МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

РАЗРАБОТКА И ВАЛИДАЦИЯ БЕНЧМАРКА NOLIMA-RU ДЛЯ ОЦЕНКИ ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ ИЗ КОНТЕКСТА РУССКОЯЗЫЧНЫМИ БОЛЬШИМИ ЯЗЫКОВЫМИ МОДЕЛЯМИ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 412 группы
направления 01.03.02 — Прикладная математика и информатика

механико-математического факультета
Пицунова Тимура Андреевича

Научный руководитель
к. ф.-м. н., доцент _____ Л. В. Борисова

Заведующий кафедрой
д. ф.-м. н., доцент _____ С. П. Сидоров

Саратов 2026

Актуальность данной работы обусловлена растущей потребностью в надёжных методах оценки способности больших языковых моделей (LLM) обрабатывать длинные контексты. Современные модели декларируют контекстные окна вплоть до миллиона токенов, однако формальное расширение контекста не гарантирует, что модель действительно способна извлекать и использовать информацию из всего доступного объёма. Существующие бенчмарки, такие как Needle In A Haystack (NIAH), основаны на задачах буквального сопоставления и в значительной степени допускают решение путём лексического сопоставления, не требуя от модели семантического рассуждения. Бенчмарк NoLiMa (Adobe Research, 2025) устраняет этот недостаток, требуя извлечения через латентные ассоциации без лексического перекрытия, но существует только для английского языка. Настоящая работа посвящена созданию такой адаптации.

Объект исследования — методология оценки способности языковых моделей к семантическому извлечению информации из длинных контекстов.

Предмет исследования — адаптация бенчмарка NoLiMa для русского языка: разработка оригинальных шаблонов, построение русскоязычного корпуса, валидация нулевого лексического перекрытия.

Целью данной работы является разработка русскоязычного бенчмарка NoLiMa-ru — полного аналога оригинального NoLiMa, адаптированного для русского языка с учётом культурной специфики и морфологических особенностей.

Для достижения поставленной цели необходимо решить следующие задачи:

- Провести анализ оригинального бенчмарка NoLiMa и определить ограничения, препятствующие прямому переносу на русский язык.
- Разработать оригинальные шаблоны на основе культурно-географически знаний стран СНГ с валидацией нулевого лексического перекрытия.
- Построить русскоязычный корпус для фонового контекста и набор

персонажей, удовлетворяющие требованиям бенчмарка.

— Реализовать программный комплекс для проведения оценки и анализа результатов.
— Провести экспериментальную оценку на модели GigaChat 3.1.

Работа состоит из двух основных разделов. В первом разделе рассматривается теоретическая база: принципы работы больших языковых моделей, существующие бенчмарки и формальная постановка задачи NoLiMa. Во втором разделе описывается практическая адаптация для русского языка и результаты экспериментальной оценки модели GigaChat 3.1 на разработанном бенчмарке.

Большая языковая модель (LLM) — нейронная сеть, обученная на корпусе текстов задаче авторегрессионного предсказания следующего токена. Модель строит условное распределение $P(t_{n+1} \mid t_1, \ldots, t_n; \theta)$, обучение заключается в минимизации перекрёстной энтропии. Архитектурной основой современных LLM является Transformer [1], ключевой компонент которого — механизм самовнимания:

$$\text{Attention}(\mathbb{Q}, \mathbb{K}, \mathbb{V}) = \text{softmax}\left(\frac{\mathbb{Q}\mathbb{K}^T}{\sqrt{d_k}}\right)\mathbb{V} \tag{1}$$

Вычислительная сложность полного внимания составляет $\mathcal{O}(n^2 \cdot d)$, что делает обработку длинных контекстов ресурсоёмкой и мотивирует разработку бенчмарков для оценки реальной способности моделей использовать длинный контекст. Максимальная длина последовательности $L_{\max}$, которую модель может обработать за один проход, называется *контекстным окном*. Для современных моделей $L_{\max}$ составляет от 32 000 до 1 000 000 токенов, однако формальное наличие длинного контекстного окна не гарантирует, что модель эффективно использует информацию из всех его частей.

Бенчмарк NIAH [2] внедряет целевой факт в длинный фоновый контекст и тестирует различные положения целевого факта и длины контекста. Однако формулировка вопроса прямо повторяет целевой фрагмент, сводя задачу к поиску по ключевым словам. LongBench [3] включает 21 задачу на шести языках, однако не специализируется на

семантическом извлечении и не контролирует лексическое перекрытие. Исследование «Lost in the Middle» [4] показало, что модели склонны игнорировать информацию из центральной части длинного документа. Это явление особенно выражено в задачах, требующих понимания, а не простого поиска.

Бенчмарк NoLiMa (Non-Literal Matching) [5] предложен Adobe Research в 2025 г. Ключевая идея: между формулировкой вопроса и целевым фактом не должно быть лексического перекрытия. Модель не может решить задачу поиском по ключевым словам и должна понимать семантику вопроса, идентифицировать релевантный фрагмент через ассоциативные связи и извлекать ответ через промежуточные знания. Для каждого шаблона проверяется условие нулевого перекрытия на множестве лемматизированных токенов:

$$\text{Lem}(n) \cap \text{Lem}(q) \setminus \text{StopWords} = \varnothing \tag{2}$$

Основная метрика — точность извлечения для заданной длины контекста L и глубины внедрения d:

$$\text{Acc}(L, d) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[\text{match}(\hat{a}_i, \chi_i)] \tag{3}$$

Дополнительные метрики: скорректированный Base Score (BS_{corr}), скорректированная эффективная длина (EL_{corr}) и доверительный интервал Вильсона. Оригинальный бенчмарк показал, что модели, демонстрирующие точность $> 0{,}9$ на NIAH/Passkey, дают точность $< 0{,}5$ на NoLiMa при той же длине контекста. Эффект «потерянной середины» значительно сильнее в задачах семантического извлечения.

Шаблоны оригинального NoLiMa укоренены в англо-американской культуре. Прямой перевод создаёт необоснованное усложнение: модель должна и понять чужую культурную отсылку, и выполнить извлечение. Специфика русского языка создаёт дополнительные требования:

1. Морфология. Русский язык имеет 6 падежей, до 12+ форм на лемму. Валидацию перекрытия проводят на уровне лемм, а не словоформ.
2. Синонимия. Одно понятие может выражаться множеством лексически различных, но семантически эквивалентных способов.
3. Свободный порядок слов. Шаблоны параметризуются порядком формулировки (default и inverted).
4. География и культура СНГ. Необходимы оригинальные ассоциации: Большой театр $\rightarrow$ Москва, Оймякон $\rightarrow$ Якутия, башня Исеть $\rightarrow$ Екатеринбург.

Таким образом, локализация NoLiMa для русского языка представляет собой не перевод, а создание оригинального бенчмарка по методологии оригинала.

Для русскоязычной адаптации разработано 58 шаблонов в 5 группах: География (20: 5 ориентиров $\times$ 2 hop $\times$ 2 порядка), Еда (12), Здоровье (8), Искусство и культура (10), Здравый смысл (8). Каждый шаблон параметризуется глубиной ассоциативного вывода (1-hop — прямая ассоциация, 2-hop — цепочка из двух шагов) и порядком формулировки (default, inverted). Все шаблоны проверены на нулевое лексическое перекрытие: средний ROUGE-1 составил $0{,}031$, ROUGE-2 — $0{,}000$, что ниже порогов $0{,}15$ и $0{,}3$ соответственно.

Для формирования русскоязычного контекста использован датасет BooksSummarizationRU с двухэтапной фильтрацией (E5 + LLM). Сформировано 5 независимых фоновых контекстов. Набор из 30 дореволюционных имён минимизирует артефакты обучающих данных. Программный комплекс включает модули подготовки данных (`build_datasets.py`), оценки (`evaluate.py`) и анализа (`gather_results.py`). Конфигурация оценки: до 26 интервалов глубины внедрения (0%–100%), до 5 фоновых контекстов, greedy-декодирование, Wilson CI.

Модель GigaChat 3.1 [6] (MoE, 10 млрд параметров, 1,8 млрд активных), развёрнутая через vLLM [7] на NVIDIA RTX PRO 6000 S, протестирована на 9 длинах контекста (250–64 000 токенов). Точность по

длине контекста. Точность монотонно снижается с 80,4% при 250 токенах до 12,2% при 64 000 токенах. Скорректированный Base Score — 78,0%, эффективная длина — лишь 1 000 токенов при декларируемом контекстном окне 262 144. Влияние глубины вывода. Разница между 1-hop и 2-hop шаблонами составляет 15,0 п. п. в среднем (53,4% vs 38,4%). При 64 000 токенах точность 2-hop — 3,3% (практически случайное угадывание). Потерянная середина. При 64 000 токенах точность в середине контекста — 0,6%, в начале — 30,5%, в конце — 43,7%. Перепад между серединой и концом составляет 43,1 п. п.

Таким образом, в ходе выполнения работы был разработан русскоязычный бенчмарк NoLiMa-ru — полная адаптация оригинального бенчмарка NoLiMa (Adobe Research, 2025) для русского языка. Показано, что прямой перевод шаблонов NoLiMa неприемлем: шаблоны укоренены в англо-американской культуре, а специфика русского языка (морфология, синонимия, свободный порядок слов) требует оригинальных решений. Локализация NoLiMa — не перевод, а создание нового бенчмарка по методологии оригинала. Разработано 58 шаблонов в 5 группах (География, Еда, Здоровье, Искусство и культура, Здравый смысл), расширяющих оригинальные две группы. Каждый шаблон параметризуется глубиной вывода (1-hop, 2-hop) и порядком (default, inverted). Все шаблоны проверены на отсутствие лексического пересечения между целевым фактом и вопросом на уровне лемм, что исключает возможность решения задачи текстовым поиском. Построен русскоязычный фоновый контекст на основе BooksSummarizationRU с двухэтапной фильтрацией (E5 + LLM), сформировано 5 независимых фоновых контекстов. Набор из 30 дореволюционных имён предположительно снижает вероятность артефактов, связанных с априорными знаниями модели. Программный комплекс обеспечивает воспроизводимость: 10 интервалов глубины, 3 фоновых контекста, Wilson CI, скорректированные BS/EL. Экспериментальная оценка модели GigaChat 3.1 (10 млрд параметров, MoE) показала, что точность семантического извлечения монотонно снижается с 80,4% при 250 токенах до 12,2% при 64 000 токенах. Эффективная длина — лишь 1 000 токенов при декларируемом кон-

текстном окне $262\,144$. Задачи с двухшаговым выводом дают точность на $15{,}0$ п. п. ниже одношаговых, а эффект «потерянной середины» усиливается с ростом контекста.

Направления будущих исследований: расширение набора шаблонов с участием лингвистов, адаптация NoLiMa для других языков (китайский, арабский, хинди), валидация бенчмарка на моделях другого масштаба и архитектуры (YandexGPT, GPT-4o, Claude) для подтверждения воспроизводимости выявленных эффектов.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Vaswani, A. Attention Is All You Need [Электронный ресурс] / A. Vaswani, N. Shazeer, N. Parmar [и др.] // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – URL: https://arxiv.org/abs/1706.03762 (дата обращения: 20.05.2026). – Загл. с экрана. – Яз. англ.

2 Needle In A Haystack — Pressure Testing LLMs [Электронный ресурс] / G. Kamradt // GitHub : [сайт]. – URL: https://github.com/gkamradt/LLMTest_NeedleInAHaystack (дата обращения: 15.01.2025). – Загл. с экрана. – Яз. англ.

3 Bai, Y. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding [Электронный ресурс] / Y. Bai, S. Lv, J. Peng [и др.] // arXiv preprint arXiv:2308.14508. – 2023. – URL: https://arxiv.org/abs/2308.14508 (дата обращения: 20.05.2026). – Загл. с экрана. – Яз. англ.

4 Liu, N. F. Lost in the Middle: How Language Models Use Long Contexts [Электронный ресурс] / N. F. Liu, K. Lin, J. Hewitt [и др.] // Transactions of the Association for Computational Linguistics. – 2023. – Vol. 12. – P. 157–173. – URL: https://arxiv.org/abs/2307.03172 (дата обращения: 20.05.2026). – Загл. с экрана. – Яз. англ.

5 Modarressi, A. NoLiMa: Long Context Evaluation Beyond Literal Matching [Электронный ресурс] / A. Modarressi, J. Lee, S. Kim [и др.] // arXiv preprint arXiv:2502.05167. – 2025. – URL: https://arxiv.org/abs/2502.05167 (дата обращения: 20.05.2026). – Загл. с экрана. – Яз. англ.

6 GigaChat 3.1: Large Language Model [Электронный ресурс] // Hugging Face : [сайт]. – URL: https://huggingface.co/ai-sage/GigaChat3.1-10B-A1.8B-bf16 (дата обращения: 20.05.2025). – Загл. с экрана. – Яз. англ.

7 Kwon, W. Efficient Memory Management for Large Language Model Serving with PagedAttention [Электронный ресурс] / W. Kwon,

Z. Li, S. Zhuang [и др.] // arXiv preprint arXiv:2309.06180. – 2023. – URL: https://arxiv.org/abs/2309.06180 (дата обращения: 20.05.2025). – Загл. с экрана. – Яз. англ.