

## КОЛИЧЕСТВЕННЫЙ АНАЛИЗ СПОСОБНОСТИ МНОГОЯЗЫЧНЫХ ЯЗЫКОВЫХ МОДЕЛЕЙ СЛЕДОВАТЬ ИНСТРУКЦИЯМ НА РУССКОМ

И.Д. Мартынов<sup>1</sup>, И.А. Килбас<sup>1</sup>, Р.А. Парингер<sup>1</sup>

<sup>1</sup>Самарский университет, г. Самара  
e-mail: ilyamartinov2004@gmail.com

### Аннотация

В работе выполнен количественный анализ способности многоязычных языковых моделей следовать инструкциям на русском языке на бенчмарке ruFEval. Сравниваются модели llama-3.2-3b-instruct-uncensored, qwen2.5-3b-instruct и gemma-3-4b-it в формате Q8\_0. Оценка проведена по метрике accuracy в режимах strict и loose на уровнях prompt-level и instruction-level, а также по категориальным показателям ruFEval.

**Ключевые слова:** многоязычные языковые модели, следование инструкциям, ruFEVAL, accuracy, русский язык, instruction-following.

### Введение

Способность больших языковых моделей корректно следовать инструкциям пользователя является одной из ключевых характеристик, определяющих их практическую применимость в прикладных сценариях генерации текста, автоматизации пользовательских запросов и решения формализованных задач. Для таких систем недостаточно демонстрировать высокую общую связность или правдоподобие ответа: в ряде случаев принципиально важно, чтобы модель соблюдала явно заданные ограничения, связанные с языком ответа, структурой текста, длиной, использованием определённых слов, форматом вывода и сочетанием нескольких условий одновременно. В связи с этим задача оценки instruction-following в последние годы стала самостоятельным направлением анализа качества языковых моделей [1].

Целью работы является количественный анализ способности многоязычных языковых моделей следовать инструкциям на русском языке на основе результатов, полученных на бенчмарке ruFEval [2]. Для достижения этой цели решаются следующие задачи: провести экспериментальное тестирование выбранных моделей на полном наборе ruFEval; сравнить модели по значениям strict и loose accuracy на уровнях prompt-level и instruction-level [3]; проанализировать показатели по отдельным категориям инструкций; выявить наиболее существенные различия между моделями и интерпретировать их с точки зрения сильных и слабых сторон instruction-following.

### 1. Теоретические основы оценки способности языковых моделей следовать инструкциям

Способность языковой модели следовать инструкциям является одной из ключевых характеристик её практической применимости. В отличие от свободной генерации текста, в задачах instruction-following модель должна не только сформировать

связный ответ, но и выполнить заданные пользователем ограничения, связанные с языком, структурой, длиной, форматом и содержанием вывода.

Для русскоязычного сценария эта задача особенно актуальна, поскольку даже многоязычные instruction-tuned модели могут существенно различаться по качеству выполнения формализованных инструкций на русском языке [4]. Это связано с особенностями обучающих данных, instruction-tuning и неоднородностью поддержки разных языков.

В настоящей работе для такой оценки используется бенчмарк ruFEval, позволяющий сравнивать модели по метрике accuracy в двух режимах – strict и loose, а также на двух уровнях – prompt-level и instruction-level. Дополнительно бенчмарк предоставляет результаты по категориям инструкций, что делает возможным более содержательный анализ причин различий между моделями.

## 2. Постановка эксперимента

В экспериментальной части работы оценивалась способность многоязычных языковых моделей следовать инструкциям на русском языке с использованием бенчмарка ruFEval. В качестве исследуемых моделей были выбраны llama-3.2-3b-instruct-uncensored, qwen2.5-3b-instruct и gemma-3-4b-it. Для обеспечения сопоставимости результатов все модели тестировались в одинаковых условиях. Основной метрикой качества являлась accuracy – доля задач, для которых модель выдала правильный итоговый ответ.

В работе рассматривались две схемы оценки: strict и loose. В режиме strict корректным считался только тот ответ, который полностью удовлетворял проверяемым требованиям. В режиме loose допускались ограниченные отклонения, не нарушающие общий смысл выполнения инструкции. Кроме того, результаты анализировались на двух уровнях: prompt-level, где успешным считался ответ, удовлетворяющий всем инструкциям запроса, и instruction-level, где оценивалась доля корректно выполненных отдельных инструкций.

## 3. Результаты экспериментов

Основные результаты сравнения моделей по метрике accuracy приведены в таблице 1. Как видно из таблицы 1, наилучшие значения по всем основным метрикам показала модель gemma-3-4b-it, второе место заняла qwen2.5-3b-instruct, а наименьшие значения были получены для llama-3.2-3b-instruct-uncensored. При этом одинаковый порядок моделей сохраняется как для strict, так и для loose оценки, что указывает на устойчивость полученного ранжирования.

Таблица 1 – Основные результаты оценки моделей на бенчмарке ruFEval

| Модель                           | Strict prompt-level | Strict instruction-level | Loose prompt-level | Loose instruction-level |
|----------------------------------|---------------------|--------------------------|--------------------|-------------------------|
| gemma-3-4b-it                    | 0,5823              | 0,6930                   | 0,6248             | 0,7254                  |
| qwen2.5-3b-instruct              | 0,5046              | 0,6019                   | 0,5453             | 0,6367                  |
| llama-3.2-3b-instruct-uncensored | 0,4418              | 0,5396                   | 0,4621             | 0,5635                  |

Как видно из таблицы 1, для всех моделей значения instruction-level accuracy выше, чем prompt-level accuracy. Это объясняется тем, что выполнение отдельной инструкции является менее сложной задачей, чем одновременное соблюдение всех ограничений в одном ответе. Кроме того, значения loose accuracy во всех случаях превышают strict accuracy, что указывает на наличие частично корректных ответов, в которых модель в целом следует инструкции, но допускает отдельные формальные отклонения. Наименьший разрыв между strict и loose наблюдается у llama-3.2-3b-instruct-uncensored, однако это связано не с большей точностью, а с более низким общим качеством следования инструкциям.

Для интерпретации причин различий между моделями были рассмотрены отдельные показатели, вычисляемые бенчмарком ruFEval. Наиболее показательные категории приведены в таблице 2.

Таблица 2 – Средняя accuracy по стратегиям для command-a-reasoning-08–2025

| Категория          | gemma-3-4b-it | qwen2,5-3b-instruct | llama-3,2-3b-instruct-uncensored |
|--------------------|---------------|---------------------|----------------------------------|
| language           | 1             | 1                   | 1                                |
| keywords           | 0,7914        | 0,669               | 0,5252                           |
| detectable_format  | 0,8153        | 0,7389              | 0,6115                           |
| length_constraints | 0,6923        | 0,5944              | 0,5035                           |
| combination        | 0,5846        | 0,3385              | 0,1538                           |

Как видно из таблицы 2, все модели безошибочно справляются с выбором языка ответа, поэтому различия в итоговой accuracy определяются выполнением более сложных ограничений. Наиболее высокие результаты gemma-3-4b-it в категориях keywords, detectable\_format, length\_constraints и особенно combination свидетельствуют о её лучшей устойчивости к лексическим, структурным, количественным и составным инструкциям. qwen2.5-3b-instruct занимает промежуточное положение, тогда как llama-3.2-3b-instruct-uncensored показывает наименьшие значения, прежде всего в категории combination, что указывает на трудности при одновременном соблюдении нескольких условий. Таким образом, результаты, представленные в табл. 1 и табл. 2, показывают, что содержательное сравнение моделей требует анализа не только интегральной accuracy, но и категориальных показателей ruFEval, позволяющих выявить сильные и слабые стороны instruction-following.

## Заключение

Результаты исследования показывают, что качество следования русскоязычным инструкциям определяется не только интегральными значениями accuracy, но и профилем выполнения отдельных типов ограничений. Использование бенчмарка ruFEval позволило выявить различия между моделями по уровням strict и loose оценки, а также по категориальным показателям. Наибольшие затруднения у исследуемых моделей вызывают составные инструкции, требующие одновременного соблюдения нескольких условий.

Проведённое сравнение показало устойчивое ранжирование моделей по всем основным метрикам: наилучшие результаты продемонстрировала модель gemma-3-4b-it,

за ней следует qwen2.5-3b-instruct, тогда как llama-3.2-3b-instruct-uncensored показала наименьшие значения accuracy. Преимущество gemma-3-4b-it проявляется как на уровне отдельных инструкций, так и на уровне целых запросов, что свидетельствует о её более высокой способности к комплексному выполнению требований. В то же время разрыв между instruction-level и prompt-level метриками для всех моделей подтверждает, что основная сложность заключается не в интерпретации отдельных ограничений, а в их одновременной реализации в одном ответе.

Таким образом, полученные результаты показывают, что для адекватной оценки качества instruction-following необходимо учитывать не только обобщённые метрики, но и структуру ошибок, связанную с типами ограничений. Практическая значимость работы заключается в выявлении сильных и слабых сторон конкретных моделей при работе с русскоязычными инструкциями, что может быть использовано при выборе модели для прикладных задач, а также при дальнейшем совершенствовании методов instruction-tuning.

#### **Список использованных источников**

1. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models / J. Wei, X. Wang, D. Schuurmans [et al.] // arXiv. – 2022. – URL: <https://arxiv.org/abs/2201.11903> (дата обращения: 24.03.2026).
2. ruFEval: Instruction Following Eval in Russian. – URL: <https://github.com/NLP-Core-Team/ruFEval> (дата обращения: 29.03.2026).
3. Zhou, J. Instruction-Following Evaluation for Large Language Models / J. Zhou, T. Lu, S. Mishra [et al.] // arXiv. – 2023. – URL: <https://arxiv.org/abs/2311.07911> (дата обращения: 29.03.2026).
4. Ouyang, L. Training language models to follow instructions with human feedback / L. Ouyang, J. Wu, X. Jiang [et al.] // arXiv. – 2022. – URL: <https://arxiv.org/abs/2203.02155> (дата обращения: 29.03.2026).

## **QUANTITATIVE ANALYSIS OF THE ABILITY OF MULTILINGUAL LANGUAGE MODELS TO FOLLOW INSTRUCTIONS IN RUSSIAN**

**I.D. Martynov<sup>1</sup>, I.A. Kilbas<sup>1</sup>, R.A. Paringer<sup>1</sup>**

*<sup>1</sup>Samara University, Samara  
e-mail: ilyamartinov2004@gmail.com*

#### **Abstract**

This paper presents a quantitative analysis of the ability of multilingual language models to follow instructions in Russian using the ruFEval benchmark. The models compared are llama-3.2-3b-instruct-uncensored, qwen2.5-3b-instruct, and gemma-3-4b-it in the Q8\_0 format. The evaluation is based on the accuracy metric in strict and loose modes at the prompt-level and instruction-level, as well as on the categorical indicators provided by ruFEval.

**Keywords:** multilingual language models, instruction following, ruFEval, accuracy, Russian language, instruction-following.