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.

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. https://psta.psiras.ru/read/psta2026_5_3-42.pdf

© Краснобаева В. П. 2026

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* (→ *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.

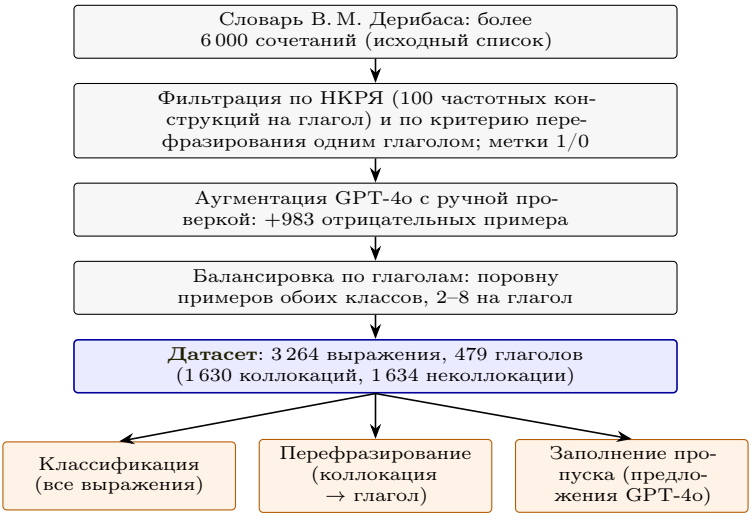


Рисунок 1. Схема построения датасета и три подзадачи над одними и теми же выражениями

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*ъ для эталона *обвинить*ъ 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.

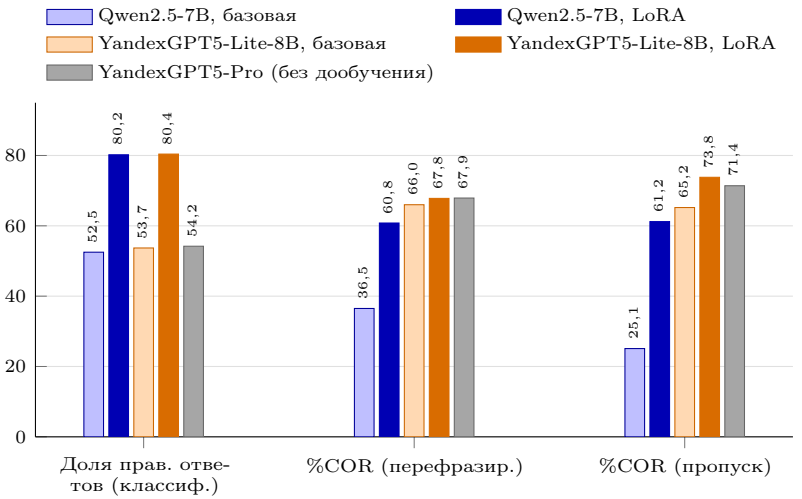


Рисунок 2. Качество двух моделей среднего размера до и после дообучения методом LoRA в сравнении с недообученной YandexGPT5-Pro (лучший режим для базовых моделей, %)

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*, Proceedings of the International Conference “Dialogue 2026” (г. Москва, Россия, 24–26 июня 2026 г.), Computational Linguistics and Intellectual Technologies.– 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.*