalance remained at the level of individual verbs: some verbs had many more examples of one class than of the other, which would prevent a clean train/test split. We therefore enforce per-verb balance: for each of the 479 verbs that had at least one positive example, we kept an

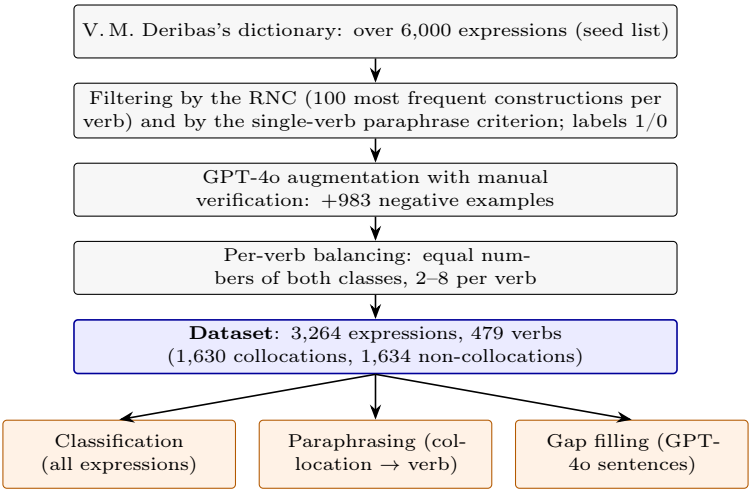


FIGURE 1. Dataset construction pipeline and the three subtasks defined over the same expressions

equal number of collocations and non-collocations (2 to 8 examples per verb), topping up scarce classes by repeating the RNC search and the GPT-4o augmentation.

The final corpus contains 3,264 expressions: 1,630 collocations and 1,634 non-collocations, balanced by class and by verb. The construction pipeline is summarized in Figure 1.

For the paraphrasing subtask, each collocation was manually assigned a synonymous verb (e.g., *zadat' napravlenie* → *napravit'*, *zalivat'sya kraskoy* → *krasnet'*, *znat' tolk* → *razbirat'sya*). For the gap-filling subtask, contextualized sentences with a masked collocation were generated for the positive examples by GPT-4o under controlled prompting and then reviewed manually; negatives are not used in this subtask. Sample entries are shown in Table 1.

The three subtasks probe complementary capabilities:

classification targets binary recognition of non-compositionality,
paraphrasing tests lexical-semantic understanding, and
gap filling tests contextual integration of form, meaning and grammatical realization.

Because they are defined over the same expressions, recognition and generation can be measured on a common footing. The resource is intended

Table 1. Sample entries from the dataset (Russian transliterated, English glosses in parentheses)

Expression	Label	Synonym	Sentence with gap
vynosit' obvinenie (bring an accusation)	1	*obvinyat'* (accuse)	*Posle dolgogo sledstviya prokuratura reshila* ⟨MASK⟩ *podozrevaemomu po tryom stat'yam.*
vynosit' poritsanie (issue censure)	1	*poritsat'* (censure)	*Za amoral'noe povedenie emu bylo resheno* ⟨MASK⟩ *pered vsemi kollegami.*
vynosit' pytku (endure torture)	0	—	—

for reuse in idiomaticity detection, paraphrase generation, controlled text-generation evaluation, and contrastive studies.

3. Evaluation of large language models

3.1. Models and setup

Seven LLMs were evaluated varying in size, language specialization and tuning regime: Mistral-7B-Instruct-v0.1 [10], Qwen2.5-7B-Instruct [11], DeepSeek-llm-7B-base [12], Llama-3.1-8B-Instruct and Llama-3.1-70B-Instruct [13], and *YandexGPT5-Lite-8B* and YandexGPT5-Pro.

This selection separates the effects of size (7B to 70B parameters), of language specialization (multilingual vs. Russian-specialized models) and of instruction tuning; the base autoregressive model DeepSeek-llm-7B-base is included as a contrast to the instruction-tuned systems.

Models up to 8B parameters were run locally with the HuggingFace Transformers library (`AutoModelForCausalLM` and `AutoTokenizer`); larger models were queried through the Yandex Cloud API. Each model was tested in zero-shot, one-shot and few-shot regimes, with in-context examples drawn from the dataset according to the subtask.

3.2. Task specifications and prompting

Classification. The model receives a verb–noun expression and predicts a label in $\{0, 1\}$, returning only the digit. In the one-shot regime two examples (one per class) were provided, since a single example would bias a binary task toward one label; in the few-shot regime six examples were used, balanced across classes and varied in case, number and presence of a preposition.

Accuracy (A), Precision (P), Recall (R) and F1 were considered for the positive class.

Paraphrasing. The model receives a collocation and must return one synonymous verb. Prompts were tuned per model family: mid-size multilingual models required an explicit "answer in Russian only" constraint to suppress English and, for Qwen, Chinese fragments.

The base DeepSeek model needed separate treatment, because an instruction-style prompt made it echo the instruction instead of producing a verb. For this model, the zero-shot regime used a forced-continuation prompt ending with an open quote that primes a verb, while the one-shot and few-shot regimes used a list of `collocation = verb` pairs ending with the target collocation followed by "=".

Gap filling. The model replaces the ⟨`MASK`⟩ token in a sentence with a verb–noun expression that fits grammatically and semantically. Mid-size multilingual models required an extended prompt specifying the structural template (verb, optional preposition, noun) and contrasting valid and invalid output formats.

For Mistral-7B-Instruct and DeepSeek-llm-7B no formulation, including system/assistant role splitting and stepwise verb-then-noun generation, produced consistently parsable output, so these models were excluded from this subtask. Representative prompts for all subtasks are given in the Appendix.

3.3. Evaluation of generative subtasks

For paraphrasing and gap filling, surface-form matching is too strict: models often produce valid synonyms or alternative but correct grammatical forms. Therefore an LLM-as-a-judge protocol with YandexGPT5-Lite-8B as the judge was used, scoring each output against the reference on a 0–6 scale (paraphrasing) or a 0–4 scale (gap filling).

The rubric explicitly penalizes non-Russian fragments, single-word outputs where a phrase is expected, over-long fragments that copy parts of the input, and outputs unrelated to the reference; it requires strict JSON (paraphrasing) or a bare integer (gap filling) so that the score can be extracted automatically.

The judge belongs to the same family as one of the evaluated systems, so we took two measures against possible bias.

First, the judge was used only as a scoring function with a fixed rubric and without information about which model produced each output.

Table 2. Classification quality (%) of the base models

Model	Regime	A	P	R	F1
Mistral-7B-Instruct	zero	50.8	50.6	55.7	53.1
	one	51.6	50.8	94.3	66.1
	few	51.1	50.6	86.5	63.8
Qwen2.5-7B-Instruct	zero	52.5	51.4	89.2	65.2
	one	52.5	51.4	92.1	65.9
	few	51.7	50.9	92.1	65.6
DeepSeek-llm-7B-base	zero	49.5	49.7	90.5	64.2
	one	50.0	50.0	100.0	66.6
	few	50.0	50.0	100.0	66.6
Llama-3.1-8B	zero	49.7	50.0	99.3	66.3
	one	49.9	50.0	96.9	65.9
	few	50.0	50.0	98.7	66.3
YandexGPT5-Lite-8B	zero	52.3	51.2	96.8	67.0
	one	53.7	52.0	96.2	67.5
	few	54.4	53.5	67.0	59.5
YandexGPT5-Pro	zero	53.5	52.0	98.5	67.9
	one	54.2	56.2	93.4	70.2
	few	55.3	62.2	53.7	57.6
Llama-3.1-70B	zero	54.5	52.5	94.3	67.4
	one	55.1	52.6	96.7	68.1
	few	56.8	54.3	95.6	69.2

Second, we manually validated a stratified sample of about 300 judge scores: the judge reliably awards high scores to near-synonyms (*osuzhdat'* for the reference *obvinit'* receives 5 of 6 points) and low scores to off-topic outputs (*oboim* for the reference *prepodnosit' podarok* receives 0), with disagreements clustering at borderline cases.

Cross-checking against YandexGPT5-Pro on a subset gave consistent system rankings, so the lighter model was retained as the judge for efficiency. From the judge scores we report two metrics: %COR, the share of outputs scoring at least 5 (paraphrasing) or at least 2 (gap filling), and the Mean Judge Score (MJS).

3.4. Results

Classification results are given in Table 2.

All base models cluster in the 49–57% accuracy band, close to random.

TABLE 3. Paraphrasing quality

Model	zero-shot		one-shot		few-shot	
	MJS	%COR	MJS	%COR	MJS	%COR
Mistral-7B-Instruct	2.2	32.6	4.0	23.8	4.1	29.8
Qwen2.5-7B-Instruct	4.2	33.8	4.2	34.5	4.2	36.5
DeepSeek-llm-7B-base	3.4	9.4	4.2	33.0	4.3	38.0
Llama-3.1-8B	4.3	36.1	4.2	34.4	4.3	37.9
YandexGPT5-Lite-8B	4.8	61.2	4.9	63.3	4.9	66.0
YandexGPT5-Pro	4.9	62.1	4.9	65.6	5.0	67.9
Llama-3.1-70B	4.7	55.0	4.7	54.9	4.7	55.6

TABLE 4. Gap-filling quality

Model	zero-shot		one-shot		few-shot	
	MJS	%COR	MJS	%COR	MJS	%COR
Qwen2.5-7B-Instruct	1.1	18.8	1.2	21.8	1.2	25.1
Llama-3.1-8B	0.8	10.7	0.9	15.3	1.0	16.3
YandexGPT5-Lite-8B	2.1	58.5	2.2	64.9	2.2	65.2
YandexGPT5-Pro	2.3	67.1	2.4	71.4	2.4	71.3
Llama-3.1-70B	1.9	54.7	2.0	56.3	2.0	58.2

Recall is consistently far higher than Precision, indicating a strong bias toward the positive (collocation) class. Few-shot prompting offers only limited and inconsistent gains, and for the Russian-specialized models it actually trades recall for precision without improving F1.

Tables 3 and 4 present the results on the generative subtasks. Mid-size multilingual models perform weakest, with frequent non-Russian fragments and format violations (most severely Mistral and, in zero-shot, DeepSeek). Russian-specialized models clearly lead, with YandexGPT5-Lite-8B nearly matching the much larger YandexGPT5-Pro.

Moving to few-shot prompting never lowers MJS relative to zero-shot for any model and helps the weaker models most: for DeepSeek on paraphrasing, %COR rises from 9.4 in zero-shot to 38.0 in few-shot. Llama-3.1-70B improves over its 8B counterpart but does not reach the Russian-specialized models, confirming that for non-compositional Russian phenomena language specialization is at least as important as raw scale.

4. Error analysis

4.1. Classification: positive-class bias

A qualitative review of misclassifications reveals three recurrent failure modes.

First, models over-predict the positive class for frequent verbs that are widely represented in dictionaries of set expressions (*brat'*, *vesti*, *davat'*), i.e., they use verb identity as a heuristic.

Second, models conflate compositional light-verb constructions (e.g., *provodit' issledovanie*, 'conduct research') with non-compositional collocations, suggesting that surface distributional cues dominate.

Third, for nouns in morphologically marked forms (for instance, abstract nouns in the instrumental case) models return inconsistent labels on near-identical inputs, pointing to instability under inflection rather than a stable decision boundary.

Together these observations explain why accuracy stays near chance: the cause is not only prompt sensitivity but genuine difficulty with the intermediate zone of non-compositionality, where neither compositional decomposition nor whole-idiom recall is sufficient. The high-recall, low-precision profile in Table 2 is the quantitative illustration of the same bias: the models find almost all true collocations but drag many compositional combinations in with them.

4.2. Generation: language mixing and format violation

On paraphrasing and gap filling, the dominant failure of the mid-size multilingual models is formal rather than conceptual. Mistral and Llama-3.1-8B frequently emit English tokens or word fragments even when the prompt forbids non-Russian output, and Qwen occasionally switches to Chinese; the judge scores such outputs as zero, which accounts for much of their low %COR.

A second formal failure, specific to gap filling, is structural: the models return only a verb or only a noun, or copy a fragment of the sentence, instead of a well-formed verb–noun phrase. This is why an extended template-and-contrast prompt was needed for Qwen and Llama, whereas no suitable prompt was found for Mistral or the base DeepSeek model.

The base DeepSeek model shows the clearest regime effect: in zero-shot it echoes the instruction and scores 9.4% %COR on paraphrasing, but with

a paired-example prompt and constrained decoding (parameters are given in the Appendix) it reaches 38.0%. Much of its apparent weakness is therefore a prompt-format artifact rather than a lack of lexical knowledge.

4.3. Generation: semantic near-misses

A separate class of apparent errors are semantic near-misses that surface-form matching would wrongly reject but the judge correctly credits: a model may answer with a synonym that differs in aspect or reflexivity, or with a very close verb.

For the reference *obvinit'* ('accuse') a model returns *osuzhdat'* ('condemn'), scored 5 of 6; for the collocation *brat' za pravilo* ('make it a rule') the answer *sledovat'* ('follow') matches the reference.

These cases motivate the LLM-as-a-judge protocol: without it, the generative competence of the stronger models would be systematically underestimated.

5. Parameter-efficient fine-tuning

5.1. Setup

To assess domain adaptation, we fine-tuned two mid-size models with complementary specialization: the multilingual Qwen2.5-7B-Instruct [11] and the Russian-specialized YandexGPT5-Lite-8B. Both models are openly available, which gives full control over training, and their moderate size fits a dataset of this scale, on which larger models are prone to overfitting.

Data were split preserving per-verb and per-class balance, yielding 2,358 training and 906 test examples (73% and 27%). Verbs with a single example of a class, and the two verbs that have only negative examples, were routed to the training split so that the test split stays balanced (451 positive and 455 negative examples).

LoRA [14] with 4-bit NF4 quantization [15], training each model was used on a single dataset for classification, paraphrasing and gap filling at once. Joint training on the three subtasks is more compute-efficient and encourages cross-task transfer. All tasks were merged into one HuggingFace dataset in which each item carries a task tag, the task-specific prompt, the reference answer and the expression; items were shuffled before training.

Hyperparameters: rank $r = 16$, $\alpha = 32$, dropout 0.05, learning rate $2 \cdot 10^{-4}$, batch size 4 with gradient accumulation to an effective batch of 16,

Table 5. Best results of all models on the three tasks (%; MJS in points)

Model	Classification				Paraphrasing		Gap filling	
	A	P	R	F1	MJS	%COR	MJS	%COR
Llama-3.1-8B	49.7	50.0	99.3	66.3	4.3	37.9	1.0	16.3
DeepSeek-7B	50.0	50.0	100.0	66.6	4.3	38.0	—	—
Mistral-7B	51.6	50.8	94.3	66.1	4.1	29.8	—	—
Qwen2.5-7B	52.5	51.4	92.1	65.9	4.2	36.5	1.2	25.1
YandexGPT-8B	53.7	52.0	96.2	67.5	4.9	66.0	2.2	65.2
Llama-3.1-70B	56.8	54.3	95.6	69.2	4.7	55.6	2.0	58.2
YandexGPT-Pro	54.2	56.2	93.4	70.2	5.0	**67.9**	2.4	71.4
Qwen-7B-LoRA	80.2	82.2	82.2	79.4	4.9	60.8	2.2	61.2
YandexGPT-8B-LoRA	**80.4**	**82.3**	82.3	**79.6**	**5.1**	67.8	**2.5**	**73.8**

three epochs, weight decay 0.01. Rank 16 balances adapter capacity against the risk of overfitting on a small set; the ratio $r : \alpha = 1 : 2$ gives a scaling factor of 2.0 and stable updates; three epochs sufficed for convergence on a held-out split, while longer schedules showed early overfitting.

Adapters were placed on seven attention and feed-forward modules, which provides the flexibility a multitask objective needs; prompt tokens were masked from the training target so that only answer tokens contribute to the loss.

5.2. Results and analysis

Fine-tuning yields substantial gains on all three tasks (Table 5). For the base models the table reports the best regime (by F1 for classification and by MJS for the generative subtasks); the best value in each column is set in bold, except for Recall, where 100% is achieved by the degenerate "always 1" answer.

Classification accuracy rises from about 53% to about 80%, and Precision and Recall become balanced: the adapted models no longer rely on the positive-class shortcut. On paraphrasing and gap filling, the fine-tuned Qwen surpasses the roughly nine times larger Llama-3.1-70B and approaches YandexGPT5-Pro, while the fine-tuned YandexGPT5-Lite-8B surpasses YandexGPT5-Pro on gap filling in both metrics.

This is direct evidence that targeted adaptation can substitute for a large increase in model scale. Figure 2 visualizes the effect of fine-tuning for both models.

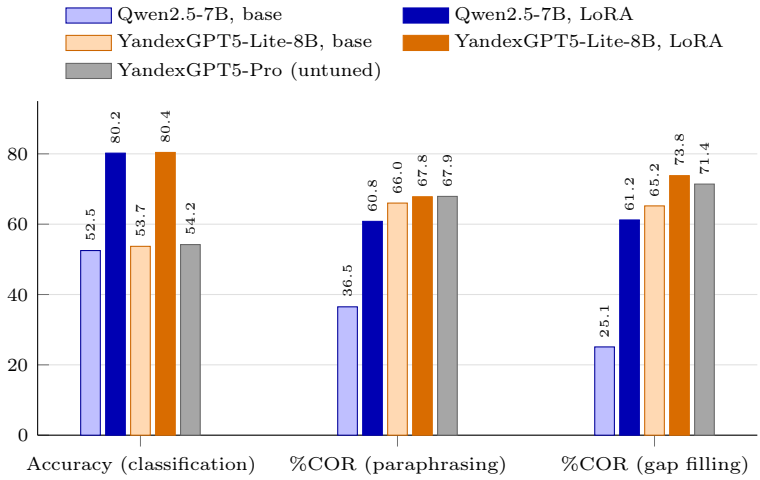


FIGURE 2. Quality of the two mid-size models before and after LoRA fine-tuning compared with the untuned YandexGPT5-Pro (best regime for base models, %)

TABLE 6. Paraphrasing: outputs of the base and LoRA-fine-tuned models

Collocation	Model	Base	LoRA
vklyuchit' um reference: *podumat'*	Qwen2.5-7B-Instruct	*aktivirovat'*	*podumat'*
	YandexGPT5-Lite-8B	*poumnet'*	*soobrazit'*
nabirat'sya smelosti reference: *reshat'sya*	Qwen2.5-7B-Instruct	*priobretat'*	*smelet'*
	YandexGPT5-Lite-8B	*priobretat'*	*reshat'sya*
zaglazhivat' ugly reference: *smyagchat'*	Qwen2.5-7B-Instruct	*shlifovat'*	*uladit'*
	YandexGPT5-Lite-8B	*sglazhivat' protivorechiya*	*sglazhivat'*
ukhodit' iz polya zreniya reference: *skryt'sya*	Qwen2.5-7B-Instruct	*vykhodit'*	*teryat'sya*
	YandexGPT5-Lite-8B	*skryvat'sya*	*skryvat'sya*

A qualitative inspection shows that, on top of the accuracy gains, the most marked improvements concern format compliance and the elimination of language mixing; semantic precision also improves. Tables 6 and 7 give side-by-side outputs of the base and fine-tuned versions of both models. On paraphrasing, for *vklyuchit' um* ('engage one's mind') the base Qwen

Table 7. Gap filling: outputs of the base and LoRA-fine-tuned models

Qwen2.5-7B-Instruct		YandexGPT5-Lite-8B	
Base	LoRA	Base	LoRA
On imel obyknovenie ⟨MASK⟩*, dazhe ne znaya vsey pravdy.* Reference: *rasprostranyat' slukhi*			
vynosit' verdikt	*rasprostranyat' slukhi*	*govorit' namyokami*	*raspuskat' slukhi*
. . . ey prishlos' ⟨MASK⟩*, svyazannye s yazykovym bar'erom i odinochestvom.* Reference: *preodolevat' trudnosti*			
stolknut'sya s trudnostyami	*perezhivat' trudnosti*	*stolknut'sya s (problemami)*	*preodolevat' trudnosti*
Postoyannye obstrely nachali ⟨MASK⟩ *nekogda protsvetayushchiy gorod.* Reference: *prevrashchat' v razvaliny*			
nachinat' razrushenie	*svodit' k ruinam*	(irrelevant continuation)	*privodit' v upadok*
On slishkom mnogo perezhil, chtoby ne ⟨MASK⟩ *nastoyashchey druzhbe.* Reference: *znat' tsenu*			
perestat' doveryat'	*znat' tsenu*	*tsenit' nastoyashchuyu druzhbu*	*utratit' veru v*

returns *aktivirovat'* ('activate'), while the fine-tuned Qwen returns the reference *podumat'* ('think'); the Russian-specialized model shows the same shift (*poumnet'* → *soobrazit'*).

On gap filling, for the sentence "*Postoyannye obstrely nachali* ⟨MASK⟩ *nekogda protsvetayushchiy gorod*" ('Constant shelling began to ⟨MASK⟩ the once-flourishing city'), the base YandexGPT5-Lite-8B produces an irrelevant continuation while its fine-tuned version returns *privodit' v upadok* ('bring into decline'), and the fine-tuned Qwen returns *svodit' k ruinam* ('reduce to ruins'), close to the reference *prevrashchat' v razvaliny* ('turn into rubble').

Beyond individual items, the fine-tuned models follow the required answer format far more reliably, which is itself a large part of the measured gain.

6. Consolidated comparison and discussion

Reading Table 5 across all models and tasks, three patterns stand out.

First, on classification the two LoRA-adapted mid-size models clearly outperform every untuned model, including the 70B one, which shows that for this task targeted adaptation yields a larger gain than model scale or language specialization.

Second, among untuned models, Russian-specialized systems beat multilingual systems of the same size, and larger versions of the same model beat smaller ones, so on classification both language specialization and scale help.

Third, on the harder generative tasks the ordering is different: the Russian-specialized models lead among untuned systems; the LoRA-adapted Russian 8B model matches or slightly surpasses the larger untuned YandexGPT5-Pro on MJS; and the LoRA-adapted multilingual Qwen catches up with the untuned YandexGPT5-Lite-8B on MJS and overtakes Llama-3.1-70B.

In other words, LoRA is effective for all three tasks and most decisively for the simpler classification task, while the advantage of Russian-specialized models on paraphrasing and gap filling confirms that these generative tasks demand a deeper command of Russian morphology and semantics than classification.

Conclusion

The article introduce a methodologically balanced dataset of 3,264 Russian verb–noun collocations and non-collocations covering 479 verbs. It is built from a specialized dictionary, the Russian National Corpus and GPT-4o augmentation, and supports three interconnected subtasks: classification, single-verb paraphrasing and gap filling.

The evaluation of seven multilingual and Russian-specialized LLMs in zero-shot, one-shot and few-shot regimes showed that general-purpose models without adaptation classify near chance and have limited robustness in generation, with substantial dependence on scale and language specialization. Targeted LoRA-based adaptation of mid-size models closes most of the gap and in places surpasses much larger general-purpose models.

Future work includes enlarging the dataset from further sources, covering additional types of set expressions, extending the model pool, and comparing parameter-efficient methods against full fine-tuning on a larger corpus.

References

[1] V. Krasnobaeva. "Automatic processing of Russian verb-noun collocations using large language models", Proceedings of the International Conference "Dialogue 2026" (Moscow, Russia, 24–26 June 2026), Computational Linguistics and Intellectual Technologies, vol. **24**, MAKS Press, M., 2026, 978-5-317-07611-5, pp. 311–317.

[2] H. Haagsma, J. Bos, M. Nissim. "MAGPIE: A large corpus of potentially idiomatic expressions", *Proceedings of the Twelfth Language Resources and Evaluation Conference*, LREC 2020 (Marseille, France, 11–16 May 2020), European Language Resources Association, 2020, 979-10-95546-34-4, pp. 279–287.

[3] H. Tayyar Madabushi, E. Gow-Smith, C. Scarton, A. Villavicencio. "AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in pre-trained language models", *Findings of the Association for Computational Linguistics: EMNLP 2021* (Punta Cana, Dominican Republic, 7–11 November 2021), ACL, 2021, 978-1-7138-9164-2, pp. 3464–3477.

[4] T. Chakrabarty, A. Saakyan, D. Ghosh, S. Muresan. "FLUTE: Figurative language understanding through textual explanations", *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2022 (Abu Dhabi, United Arab Emirates, 7–11 December 2022), ACL, 2022, 978-1-959429-40-1, pp. 7139–7159.

[5] H. Tayyar Madabushi, E. Gow-Smith, M. Garcia, C. Scarton, M. Idiart, A. Villavicencio. "SemEval-2022 Task 2: Multilingual idiomaticity detection and sentence embedding", *Proceedings of the 16th International Workshop on Semantic Evaluation*, SemEval-2022 (Seattle, Washington, United States, 14–15 July 2022), ACL, 2022, 978-1-955917-80-3, pp. 107–121.

[6] D. Phelps, T. Pickard, M. Mi, E. Gow-Smith, A. Villavicencio. "Sign of the times: Evaluating the use of large language models for idiomaticity detection", *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024* (Torino, Italia, 25 May 2024), ELRA and ICCL, 2024, 978-2-493814-20-3, pp. 178–187, 2405.09279.

[7] J. Zhou, Z. Zeng, H. Gong, S. Bhat. "Enhancing language models with idiomatic reasoning", *Proceedings of the 1st Conference on Language Modeling*, COLM 2024 (University of Pennsylvania, Philadelphia, PA, 7–9 October 2024), 2024, 19 pp.

[8] V. M. Deribas. *Stable Verbal-Nominal Collocations of the Russian Language*, Russkij yazyk, M., 1979, 252 pp. (in Russian).

[9] R. Islam, O. M. Moushi. "GPT-4o: The cutting-edge advancement in multimodal LLM", *Intelligent Computing. CompCom 2025*, Proceedings

of the 2025 Computing Conference. V. 4 (London, United Kingdom, 19–20 June 2025), Lecture Notes in Networks and Systems, vol. **1426**, Springer, Cham, 2025, 978-3-031-92610-5, pp. 47–60 .

[10] 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. Le Scao, T. Lavril, T. Wang, T. Wang, W. El Sayed. *Mistral 7B*, 2023, 9 pp., 2310.06825

[11] 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, Z. Qiu. *Qwen2.5 technical report*, 2024, 26 pp., 2412.15115

[12] X. Bi, D. Chen, G. Chen, S. Chen, D. Dai, C., Deng Ding H., Dong K., Du Q., Fu Z., Gao H., Gao K., Gao W., Ge R., Guan K., Guo D., Guo J., Hao G., Hao Z., He Y., Hu W., Huang P., Li E., Li G., Li J., Li Y., Li Y. K., Liang W., Lin F., Liu A. X., Liu B., Liu W., Liu X., Liu X., Liu Y., Lu H., Lu S., Luo F., Ma S., Nie X., Pei T., Piao Y., Qiu J., Qu H., Ren T., Ren Z., Ruan C., Sha Z., Shao Z., Song J., Su X., Sun J., Sun Y., Tang M., Wang B., Wang P., Wang S., Wang Y., Wang Y., Wu T., Wu Y., Xie X., Xie Z., Xie Z., Xiong Y., Xu H., Xu R. X., Xu Y., Yang D., You Y., Yu S., Yu X., Zhang B., Zhang H., Zhang L., Zhang L., Zhang M., Zhang M., Zhang W., Zhang Y., Zhao C., Zhao Y., Zhou S., Zhou S., Zhu Q., Zou Y.. *DeepSeek LLM: Scaling open-source language models with longtermism*, 2024, 48 pp., 2401.02954

[13] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, et al.. *The Llama 3 herd of models*, 2024, 92 pp., 2407.21783

[14] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen. "LoRA: Low-rank adaptation of large language models", *The Tenth International Conference on Learning Representations*, ICLR 2022 (Virtual Event, 25–29 April 2022), ICLR, 2022, 979-8-3313-0009-8, 2106.09685.

[15] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer. "QLoRA: Efficient finetuning of quantized LLMs", *NIPS'23: Proceedings of the 37th International Conference on Neural Information Processing Systems*, NeurIPS 2023 (New Orleans, LA, USA, 10–16 December 2023), Advances in Neural Information Processing Systems, vol. **36**, Curran Associates, Inc., 2023, 9781713899921, pp. 10088–10115, id. 441, 2305.14314.

Appendix. Representative prompts

The prompts are abridged and translated into English; the full Russian originals are released with the dataset.

Classification (zero-shot). "Below is a verb–noun combination. Decide whether it is a collocation. Reply with ONLY `1` (collocation) or `0` (non-collocation)."

Paraphrasing (Russian-specialized models and Llama-3.1-70B). "Below is a verb–noun combination. Replace it with a single verb of the same meaning. Reply with ONLY the verb." For mid-size multilingual models the line "Your answer must be ONLY in Russian" was appended.

Paraphrasing (DeepSeek-llm-7B-base). The instruction-style prompt failed, as the model echoed the examples, so the prompt differed by regime. The zero-shot regime used forced-continuation completion: "Question: How can '{col}' be replaced by a single verb? Answer: '{col}' can be replaced by the verb "', with the trailing open quote priming the verb. The one-shot and few-shot regimes used a list of paired examples:

> Replace a verb–noun combination with a single verb.
> Examples:
> *vskryvat' prichiny* = *vyyasnyat'*
> ...
> {col} =

With the decoding parameters

```
do_sample=False, num_beams=3,
repetition_penalty=1.5, max_new_tokens=7
```

this gave a one-word verb in nearly all cases.

Gap filling (extended prompt for mid-size multilingual models). "You are a specialist in Russian set expressions. Fill ⟨`MASK`⟩ with a SET PHRASE consisting of a VERB and a NOUN (optional preposition). CORRECT: *razbudit' interes*. INCORRECT: *pozhertvovat'* (verb only). The phrase must be natural Russian and fit the sentence. Reply ONLY in Russian."

Augmentation of negative examples (GPT-4o). "A collocation is a stable two-word expression whose meaning is non-compositional and which can be replaced by a single verb. Collocations: *vesti diskussiyu, vzvalivat' bremya, vpadat' v isteriku* (= *diskutirovat', obremenyat', isterit'*). Non-collocations are compositional (*gotovit' obed, organizovyvat' meropriyatie*) and cannot be paraphrased with one verb. Generate 100 verb–noun pairs that are *not* collocations." The outputs were filtered manually.

Generation of gap sentences (GPT-4o). "For each collocation, produce one Russian sentence of medium length, natural and grammatical, with enough context that the collocation is uniquely recoverable. Use the correct aspect. Return two columns: the sentence with ⟨MASK⟩ in place of the collocation, and the same sentence with the collocation restored."

Judge for paraphrasing (0–6 scale). "Evaluate the semantic and grammatical proximity of two Russian verbs. *Semantic* (0–3): 3 = identical; 2 = close synonyms interchangeable in one sentence (*vzglyanut'/posmotret'*); 1 = shared meaning with notable differences (*zashchishchat'/tsenit'*); 0 = unrelated or non-Russian. *Grammatical* (0–3): 3 = identical form; 2 = minor differences (aspect, reflexivity); 1 = substantial differences; 0 = not a verb or non-Russian. Return strict JSON with the two scores and their sum."

Judge for gap filling (0–4 scale). "Evaluate the proximity of two text fragments (predicted vs. reference). 4 = identical; 3 = very close in meaning, same length, interchangeable (*prinyat' na sebya otvetstvennost' / vzyat' vinu na sebya*); 2 = the same words in different forms (*bit' trevogu / zabit' trevogu*); 1 = loosely related (*derzhat' uchet / berech' den'gi*); 0 = unrelated, OR a single word, OR too long, OR contains non-Russian inserts. Return only the integer score."

Received 11.10.2026;
approved after reviewing 13.10.2026;
accepted for publication 17.10.2026;
published online 06.11.2026.

Recommended by *Prof. N. V. Loukashevich*

Information about the author:



Vera Pavlovna Krasnobaeva

Master's student in Artificial Intelligence at the AI Talent Hub, ITMO University; NLP engineer and researcher. Research interests: natural language processing, large language models, non-compositional language, parameter-efficient fine-tuning, AI security and safety

0009-0000-3109-9101

e-mail: krasnobaevavera2022@gmail.com

The author declare no conflicts of interests.

RU EN ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ ISSN 2079-3316
научная статья ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И МАШИННОЕ ОБУЧЕНИЕ

УДК 004.912:81'322.2
10.25209/2079-3316-2026-17-5-3-42

Обучение больших языковых моделей обработке русских глагольно-именных коллокаций

Вера Павловна **Краснобаева**✉
Университет ИТМО, Санкт-Петербург, Россия
✉*krasnobaevavera2022@gmail.com*

Аннотация. Некомпозициональная лексика – коллокации, идиомы, глагольно-именные сочетания – остаётся слабым местом больших языковых моделей (LLM), особенно для морфологически богатых языков, таких как русский, где граница между композициональным и некомпозициональным прочтением градуальна. Русские глагольно-именные коллокации рассматриваются как контролируемый полигон для изучения этой границы.

Построен методологически сбалансированный датасет из 3 264 выражений на основе 479 глаголов с равным числом коллокаций и композициональных сочетаний и балансом классов для каждого глагола; над одними и теми же выражениями определены три подзадачи: бинарная классификация, перефразирование одним глаголом и заполнение пропуска. Семь мультиязычных и русскоязычных LLM оценены в режимах zero-shot, one-shot и few-shot; две модели среднего размера дообучены методом LoRA в мультизадачной постановке.

Базовые модели классифицируют выражения на уровне, близком к случайному, и неустойчивы в генерации; размер и языковая специализация помогают, но не устраняют разрыв. Дообучение повышает точность классификации примерно с 53 до 80%, а на генеративных подзадачах дообученная модель на 8 млрд параметров сравнивается с гораздо более крупными моделями или превосходит их.

Статья является расширенной версией статьи, опубликованной в трудах конференции «Диалог-2026»; датасет и код находятся в открытом доступе. (*Связанные тексты статьи на английском и на русском языках*)

Ключевые слова и фразы: глагольно-именные коллокации, некомпозициональность, распознавание идиоматичности, большие языковые модели, параметрически эффективное дообучение, LoRA, перефразирование, заполнение пропусков, обработка русского языка

Для цитирования: Краснобаева В. П. *Обучение больших языковых моделей обработке русских глагольно-именных коллокаций* // Программные системы: теория и приложения. 2026. **Т. 17**. № 5(74). С. 3–42. (*Англ.+русс.*) https://psta.psiras.ru/read/psta2026_5_3-42.pdf

© Краснобаева В. П. 2026 CC BY

Введение

Автоматическая обработка идиоматичного и некомпозиционального языка остаётся одной из наиболее устойчивых проблем компьютерной лингвистики. Хотя современные большие языковые модели демонстрируют высокую общую лингвистическую компетентность, их способность надёжно распознавать и интерпретировать некомпозициональные выражения изучена недостаточно.

Русские глагольно-именные коллокации представляют собой особенно показательный случай: в них сочетаются лексические, семантические и грамматические ограничения, требующие от модели выйти за пределы поверхностной композициональной интерпретации. Так, коллокация *одержать победу* не сводится к сумме значений своих компонентов, но не является и застывшей идиомой: она изменяется по формам, допускает ограниченную модификацию и перефразируется одним глаголом (*победить*).

Исследования распознавания идиоматичности проводились преимущественно на английском материале и в основном в постановке классификации; большого методологически сбалансированного корпуса русских глагольно-именных коллокаций до сих пор нет. Большинство работ ориентировано на полностью идиоматичные выражения, из-за чего промежуточная зона остаётся малоизученной. Лишь немногие исследования объединяют классификацию, генерацию и восстановление в единой постановке, поэтому способности к распознаванию и к генерации обычно измеряются на разных данных и не поддаются прямому сравнению.

Цель работы – построить сбалансированный датасет русских глагольно-именных коллокаций, определить над одними и теми же выражениями три взаимосвязанные подзадачи, систематически оценить на них мультиязычные и русскоязычные LLM в режимах zero-shot, one-shot и few-shot и исследовать, в какой мере параметрически эффективная доменная адаптация методом LoRA устраняет разрыв, остающийся после увеличения размера модели и её языковой специализации. Датасет и код опубликованы по адресу `https://github.com/Vera-bahval/russian-verb-noun-collocations`.

Отличия от версии для конференции. Настоящая статья является расширенной версией статьи «Automatic Processing of Russian Verb-Noun Collocations Using Large Language Models», опубликованной в трудах конференции «Диалог-2026» [1]. В ней были представлены датасет, постановка из трёх задач, оценка семи LLM и результаты дообучения. Датасет, определения задач, набор моделей, гиперпараметры дообучения и ключевые метрики не изменились, поэтому числа в обеих публикациях согласованы. Новыми являются:

(*1*) подробное описание методологии построения датасета: рабочего определения коллокации, роли исходного словаря, устройства запросов к Национальному корпусу русского языка, аугментации и балансировки (раздел 2);

(*2*) разбор характерных ошибок моделей общего назначения по каждой задаче с примерами (раздел 4);

(*3*) сопоставление ответов базовых и дообученных версий обеих моделей на генеративных подзадачах (раздел 5);

(*4*) сводное сравнение всех моделей на всех задачах и обсуждение взаимодействия размера модели, языковой специализации и целевой адаптации (раздел 6).

1. Обзор существующих работ

Основополагающим шагом в исследованиях идиоматичности стало создание больших размеченных корпусов. Корпус MAGPIE [2] объединил извлечение кандидатов по словарям, автоматический поиск и многоуровневую краудсорсинговую разметку английских выражений; он содержит более 50 000 размеченных вхождений и закрепил связку «словари и большие корпуса» как рабочий конвейер построения данных.

В работе AStitchInLanguageModels [3] была введена многоуровневая классификация многословных выражений, перефразирование использовалось как индикатор семантической чувствительности в режимах zero-shot, one-shot и few-shot; кроме того, было показано, что специальная токенизация целевого выражения и режим обучения влияют на измеряемое качество.

В работе FLUTE [4] понимание образного языка сформулировано как задача логического следования (NLI) с генерацией объяснений; предложения-кандидаты порождались моделью GPT-3 и затем проверялись экспертами.

Соревнование SemEval-2022 Task 2 [5] задало мультиязычный бенчмарк распознавания идиоматичности для английского, португальского и галисийского языков. Лидировали адаптированные энкодерные архитектуры, а победившее решение в режиме one-shot опиралось на ансамблирование, состязательное и контрастивное обучение и явную разметку многословного выражения.

Д. Фелпс и соавторы [6] показали, что LLM общего назначения без адаптации уступают меньшим энкодерным моделям, дообученным под задачу; вместе с тем способность LLM отличать идиоматичное употребление от буквального без целевого обучения свидетельствует об их общей латентной компетентности, причём качество растёт с числом параметров.

Ц. Чжоу и соавторы [7] объединили генерацию определения идиомы, её идентификацию и восстановление в единой постановке и адаптировали модель с помощью параметрически эффективных модулей вместо полного дообучения; именно это разделение на три подзадачи послужило прообразом трёх задач нашей работы.

Из этих работ следуют два вывода, определивших наш подход:

Во-первых, словари и большие эталонные корпуса служат естественной основой для построения ресурсов некомпозициональной лексики, а аугментация с помощью LLM масштабирует процесс при умеренном ручном редактировании.

Во-вторых, некомпозициональность лучше всего проверять, сочетая распознавание и генерацию и тестируя несколько моделей в нескольких режимах.

Вместе с тем отсутствует большой методологически проверенный корпус русских глагольно-именных коллокаций, промежуточная зона между композициональными и идиоматичными выражениями остаётся малоизученной, и лишь немногие работы объединяют классификацию, перефразирование и восстановление над одними и теми же единицами. Наша работа направлена на устранение этих пробелов.

2. Датасет

2.1. Рабочее определение

Глагольно-именной коллокацией в данной работе называется сочетание глагола и существительного (прямого или косвенного дополнения, возможно, отделённого одним предлогом), значение которого некомпозиционально. Диагностическим критерием служит возможность перефразировать всю конструкцию одним синонимичным глаголом: *одержать победу* удовлетворяет критерию, поскольку сводится к *победить*, а *бросить взгляд* – к *посмотреть*, тогда как сочетание *исследовать вопрос* композиционально. Критерий выбран намеренно: он выделяет именно промежуточную зону между полностью композициональными и полностью идиоматичными выражениями.

2.2. Источники и построение

Исходным источником положительных примеров послужил словарь «Устойчивые глагольно-именные словосочетания русского языка» [8], содержащий более 6 000 глагольно-именных выражений с информацией об управлении. Узкая направленность словаря и явная грамматическая информация сделали его удобным стартовым списком, однако он использовался лишь как основа: все кандидаты проверялись по современным

корпусным данным и по критерию перефразирования, так что выражения, отсутствующие или маргинальные в современном употреблении либо не сводимые к одному глаголу (например, *брать такси*), исключались. Автоматическая предобработка (регулярные выражения и строковые операции) оставляла глагол, существительное и, при наличии, единственный предлог; если в словаре приводились обе видовые формы глагола, сохранялась форма несовершенного вида.

Национальный корпус русского языка[URL] (НКРЯ) использовался и как современный частотный фильтр, и как источник новых положительных и отрицательных кандидатов. Грамматические признаки в запросах к корпусу отражали синтаксический профиль каждого глагола: падеж соответствовал управлению глагола; единственный промежуточный предлог допускался только для предложных дополнений; число не ограничивалось, чтобы охватить морфологическую вариативность.

Для каждого глагола вручную просматривались 100 наиболее частотных конструкций: выражения, удовлетворяющие критерию перефразирования одним глаголом, размечались как коллокации (метка 1), остальные – как композициональные примеры (метка 0). Например, для глагола *водить* сочетание *водить за нос* ($\rightarrow$ *обманывать*) добавлено как положительный пример в силу частотности и однозначного перефразирования одним глаголом, а *водить кистью* – как отрицательный, будучи и менее частотным, и композициональным.

Поскольку эта процедура давала избыток положительных примеров, множество балансировалось аугментацией данных с помощью модели GPT-4o [9]: модели предъявлялись явное определение и примеры обоих классов и предлагалось порождать глагольно-именные сочетания пакетами по 100 выражений (промпт приведён в приложении).

Все выходы вручную проверялись на композициональность, грамматичность и отсутствие дубликатов; отбрасывались, например, порождения, сами являющиеся коллокациями (*водить дружбу*), и семантически странные сочетания. Так было добавлено 983 проверенных отрицательных примера.

2.3. Балансировка и итоговый состав

На уровне отдельных глаголов сохранялся остаточный дисбаланс: у некоторых глаголов примеров одного класса было заметно больше, чем другого, что не позволило бы корректно разбить данные на обучающую и тестовую выборки. Поэтому мы обеспечили баланс для каждого глагола: для 479 глаголов, имевших хотя бы один положительный пример, оставлялось равное число коллокаций и неколлокаций (от 2 до 8 примеров

на глагол), а недостающие классы пополнялись повторным поиском по НКРЯ и аугментацией с помощью GPT-4o.

Итоговый корпус содержит 3 264 выражения: 1 630 коллокаций и 1 634 неколлокации, сбалансированных по классу и по глаголу. Схема построения датасета показана на рисунке 1.

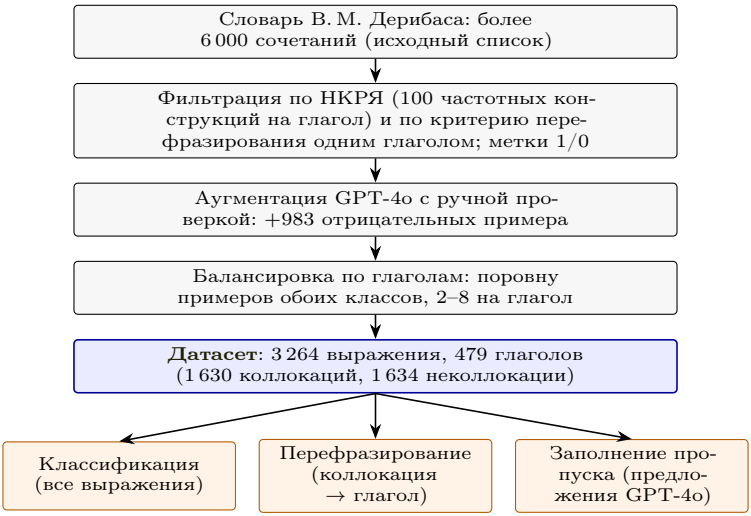


Рисунок 1. Схема построения датасета и три подзадачи над одними и теми же выражениями

Для подзадачи перефразирования каждой коллокации вручную сопоставлялся синонимичный глагол (например, *задать направление* → *направить*, *заливаться краской* → *краснеть*, *знать толк* → *разбираться*). Для подзадачи заполнения пропуска контекстные предложения с замаскированной коллокацией порождались для положительных примеров моделью GPT-4o при контролируемом промптинге и затем проверялись вручную; отрицательные примеры в этой подзадаче не используются. Образцы записей приведены в таблице 1.

Три подзадачи проверяют взаимодополняющие способности:

классификация – бинарное распознавание некомпозициональности,

перефразирование – лексико-семантическое понимание,

заполнение пропуска – контекстное согласование формы, значения и грамматического оформления.

Поскольку подзадачи определены над одними и теми же выражениями, распознавание и генерацию можно измерять на общей основе. Ресурс может использоваться в задачах распознавания идиоматичности, генерации

Таблица 1. Образцы записей датасета

Выражение	Метка	Синоним	Предложение с пропуском
выносить обвинение	1	*обвинять*	*После долгого следствия прокуратура решила* ⟨MASK⟩ *подозреваемому по трём статьям.*
выносить порицание	1	*порицать*	*За аморальное поведение ему было решено* ⟨MASK⟩ *перед всеми коллегами.*
выносить пытку	0	–	–

парафраз, контролируемой оценки генерации текста и в контрастивных исследованиях.

3. Оценка больших языковых моделей

3.1. Модели и постановка

Оценивались семь LLM, различающихся размером, языковой специализацией и режимом настройки: Mistral-7B-Instruct-v0.1 [10], Qwen2.5-7B-Instruct [11], DeepSeek-llm-7B-base [12], Llama-3.1-8B-Instruct и Llama-3.1-70B-Instruct [13], а также *YandexGPT5-Lite-8B* и YandexGPT5-Pro.

Такой отбор позволяет разделить эффекты размера (от 7 до 70 млрд параметров), языковой специализации (мультиязычные и русскоязычные модели) и инструктивной настройки; базовая авторегрессионная модель DeepSeek-llm-7B-base включена для контраста с инструктивными системами. Модели до 8 млрд параметров запускались локально с помощью библиотеки HuggingFace Transformers (классы `AutoModelForCausalLM` и `AutoTokenizer`), более крупные модели – через API Yandex Cloud. Каждая модель тестировалась в режимах zero-shot, one-shot и few-shot; примеры для контекста отбирались из датасета в соответствии с подзадачей.

3.2. Спецификация задач и промптинг

Классификация. Модель получает глагольно-именное сочетание и предсказывает метку из множества $\{0, 1\}$, возвращая только цифру. В режиме one-shot предъявлялись два примера (по одному на класс), поскольку единственный пример смещал бы бинарную задачу к одной метке; в режиме few-shot – шесть примеров, сбалансированных по классам и различающихся падежом, числом и наличием предлога.

Рассматривалась доля правильных ответов (Accuracy, A), точность (Precision, P), полноту (Recall, R) и F1-меру для положительного класса.

Перефразирование. Модель получает коллокацию и должна вернуть один синонимичный глагол. Промпты подбирались под каждое семейство моделей: мультиязычным моделям среднего размера требовалось явное указание «отвечать только по-русски», чтобы подавить английские, а у Qwen и китайские вставки. Базовой модели DeepSeek потребовался отдельный подход, поскольку инструктивный промпт вызывал повторение инструкции вместо порождения глагола. Поэтому для неё в режиме zero-shot использовался промпт-продолжение, оканчивающийся открывающей кавычкой, которая «подсказывает» глагол, а в режимах one-shot и few-shot – список пар вида `коллокация = глагол`, оканчивающийся целевой коллокацией и знаком «=».

Заполнение пропуска. Модель заменяет токен ⟨MASK⟩ в предложении глагольно-именным сочетанием, подходящим грамматически и по смыслу. Мультиязычным моделям среднего размера требовался расширенный промпт, задающий структурный шаблон (глагол, необязательный предлог, существительное) и противопоставляющий корректный и некорректный форматы ответа. Для Mistral-7B-Instruct и DeepSeek-llm-7B ни одна формулировка, включая разделение ролей system/assistant и пошаговое порождение «сначала глагол, затем существительное», не давала стабильно разбираемого ответа, поэтому эти модели исключены из данной подзадачи. Представительные промпты для всех подзадач приведены в приложении.

3.3. Оценка генеративных подзадач

Для перефразирования и заполнения пропуска точное совпадение по форме слишком строго: модели часто выдают корректные синонимы или иные, но верные грамматические формы. Поэтому мы применили протокол LLM-as-a-judge: модель-судья YandexGPT5-Lite-8B оценивает каждый ответ относительно эталона по шкале от 0 до 6 (перефразирование) или от 0 до 4 (заполнение пропуска).

Рубрика оценивания явно штрафует нерусские фрагменты, однословные ответы там, где ожидается словосочетание, чрезмерно длинные фрагменты, копирующие части входа, и ответы, не связанные с эталоном; она требует строгого формата JSON (перефразирование) или одного целого числа (заполнение пропуска), чтобы оценку можно было извлекать автоматически.

Судья принадлежит тому же семейству, что и одна из оцениваемых систем, поэтому для исключения возможной предвзятости мы приняли две меры.

Во-первых, модель-судья использовалась только как оценивающая функция с фиксированной рубрикой и без сведений о том, какая модель породила ответ.

Во-вторых, мы вручную проверили стратифицированную выборку примерно из 300 оценок судьи: судья надёжно ставит высокие баллы близким синонимам (*осуждать* для эталона *обвинить* получает 5 баллов из 6) и низкие – ответам не по теме (*обоим* для эталона *преподносить подарок* получает 0 баллов), причём расхождения с ручной оценкой сосредоточены на пограничных случаях.

Перекрёстная проверка на подмножестве данных с помощью YandexGPT5-Pro дала хорошо согласованное ранжирование систем, поэтому из соображений эффективности в качестве судьи сохранена меньшая модель. По оценкам судьи вычисляются две метрики: %COR – доля ответов с оценкой не ниже 5 (перефразирование) или не ниже 2 (заполнение пропуска) – и средняя оценка судьи MJS (Mean Judge Score).

3.4. Результаты

Результаты классификации приведены в таблице 2.

Таблица 2. Качество классификации (%) для базовых моделей

Модель	Режим	A	P	R	F1
Mistral-7B-Instruct	zero	50,8	50,6	55,7	53,1
	one	51,6	50,8	94,3	66,1
	few	51,1	50,6	86,5	63,8
Qwen2.5-7B-Instruct	zero	52,5	51,4	89,2	65,2
	one	52,5	51,4	92,1	65,9
	few	51,7	50,9	92,1	65,6
DeepSeek-llm-7B-base	zero	49,5	49,7	90,5	64,2
	one	50,0	50,0	100,0	66,6
	few	50,0	50,0	100,0	66,6
Llama-3.1-8B	zero	49,7	50,0	99,3	66,3
	one	49,9	50,0	96,9	65,9
	few	50,0	50,0	98,7	66,3
YandexGPT5-Lite-8B	zero	52,3	51,2	96,8	67,0
	one	53,7	52,0	96,2	67,5
	few	54,4	53,5	67,0	59,5
YandexGPT5-Pro	zero	53,5	52,0	98,5	67,9
	one	54,2	56,2	93,4	70,2
	few	55,3	62,2	53,7	57,6
Llama-3.1-70B	zero	54,5	52,5	94,3	67,4
	one	55,1	52,6	96,7	68,1
	few	56,8	54,3	95,6	69,2

Все базовые модели находятся в диапазоне доли правильных ответов 49–57%, близком к случайному.

Полнота устойчиво и заметно превышает точность, что указывает на сильное смещение в сторону положительного класса (коллокаций). Режим few-shot даёт лишь ограниченный и непоследовательный прирост, а у русскоязычных моделей он фактически обменивает полноту на точность, не улучшая F1.

Результаты на генеративных подзадачах представлены в таблицах 3 и 4. Мультиязычные модели среднего размера показывают наиболее

Таблица 3. Качество перефразирования

Модель	zero-shot		one-shot		few-shot	
	MJS	%COR	MJS	%COR	MJS	%COR
Mistral-7B-Instruct	2,2	32,6	4,0	23,8	4,1	29,8
Qwen2.5-7B-Instruct	4,2	33,8	4,2	34,5	4,2	36,5
DeepSeek-llm-7B-base	3,4	9,4	4,2	33,0	4,3	38,0
Llama-3.1-8B	4,3	36,1	4,2	34,4	4,3	37,9
YandexGPT5-Lite-8B	4,8	61,2	4,9	63,3	4,9	66,0
YandexGPT5-Pro	4,9	62,1	4,9	65,6	5,0	67,9
Llama-3.1-70B	4,7	55,0	4,7	54,9	4,7	55,6

слабые результаты, с частыми нерусскими фрагментами и нарушениями формата (сильнее всего у Mistral и, в режиме zero-shot, у DeepSeek). Русскоязычные модели уверенно лидируют, причём YandexGPT5-Lite-8B почти не уступает гораздо более крупной YandexGPT5-Pro.

Таблица 4. Качество заполнения пропуска

Модель	zero-shot		one-shot		few-shot	
	MJS	%COR	MJS	%COR	MJS	%COR
Qwen2.5-7B-Instruct	1,1	18,8	1,2	21,8	1,2	25,1
Llama-3.1-8B	0,8	10,7	0,9	15,3	1,0	16,3
YandexGPT5-Lite-8B	2,1	58,5	2,2	64,9	2,2	65,2
YandexGPT5-Pro	2,3	67,1	2,4	71,4	2,4	71,3
Llama-3.1-70B	1,9	54,7	2,0	56,3	2,0	58,2

Переход к режиму few-shot ни у одной модели не снижает MJS по сравнению с zero-shot и сильнее всего помогает слабым моделям: у DeepSeek на перефразировании %COR растёт с 9,4 в режиме zero-shot до 38,0 в режиме few-shot. Модель Llama-3.1-70B превосходит свой аналог на 8 млрд параметров, но не достигает уровня русскоязычных моделей; это подтверждает, что для некомпозициональных явлений русского языка языковая специализация не менее важна, чем размер модели.

4. Анализ ошибок

4.1. Классификация: смещение к положительному классу

Качественный разбор ошибок классификации выявляет три повторяющихся типа сбоев.

Во-первых, модели переоценивают положительный класс для частотных глаголов, широко представленных в словарях устойчивых сочетаний (*брать*, *вести*, *давать*), т. е. используют идентичность глагола как эвристику.

Во-вторых, модели смешивают композициональные конструкции с семантически ослабленным глаголом (например, *проводить исследование*) с некомпозициональными коллокациями, что указывает на преобладание поверхностных дистрибутивных признаков.

В-третьих, для существительных в морфологически маркированных формах (например, абстрактных существительных в творительном падеже) модели дают несогласованные метки на почти идентичных входах, что указывает на неустойчивость при словоизменении, а не на устойчивую границу решения.

Вместе эти наблюдения объясняют, почему доля правильных ответов держится около случайного уровня: дело не только в чувствительности к промпту, но и в подлинной трудности промежуточной зоны некомпозициональности, где недостаточно ни композиционального разбора, ни воспроизведения идиомы целиком. Профиль «высокая полнота – низкая точность» в таблице 2 количественно иллюстрирует то же смещение: модели находят почти все истинные коллокации, но относят к ним и множество композициональных сочетаний.

4.2. Генерация: смешение языков и нарушение формата

На перефразировании и заполнении пропуска доминирующий сбой мультиязычных моделей среднего размера носит не концептуальный, а формальный характер. Mistral и Llama-3.1-8B часто выдают английские токены или фрагменты слов, даже когда промпт запрещает нерусский ответ, а Qwen иногда переключается на китайский; судья оценивает такие ответы нулём, что объясняет значительную часть их низкого %COR. Второй формальный сбой, специфичный для заполнения пропуска, – структурный: модели возвращают только глагол или только существительное либо копируют фрагмент предложения вместо корректного глагольно-именного словосочетания. Поэтому для Qwen и Llama потребовался расширенный промпт с шаблоном и противопоставлением форматов, тогда как для Mistral и базовой модели DeepSeek подходящего промпта найти не удалось.

Базовая модель DeepSeek демонстрирует наиболее наглядную зависимость от режима: в zero-shot она повторяет инструкцию и получает

9,4% %COR на перефразировании, а с промптом из пар-образцов и ограниченным декодированием (параметры приведены в приложении) достигает 38,0%. Следовательно, значительная часть её видимой слабости — артефакт формата промпта, а не отсутствие лексических знаний.

4.3. Генерация: семантические «почти попадания»

Отдельный класс мнимых ошибок составляют семантические «почти попадания», которые сопоставление по форме ошибочно отвергло бы, но судья корректно засчитывает: модель может ответить синонимом с иной формой вида или возвратности либо очень близким по значению глаголом.

Для эталона *обвинить* модель возвращает *осуждать*, что оценивается в 5 баллов из 6; для коллокации *брать за правило* ответ *следовать* совпадает с эталоном.

Такие случаи обосновывают применение протокола LLM-as-a-judge: без него генеративная компетентность более сильных моделей систематически недооценивалась бы.

5. Параметрически эффективное дообучение

5.1. Постановка

Для оценки доменной адаптации мы дообучили две модели среднего размера с взаимодополняющей специализацией: мультиязычную Qwen2.5-7B-Instruct [11] и русскоязычную YandexGPT5-Lite-8B. Обе модели находятся в открытом доступе, что даёт полный контроль над обучением, а их умеренный размер соответствует объёму датасета: более крупные модели склонны к переобучению на данных такого масштаба.

Данные разбивались с сохранением баланса по глаголам и по классам: 2 358 обучающих и 906 тестовых примеров (73 и 27%). Глаголы с единственным примером какого-либо класса, а также два глагола, имеющих только отрицательные примеры, направлялись в обучающую выборку, чтобы тестовая выборка оставалась сбалансированной (451 положительный и 455 отрицательных примеров).

Использовались LoRA [14] с 4-битной квантизацией NF4 [15], обучая модели на едином датасете сразу для классификации, перефразирования и заполнения пропуска. Совместное обучение на трёх подзадачах экономнее по вычислениям и способствует переносу знаний между задачами. Все задачи объединены в единый датасет формата HuggingFace, каждый элемент которого содержит тег задачи, соответствующий промпт, эталонный ответ и само выражение; перед обучением элементы перемешивались.

Гиперпараметры: ранг $r = 16$, $\alpha = 32$, dropout 0,05, скорость обучения $2 \cdot 10^{-4}$, размер батча 4 с аккумуляцией градиентов до эффективного

размера 16, три эпохи, weight decay 0,01. Ранг 16 уравновешивает ёмкость адаптеров и риск переобучения на небольшом множестве; отношение $r : \alpha = 1 : 2$ даёт коэффициент масштабирования 2,0 и устойчивые обновления; трёх эпох хватило для сходимости на отложенной выборке, а при более длительном обучении наблюдались первые признаки переобучения.

Адаптеры размещались на семи модулях блоков внимания и feed-forward, что обеспечивает гибкость, необходимую в мультизадачной постановке; токены промпта исключались из целевой последовательности, так что в функцию потерь вносят вклад только токены ответа.

5.2. Результаты и анализ

Дообучение даёт существенный прирост на всех трёх задачах (таблица 5). Для базовых моделей в таблице указан лучший режим: по F1 для классификации и по MJS для генеративных подзадач; полужирным выделен лучший результат в столбце (для полноты он не выделяется, поскольку 100% достигается вырожденным ответом «всегда 1»).

Таблица 5. Лучшие результаты моделей на трёх задачах (%, для MJS – баллы)

Модель	Классификация				Перефразирование		Пропуск	
	A	P	R	F1	MJS	%COR	MJS	%COR
Llama-3.1-8B	49,7	50,0	99,3	66,3	4,3	37,9	1,0	16,3
DeepSeek-7B	50,0	50,0	100,0	66,6	4,3	38,0	–	–
Mistral-7B	51,6	50,8	94,3	66,1	4,1	29,8	–	–
Qwen2.5-7B	52,5	51,4	92,1	65,9	4,2	36,5	1,2	25,1
YandexGPT-8B	53,7	52,0	96,2	67,5	4,9	66,0	2,2	65,2
Llama-3.1-70B	56,8	54,3	95,6	69,2	4,7	55,6	2,0	58,2
YandexGPT-Pro	54,2	56,2	93,4	70,2	5,0	**67,9**	2,4	71,4
Qwen-7B-LoRA	80,2	82,2	82,2	79,4	4,9	60,8	2,2	61,2
YandexGPT-8B-LoRA	**80,4**	**82,3**	82,3	**79,6**	**5,1**	67,8	**2,5**	**73,8**

Доля правильных ответов при классификации растёт примерно с 53 до 80%, а точность и полнота выравниваются: адаптированные модели больше не опираются на смещение к положительному классу. На перефразировании и заполнении пропуска дообученная Qwen превосходит примерно в девять раз более крупную Llama-3.1-70B и приближается к YandexGPT5-Pro, а дообученная YandexGPT5-Lite-8B превосходит YandexGPT5-Pro на заполнении пропуска по обеим метрикам.

Это прямое свидетельство того, что целевая адаптация способна заменить значительное увеличение размера модели. Эффект дообучения для обеих моделей наглядно показан на рисунке 2.

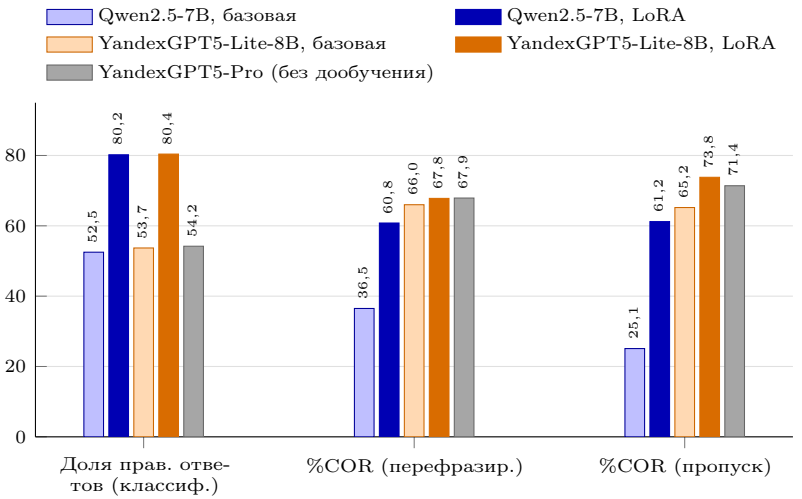


Рисунок 2. Качество двух моделей среднего размера до и после дообучения методом LoRA в сравнении с недообученной YandexGPT5-Pro (лучший режим для базовых моделей, %)

Качественный разбор ответов показывает, что, помимо роста точности, наиболее заметные улучшения касаются соблюдения формата и устранения смешения языков; повышается и семантическая точность ответов. В таблицах 6 и 7 сопоставлены ответы базовых и дообученных версий обеих моделей.

Таблица 6. Перефразирование: ответы базовых и дообученных методом LoRA моделей

Коллокация	Модель	Базовая	LoRA
включить ум эталон: *подумать*	Qwen2.5-7B-Instruct	*активировать*	*подумать*
	YandexGPT5-Lite-8B	*поумнеть*	*сообразить*
набираться смелости эталон: *решаться*	Qwen2.5-7B-Instruct	*приобретать*	*смелеть*
	YandexGPT5-Lite-8B	*приобретать*	*решаться*
заглаживать углы эталон: *смягчать*	Qwen2.5-7B-Instruct	*шлифовать*	*уладить*
	YandexGPT5-Lite-8B	*сглаживать противоречия*	*сглаживать*
уходить из поля зрения эталон: *скрыться*	Qwen2.5-7B-Instruct	*выходить*	*теряться*
	YandexGPT5-Lite-8B	*скрываться*	*скрываться*

На перефразировании для *включить ум* базовая Qwen возвращает *активировать*, а дообученная – эталонный ответ *подумать*; у русскоязычной модели наблюдается тот же сдвиг (*поумнеть* → *сообразить*).

На заполнении пропуска в предложении «*Постоянные обстрелы начали* ⟨MASK⟩ *некогда процветающий город*» базовая YandexGPT5-Lite-8B выдаёт нерелевантное продолжение, а дообученная версия – сочетание *приводить в упадок*; дообученная Qwen возвращает *сводить к руинам*, близкое к эталону *превращать в развалины*.

Таблица 7. Заполнение пропуска: ответы базовых моделей и моделей, дообученных методом LoRA

Qwen2.5-7B-Instruct		YandexGPT5-Lite-8B	
базовая	LoRA	базовая	LoRA
Он имел обыкновение ⟨MASK⟩*, даже не зная всей правды.* Эталон: *распространять слухи*			
выносить вердикт	*распространять слухи*	*говорить намёками*	*распускать слухи*
. . . ей пришлось ⟨MASK⟩*, связанные с языковым барьером и одиночеством.* Эталон: *преодолевать трудности*			
столкнуться с трудностями	*переживать трудности*	*столкнуться с (проблемами)*	*преодолевать трудности*
Постоянные обстрелы начали ⟨MASK⟩ *некогда процветающий город.* Эталон: *превращать в развалины*			
начинать разрушение	*сводить к руинам*	(нерелевантное продолжение)	*приводить в упадок*
Он слишком много пережил, чтобы не ⟨MASK⟩ *настоящей дружбе.* Эталон: *знать цену*			
перестать доверять	*знать цену*	*ценить настоящую дружбу*	*утратить веру в*

Помимо отдельных примеров, дообученные модели гораздо надёжнее соблюдают требуемый формат ответа, что само по себе составляет значительную часть измеренного прироста.

6. Сводное сравнение и обсуждение

Сравнение всех моделей на всех задачах (таблица 5) выявляет три закономерности.

Во-первых, при классификации две модели среднего размера, адаптированные методом LoRA, уверенно превосходят все недообученные

модели, включая модель на 70 млрд параметров; следовательно, для этой задачи целевая адаптация даёт больший прирост, чем размер и языковая специализация.

Во-вторых, среди недообученных моделей русскоязычные превосходят мультиязычные того же размера, а более крупные версии одной модели превосходят меньшие, т. е. при классификации помогают и языковая специализация, и размер.

В-третьих, на более трудных генеративных задачах картина иная: среди недообученных моделей лидируют русскоязычные; дообученная русскоязычная модель на 8 млрд параметров сравнивается с более крупной недообученной YandexGPT5-Pro или слегка превосходит её по метрике MJS, а дообученная мультиязычная Qwen догоняет недообученную YandexGPT5-Lite-8B по MJS и обходит Llama-3.1-70B.

Таким образом, LoRA эффективна для всех трёх задач и наиболее значительно — для более простой задачи классификации, а преимущество русскоязычных моделей на перефразировании и заполнении пропуска подтверждает, что генеративные задачи требуют более глубокого владения русской морфологией и семантикой, чем классификация.

Заключение

Представлен методологически сбалансированный датасет из 3 264 русских глагольно-именных коллокаций и неколлокаций, охватывающий 479 глаголов. Он построен на основе специализированного словаря, Национального корпуса русского языка и аугментации с помощью GPT-4o и поддерживает три взаимосвязанные подзадачи: классификацию, перефразирование одним глаголом и заполнение пропуска.

Оценка семи мультиязычных и русскоязычных LLM в режимах zero-shot, one-shot и few-shot показала, что модели общего назначения без адаптации классифицируют выражения на уровне, близком к случайному, и ограниченно устойчивы в генерации, а их качество существенно зависит от размера и языковой специализации.

Целевая адаптация методом LoRA моделей среднего размера устраняет большую часть разрыва и на отдельных задачах позволяет превзойти гораздо более крупные модели общего назначения.

Дальнейшая работа включает расширение датасета данными из дополнительных источников, охват других типов устойчивых сочетаний, расширение набора моделей и сравнение параметрически эффективных методов с полным дообучением на более объёмном корпусе.

Список использованных источников

[1] Краснобаева В. *Автоматическая обработка русских глагольно-именных коллокаций с помощью больших языковых моделей*, По материалам ежегодной международной конференции «Диалог» (г. Москва, Россия, 24–26 июня 2026 г.), Компьютерная лингвистика и интеллектуальные технологии.– т. **24**, М.: МАКС Пресс.– 2026.– 978-5-317-07611-5.– С. 311–317 (англ.).

[2] Haagsma H., Bos J., Nissim M. *MAGPIE: A large corpus of potentially idiomatic expressions // Proceedings of the Twelfth Language Resources and Evaluation Conference*, LREC 2020 (Marseille, France, 11–16 May 2020).– European Language Resources Association.– 2020.– 979-10-95546-34-4.– Pp. 279–287.

[3] Tayyar Madabushi H., Gow-Smith E., Scarton C., Villavicencio A. *AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in pre-trained language models // Findings of the Association for Computational Linguistics: EMNLP 2021* (Punta Cana, Dominican Republic, 7–11 November 2021).– ACL.– 2021.– 978-1-7138-9164-2.– Pp. 3464–3477.

[4] Chakrabarty T., Saakyan A., Ghosh D., Muresan S. *FLUTE: Figurative language understanding through textual explanations // Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2022 (Abu Dhabi, United Arab Emirates, 7–11 December 2022).– ACL.– 2022.– 978-1-959429-40-1.– Pp. 7139–7159.

[5] Tayyar Madabushi H., Gow-Smith E., Garcia M., Scarton C., Idiart M., Villavicencio A. *SemEval-2022 Task 2: Multilingual idiomaticity detection and sentence embedding // Proceedings of the 16th International Workshop on Semantic Evaluation*, SemEval-2022 (Seattle, Washington, United States, 14–15 July 2022).– ACL.– 2022.– 978-1-955917-80-3.– Pp. 107–121.

[6] Phelps D., Pickard T., Mi M., Gow-Smith E., Villavicencio A. *Sign of the times: Evaluating the use of large language models for idiomaticity detection // Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024* (Torino, Italia, 25 May 2024).– ELRA and ICCL.– 2024.– 978-2-493814-20-3.– Pp. 178–187, 2405.09279.

[7] Zhou J., Zeng Z., Gong H., Bhat S. *Enhancing language models with idiomatic reasoning // Proceedings of the 1st Conference on Language Modeling*, COLM 2024 (University of Pennsylvania, Philadelphia, PA, 7–9 October 2024).– 2024.– 19 pp.

[8] Дерибас В. М. *Устойчивые глагольно-именные словосочетания русского языка.*– М.: Русский язык.– 1979.– 252 с.

[9] Islam R., Moushi O. M. *GPT-4o: The cutting-edge advancement in multimodal LLM* // *Intelligent Computing. CompCom 2025*, Proceedings of the 2025 Computing Conference.– V. 4 (London, United Kingdom, 19–20 June 2025), Lecture Notes in Networks and Systems.– vol. **1426**, Cham: Springer.– 2025.– 978-3-031-92610-5.– Pp. 47–60 .

[10] Jiang A. Q., Sablayrolles A., Mensch A., Bamford C., Chaplot D. S., de las Casas D., Bressand F., Lengyel G., Lample G., Saulnier L., Lavaud L. R., Lachaux M.-A., Stock P., Le Scao T., Lavril T., Wang T., LacroixT., El Sayed W. *Mistral 7B.*– 2023.– 9 pp., 2310.06825

[11] Yang A., Yang B., Zhang B., Hui B., Zheng B., Yu B., Li C., Liu D., Huang F., Wei H., Lin H., Yang J., Tu J., Zhang J., Yang J., Yang J., Zhou J., Lin J., Dang K., Lu K., Bao K., Yang K., Yu L., Li M., Xue M., Zhang P., Zhu Q., Men R., Lin R., Li T., Tang T., Xia T., Ren X., Ren X., Fan Y., Su Y., Zhang Y., Wan Y., Liu Y., Cui Z., Zhang Z., Qiu Z. *Qwen2.5 technical report.*– 2024.– 26 pp., 2412.15115

[12] Bi X., Chen D., Chen G., Chen S., Dai D., Deng C., Ding H., Dong K., Du Q., Fu Z., Gao H., Gao K., Gao W., Ge R., Guan K., Guo D., Guo J., Hao G., Hao Z., He Y., Hu W., Huang P., Li E., Li G., Li J., Li Y., Li Y. K., Liang W., Lin F., Liu A. X., Liu B., Liu W., Liu X., Liu X., Liu Y., Lu H., Lu S., Luo F., Ma S., Nie X., Pei T., Piao Y., Qiu J., Qu H., Ren T., Ren Z., Ruan C., Sha Z., Shao Z., Song J., Su X., Sun J., Sun Y., Tang M., Wang B., Wang P., Wang S., Wang Y., Wang Y., Wu T., Wu Y., Xie X., Xie Z., Xie Z., Xiong Y., Xu H., Xu R. X., Xu Y., Yang D., You Y., Yu S., Yu X., Zhang B., Zhang H., Zhang L., Zhang L., Zhang M., Zhang M., Zhang W., Zhang Y., Zhao C., Zhao Y., Zhou S., Zhou S., Zhu Q., Zou Y. *DeepSeek LLM: Scaling open-source language models with longtermism.*– 2024.– 48 pp., 2401.02954

[13] Grattafiori A., Dubey A., Jauhri A., Pandey A., Kadian A., et al. *The Llama 3 herd of models.*– 2024.– 92 pp., 2407.21783

[14] Hu E. J., Shen Y., Wallis P., Allen-Zhu Z., Li Y., Wang S., Wang L., Chen W. *LoRA: Low-rank adaptation of large language models* // *The Tenth International Conference on Learning Representations*, ICLR 2022 (Virtual Event, 25–29 April 2022).– ICLR.– 2022.– 979-8-3313-0009-8, 2106.09685 .

[15] Dettmers T., Pagnoni A., Holtzman A., Zettlemoyer L. *QLoRA: Efficient finetuning of quantized LLMs* // *NIPS'23: Proceedings of the 37th International Conference on Neural Information Processing Systems*, NeurIPS 2023 (New Orleans, LA, USA, 10–16 December 2023), Advances in Neural Information Processing Systems.– vol. **36**.– Curran Associates, Inc..– 2023.– 9781713899921.– Pp. 10088–10115.– id. 441, 2305.14314 .

Приложение. Примеры промптов

Промпты приведены в сокращённом виде; полные формулировки опубликованы вместе с датасетом.

Классификация (zero-shot). «Ниже дано сочетание глагола и существительного. Определи, является ли оно коллокацией. Ответь только `1` (коллокация) или `0` (не коллокация).»

Перефразирование (русскоязычные модели и Llama-3.1-70B). «Ниже дано сочетание глагола и существительного. Замени всё сочетание на один глагол с таким же значением. В ответе напиши ТОЛЬКО глагол.» Для мультиязычных моделей среднего размера добавлялась строка «Твой ответ должен быть ТОЛЬКО на русском языке».

Перефразирование (DeepSeek-llm-7B-base). Инструктивный промпт не сработал: модель повторяла примеры, поэтому промпт различался по режимам. В режиме zero-shot использовалось продолжение с принудительным началом: «Вопрос: Как одним глаголом заменить выражение ‘{col}’? Ответ: Выражение ‘{col}’ можно заменить глаголом ‘», где завершающая открывающая кавычка «подсказывает» глагол. В режимах one-shot и few-shot применялся список пар-образцов:

> Замени сочетание глагола и существительного одним глаголом. Примеры: *вскрывать причины* = *выяснять* … {col} =

При параметрах декодирования

```
do_sample=False, num_beams=3, repetition_penalty=1.5, max_new_tokens=7
```

это почти всегда давало однословный ответ-глагол.

Заполнение пропуска (расширенный промпт для мультиязычных моделей среднего размера). «Ты — специалист по русским устойчивым выражениям. Заполни ⟨MASK⟩ УСТОЙЧИВЫМ СОЧЕТАНИЕМ из ГЛАГОЛА и СУЩЕСТВИТЕЛЬНОГО (возможен предлог). ПРАВИЛЬНО: *разбудить интерес*. НЕПРАВИЛЬНО: *пожертвовать* (только глагол). Сочетание должно быть естественным для русского языка и подходить в предложении. Отвечай ТОЛЬКО на русском.»

Аугментация отрицательных примеров (GPT-4o). «Коллокация — это устойчивое сочетание двух слов, значение которого некомпозиционально и которое можно заменить одним глаголом. Коллокации: *вести дискуссию, взваливать бремя, впадать в истерику* (= *дискутировать, обременять, истерить*). Неколлокации композициональны (*готовить обед, организовывать мероприятие*) и не заменяются одним глаголом. Напиши 100 сочетаний глагола с существительным, которые *не* являются коллокациями.» Ответы фильтровались вручную.

Порождение предложений с пропуском (GPT-4o). «Для каждой коллокации придумай одно русское предложение средней длины, естественное и грамматичное, с достаточным контекстом, чтобы коллокация однозначно восстанавливалась. Ставь глагол в правильной видовой форме. Верни два столбца: предложение с ⟨MASK⟩ на месте коллокации и то же предложение с восстановленной коллокацией.»

Судья для перефразирования (шкала 0–6). «Оцени семантическую и грамматическую близость двух русских глаголов. *Семантическая* близость (0–3): 3 – полностью совпадают; 2 – близкие синонимы, взаимозаменяемые в одном предложении (*взглянуть/посмотреть*); 1 – общее значение с заметными отличиями (*защищать/ценить*); 0 – не связаны или не на русском. *Грамматическая* близость (0–3): 3 – формы совпадают; 2 – незначительные отличия (вид, возвратность); 1 – существенные отличия; 0 – не глагол или не на русском. Верни строгий JSON с двумя оценками и их суммой.»

Судья для заполнения пропуска (шкала 0–4). «Оцени близость двух фрагментов (ответ и эталон). 4 – совпадают; 3 – очень близки по смыслу, одинаковой длины, взаимозаменяемы (*принять на себя ответственность / взять вину на себя*); 2 – те же слова в разных формах (*бить тревогу / забить тревогу*); 1 – отдалённо связаны (*держать учёт / беречь деньги*); 0 – не связаны, ИЛИ одно слово, ИЛИ слишком длинно, ИЛИ содержит нерусские вставки. Верни только целое число.»

Поступила в редакцию 11.10.2026;
одобрена после рецензирования 13.10.2026;
принята к публикации 17.10.2026;
опубликована онлайн 06.11.2026.

Рекомендовал к публикации *д.ф.м.-н. Н. В. Лукашевич*

Информация об авторе:



Вера Павловна Краснобаева

магистрант программы «Искусственный интеллект» AI Talent Hub Университета ИТМО, NLP-инженер и исследователь. Научные интересы: обработка естественного языка, большие языковые модели, некомпозициональная лексика, параметрически эффективное дообучение, безопасность ИИ

0009-0000-3109-9101

e-mail: krasnobaevavera2022@gmail.com

Декларация об отсутствии личной заинтересованности: *благополучие автора не зависит от результатов исследования.*