

Правительство Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «МОСКОВСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ М.В.ЛОМОНОСОВА»
(МГУ)

Научно-исследовательский вычислительный центр

УДК
004.9
Рег. № НИОКТР
Рег. № ИКРБС

УТВЕРЖДАЮ
Директор
Научно-исследовательский
вычислительный центр,
д.ф.-м.н., проф., член-корр. РАН

_____ В.В. Воеводин

« ____ » _____ Г.

ОТЧЕТ
О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

по теме:

Методы адаптации больших языковых моделей на русский язык и
конкретные предметные области
(промежуточный)

Руководитель НИР:

старший
научный сотрудник, кандидат
физико-математических наук,



Тихомиров М.М.

СПИСОК ИСПОЛНИТЕЛЕЙ

Руководитель НИР,
старший науч-
ный сотрудник,
кандидат физико-
математических наук



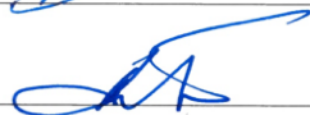
Тихомиров М.М.
(все разделы)

Исполнители:



Григорьев Д.А.
(все разделы)

заведующий ла-
бораторией, кан-
дидат физико-
математических наук



Добров Б.В.
(все разделы)

стажёр-
исследователь



Ковалев Г.П.
(все разделы)

специалист



Рожков И.С.
(все разделы)

стажер



Чернышев Д.И.
(все разделы)

научный сотрудник,
кандидат филологи-
ческих наук



Ярошенко П.В.
(все разделы)

РЕФЕРАТ

Ключевые слова:

русский язык, глубокое обучение, машинное обучение, большие языковые модели, искусственный интеллект

Ключевые слова по-английски:

russian language, deep learning, artificial intelligence, large language models, machine learning

В данном проекте рассматривается комплексная проблема адаптации больших языковых моделей на русский язык. Решение подобной задачи требует разработки стабильного метода по смене токенизации у уже существующих больших языковых моделей, так как в большинстве случаев, исходная токенизация мультязычных LLM имеет низкую эффективность представления текстов на русском языке. Не менее важной связанной задачей является разработка бенчмарка по оценке качества работы языковых моделей на русском языке с точки зрения навыков понимания, анализа и др. так как большинство современных бенчмарков сконцентрированы вокруг английского языка, либо же недостаточно полно рассматривают возможности моделей применительно к русскому языку. Разработка подобных открытых бенчмарков (то есть, которые не требуют отправки результатов на какие-либо платформы), является необходимым для эффективного развития методов адаптации и обучения языковых моделей на русском языке.

В первый год выполнения проекта были получены следующие ключевые результаты:

1. Разработана улучшенная версия методологии адаптации - Ruadapt.

Предложен и верифицирован комплексный подход к адаптации моделей (LLM) на русский язык, включающий расширение словаря для лучшего учета морфологии русского языка и многоэтапное обучение. Ключевой особенностью методологии является использование стратегии двухэтапной адаптации, что позволяет модели эффективно усваивать русский язык без «катастрофического забывания» исходных знаний, а затем метода проекции эмбедингов (LER - Learned Embedding Propagation), идея которого в том, чтобы спроецировать адаптированные эмбединги базовой версии (которая не умеет вести диалог) в инструктивное пространство. На основе разработанной методологии адаптированы и выложены в открытый доступ модели серии RuadaptQwen3 (размером 4B, 8B и 32B параметров).

2. Созданы новые инструменты оценки и наборы данных (бенчмарки).

Для всесторонней проверки качества моделей были разработаны уникальные для русского языка ресурсы:

RuWikiBench: бенчмарк для проверки аналитических способностей моделей в рамках задачи составления энциклопедических статей по стандартам

«Рувикис». Показано, что при меньших вычислительных затратах адаптированные модели способны воспроизводить статьи на уровне лучших открытых аналогов.

RuBookSum: набор данных для реферирования длинных художественных текстов со сложными нарративными структурами. Для эффективного решения задачи посредством больших языковых моделей разработан вычислительно оптимизированный вариант алгоритма иерархического реферирования.

Оценка эмоционального интеллекта: проведено сравнение восприятия эмоциональной окраски слов моделями и людьми. Выявлено, что базовые модели склонны гиперболизировать негативные эмоции, в то время как процедура адаптации на русский язык приближает оценки моделей к человеческому восприятию.

Синтаксический анализ: исследованы возможности моделей в построении деревьев зависимостей. Установлено, что LLM пока уступают специализированным парсерам в задачах строгого структурного анализа.

3. Разработан новый метод сжатия моделей.

Предложен алгоритм итерационного пруннинга (удаления слоев) с последующей дистилляцией знаний. Метод позволяет уменьшить размер модели и ускорить ее работу минимизируя потерю качества (менее 10% потерь при сокращении числа параметров с 3 до 2.47 млрд), превосходя существующие статические подходы.

Были подготовлены рукописи и опубликованы и приняты в печать 6 научных статей:

Тихомиров М. М., Чернышев Д. И. Ruadapt: вычислительно эффективная языковая адаптация больших языковых моделей // Доклады Российской академии наук. Математика, информатика, процессы управления. – 2025. – Т.

527. – С. 290-300. (WoS, Scopus)

Григорьев Д. А., Чернышев Д. И. RuWikiBench: оценка больших языковых моделей посредством воспроизведения энциклопедических статей // Доклады Российской академии наук. Математика, информатика, процессы управления. – 2025. – Т. 527. – С. 171-181. (WoS, Scopus)

Iaroshenko P. V., Loukachevitch N. V. Large Language Models versus Native Speakers in Emotional Assessment of Russian Words // Supercomputing Frontiers and Innovations. – 2025. – Т. 12. – №. 3. – С. 20–30. (Scopus)

Shamaeva E. D., Tikhomirov M. M., Loukachevitch N. V. One-Shot Prompting for Russian Dependency Parsing // Supercomputing Frontiers and Innovations. – 2025. – Т. 12. – №. 3. – С. 108–120. (Scopus)

Grigoriev D. A., Khudiakov D. V., Chernyshev D. I. RuBookSum: Dataset for Russian Literature Abstractive Summarization // Supercomputing Frontiers and Innovations. – 2025. – Т. 12. – №. 3. – С. 76–89-76–89. (Scopus)

Grigory Kovalev, Mikhail Tikhomirov. Iterative Layer-wise Distillation for Efficient Compression of Large Language Models // Pattern Recognition and Image Analysis 2025 (Принята к публикации, Scopus)

ВВЕДЕНИЕ

Проект направлен на исследование методов адаптации и переноса знаний в больших языковых моделях с целью повышения вычислительной эффективности в условиях ограниченных вычислительных ресурсов, а также качества при работе на русском языке и конкретных предметных областях. На сегодняшний день существует значительная потребность в русскоязычных языковых моделях, которые бы соответствовали стандартам качества и могли эффективно использоваться в различных приложениях, таких как чат-боты, системы автоматического перевода, анализ текстов и другие. Адаптация существующих моделей на русский язык является компромиссом созданию таких моделей с нуля, позволяя существенно сэкономить на процессе обучения языковых моделей, при этом получая качество, не уступающее исходным версиям языковых моделей. Адаптированная под язык токенизация позволит использовать подобные модели с повышенной вычислительной эффективностью, а внедрение в модель новых знаний, в частности, ориентированных на актуальные для русскоязычного сообщества темы, позволит поднять уровень понимания русскоязычных текстов.

Существующие языковые модели часто рассматриваются как универсальные инструменты, но на практике возникают ситуации, когда их знания в специализированных областях недостаточны. Решение проблемы специализации моделей позволит повысить их точность и эффективность в решении задач в конкретных областях, таких как юриспруденция, медицина, финансы и другие.

Традиционные методы оценки качества языковых моделей часто не учитывают специфику русского языка и культурные аспекты, что приводит к недостаточной точности оценки с точки зрения применения LLM на русскоязычных текстах. Под культурными аспектами в рамках данного проекта подразумеваются специфические знания, актуальные для русскоязычного сообщества, явно выражаемые в текстах. Несмотря на то, что на данном этапе невозможно полностью охватить систему знаний о мире с точки зрения русскоязычного сообщества, будут предприняты первые попытки оценить связанные с этим направления свойства языковых моделей.

Разработка новых бенчмарков и методик оценки качества моделей позволит более точно оценивать их производительность и направлять дальнейшие исследования.

Таким образом, решение обозначенной научной проблемы имеет значительную научную и практическую значимость, что делает данный проект актуальным и необходимым для развития технологий больших языковых моделей.

ОСНОВНАЯ ЧАСТЬ

Первый год проекта был направлен на два ключевых направления: исследование методов повышения качества и эффективности адаптации LLM (больших языковых моделей) на русский язык, а также разработку инструментов для их всесторонней лингвистической оценки. Ниже представлены конкретные сведения о выполненных работах.

1. Исследование и разработка методик адаптации LLM на русский язык В основе проведенных исследований лежала задача качественного развития подходов к языковой адаптации, предварительно намеченных в ранних работах (Tikhomirov, Chernyshev, 2024). Предыдущая работа ограничивалась демонстрацией концепции переноса эмбедингов (LER - Learned Embedding Propagation), идея которой в том, чтобы спроецировать адаптированные эмбединги базовой версии в инструктивное пространство. Целью текущего этапа стала разработка и верификация комплексной методологии Ruadapt, применимой к современным моделям, а также анализ влияния гиперпараметров и различных компонент подхода.

Разработанный подход состоит из 4 основных этапов:

- 1) Построение морфологически-ориентированного словаря и расширение исходной токенизации. На данном этапе, применительно к русскому языку, а также другим языкам с богатой морфологией, рекомендуется построить новые токены алгоритмом Unigram и затем расширить исходный токенизатор новыми токенами целевого языка.
- 2) Адаптация базовой версии модели в два этапа - сначала адаптация только слоев эмбедингов, которые были инициализированы с использованием новой токенизации, а затем совместная адаптация внутренних слоев и эмбедингов с использованием метода LoRA (Hu E. J. et al. 2022). LoRA - это альтернатива полному дообучению, которая позволяет существенно сократить количество обучаемых параметров без потерь в качестве. Под адаптацией подразумевается подход “продолжения предобучения” на качественном наборе данных целевого языка (книги, Вики и др.).
- 3) Проекция эмбедингов (LER - Learned Embedding Propagation) из базовой адаптированной версии модели в ее инструктивную версию. В случае двухэтапного подхода к адаптации базовой версии, LoRA адаптеры накладываются на тело модели после процедуры проецирования эмбедингов и совмещения их с телом инструктивной версии модели.
- 4) Калибровка. Дообучение полученной “LER версии” модели на инструктивных данных из распределения исходной модели.

В работе был проведен анализ влияния компонентов методологии и гиперпараметров на итоговое качество:

- 1) Для решения проблемы баланса между усвоением новой лингвистической информации и сохранением исходных интеллектуальных способностей модели (с целью избежать феномена “катастрофической забывчивости”, из-за которого предыдущие знания модели переписываются новыми, что негативно влияет на итоговое качество) была экспериментально проверена стратегия двухэтапной адаптации базовой версии модели. Сначала производится обучение исключительно новых слоев эмбедингов (на базе нового словаря, построенного методом Unigram для лучшего учета русской морфологии), и лишь затем осуществляется корректировка внутренних слоев модели. Подобная схема впервые проверялась в комбинации с последующим применением метода LER.
- 2) Было показано, что важным является подбор скорости обучения, особенно на

первом этапе адаптации базовой версии модели, в частности в случае моделей серии Qwen3 рекомендуется использовать скорость обучения равную $5e-4$ (что выше, чем при обычных этапах дообучения моделей), а на втором шаге адаптации скорость обучения необходимо снизить до значений $5e-5$. С точки зрения гиперпараметров LoRA рекомендуется использовать ранг 128 и альфу 256. В целом, ранг LoRA адаптеров оказывает минорное влияние на итоговое качество, однако, важно подбирать альфу в соответствии с рангом и скоростью обучения (обычно $\times 2$).

- 3) Была показана переносимость подхода между разными размерами моделей, а также между различными инструктивными наборами данных, однако, рекомендуется для этапа калибровки использовать инструктивный набор из схожего с исходной моделью распределения.

Для оценки качества различных версий моделей использовался разрабатываемый в рамках выполнения проекта бенчмарк с разделением на разные категории задач: извлечение именованных сущностей, знания, суммаризация, перевод (ру-англ/англ-ру) и др. Полученные по разработанной методологии адаптированные модели RuadaptQwen3-4B, RuadaptQwen3-8B и RuadaptQwen3-32B были выложены в открытый доступ.

2. Разработка новых методик оценки и создание комплексных бенчмарков для русского языка

Значительный объем работ первого года был сосредоточен на создании новых инструментов и наборов данных для оценки языковых моделей на русском языке.

2.1. Исследование соответствия эмоционального восприятия слов большими языковыми моделями и носителями русского языка

Целью работы является оценка понимания эмоциональной составляющей русского языка большими языковыми моделями. Было проведено сопоставление того, каким образом носители русского языка оценивают русские слова с точки зрения их связи с той или иной эмоцией, с оценкой тех же слов с помощью больших языковых моделей. Для исследования была использована база данных ENRuN (Emotional Norms for Russian Nouns), которая включает эталонные эмоциональные оценки для 1800 русских существительных носителями русского языка. Эталонные оценки подразумевают разметку существительных, которым были проставлены оценки по шкале от 0 до 5 баллов по пяти категориям эмоций (радость, грусть, злость, страх, отвращение). Для выявления корреляции между “эмоциональной картиной мира моделей” и восприятием человека был проведен эксперимент, где слова из базы данных оценивались семью большими языковыми моделями, которые различаются по размеру (от 7 до 70 млрд параметров) и по происхождению (отечественные и иностранные). Затем результаты оценки моделей были сопоставлены с человеческой оценкой.

Анализ выявил тенденцию современных LLM к систематической гиперболизации негативных эмоций, особенно при обработке лексики, связанной с чувствительными темами (религия, этика, проявления физического насилия и др.), что, скорее всего, свидетельствует о влиянии этапов дообучения моделей безопасности (safety alignment).

Важным результатом исследования стало подтверждение позитивного влияния процедур адаптации моделей на русский язык. Было показано,

что адаптированные (или же исходно русскоязычные модели, такие как YandexGPT 5 Lite) показывают в среднем более высокую степень соответствия человеческим оценкам. В частности, адаптированная на русский язык модель RuadaptQwen2.5-32B-Pro-Beta показала наивысшее качество среди проверенных моделей, превзойдя исходную неадаптированную модель Qwen 2.5-32B.

2.2. Синтаксический анализ

С целью оценить понимание языковыми моделями структуры текстов на русском языке было проведено исследование возможностей языковых моделей к синтаксическому анализу. Данный тип тестирования позволяет определить, насколько глубоко модель понимает синтаксическую структуру русского предложения. Для этого от моделей требовалось генерировать полное синтаксическое дерево в формате CoNLL-U в режиме one-shot (по одному примеру), что является задачей высокого уровня сложности, так как современные большие языковые модели плохо справляются с задачами, когда необходимо детально проанализировать каждое слово текстовой последовательности.

На базе пяти корпусов данных (SynTagRus, GSD, PUD, Taiga, Poetry), было показано, что хотя LLM и способны выполнять поверхностный анализ, они все еще существенно уступают специализированным парсерам. Наибольшее влияние на итоговое качество оказывает в первую очередь размер моделей, однако при этом было замечено, что адаптация языковых моделей может негативно сказываться на качестве работы в подобной постановке, во многом, из-за несоответствия эталонных наборов слов и генерируемых наборов слов для последовательности. Данная проблема отдельно подсвечивает необходимость подобных наборов данных для оценки языковых моделей на русском языке, так как большинство классических задач не способны выявлять возможные проблемы со стабильностью генерации новых добавленных в модель токенов. Полученные результаты могут сигнализировать о том, что текущая методика адаптации должна быть улучшена с точки зрения повышения стабильности генерации новых токенов в различных контекстах.

2.3. Оценка качества понимания длинных текстов

Другим направлением стало исследование того, насколько модели хорошо понимают длинные тексты на русском языке. Одной из наиболее показательных задач для комплексной оценки навыков работы с длинными текстами является абстрактное реферирование, которое требует глубокого анализа текста на всех уровнях для выделения и обобщения наиболее важной информации. Поскольку для русского языка до сих пор не существовало открытых коллекций необходимого качества для оценки реферирования многостраничных документов был разработан новый набор данных для реферирования (summarization) длинных текстов художественной и научно-популярной литературы - RuBookSum. Важным аспектом набора данных является наличие сложных нарративных структур и нелинейных сюжетных линий, которые нельзя встретить в распространенных коллекциях реферирования коротких документов, таких как новостные или научные статьи, что предъявляет качественно более высокие требования к когнитивным способностям моделей по удержанию и анализу глобального контекста.

Кроме самого набора данных был предложен ряд модификаций наиболее популярного подхода к реферированию подобных текстов, иерархического,

когда текст сначала разбивается на сегменты, которые затем рекурсивно объединяются в группы по принципу дендрограммы для построения рефератов различных уровней детализации. Модификации касались как вопроса стабилизации работы метода посредством псевдо-планов реферата так и общей оптимизации вычислительных операций для снижения издержек и сокращения времени работы. Наиболее совершенная модифицированная версия иерархического алгоритма реферирования позволила в среднем ускорить генерацию реферата до 300% без существенных потерь в качестве.

Результаты оценки различных языковых моделей показали, что наилучшего качества достигают такие модели как DeepSeek V3 и Qwen3-235BA22B, однако средний разрыв в качестве с моделями меньших размеров (32 миллиарда параметров и 8 миллиардов параметров) не такой существенный, а в ряде случаев последние и вовсе демонстрируют конкурентоспособные результаты. Так, например, русскоязычная модель YandexGPT 5 Lite показала результаты сопоставимые с адаптированными моделями в 32 миллиарда параметров, что подчеркивает важность повышения степени знания русского языка в существующих моделях.

2.4. Оценка способностей к структурированию знаний

С целью диагностики аналитических способностей моделей при создании сложного по структуре и содержанию контента был разработан новый набор данных RuWikiBench – открытый бенчмарк на основе “Рувики” для оценки способностей больших языковых моделей воспроизводить логику построения статей в стиле Википедии. В основу бенчмарка легли три фундаментальных задачи построения энциклопедических статей: поиск релевантных источников по заданной тематике, планирование структуры статьи в соответствии с выделенными подтемами и обобщение источников в краткое и информативное содержание соответствующих секций плана итоговой статьи. Таким образом бенчмарк требует от моделей навыков понимания релевантности документов, умения структурировать внешние знания, а также анализировать и выделять нужные элементы знаний для формирования лаконичного представления (реферата).

Главный принцип бенчмарка - воспроизводимость и чистота оценки. Поэтому все этапы построения статьи по заданной тематике оцениваются по независимым эталонам, собранным напрямую из реально существующих статей Рувики: отобранные источники должны совпадать с реальным списком источников, сгенерированный план должен быть максимально близок по содержанию к экспертному, а текст статьи должен совпадать по содержанию с текстом настоящей статьи. При этом, каждый этап построения получает на вход эталонный результат работы предыдущего этапа, что позволяет исключить влияние сторонних факторов вроде низкую точность поиска релевантных документов или ошибки группировки источников по пунктам плана. При этом каждый этап учитывает несколько возможных стратегий решения задачи: на этапе поиска релевантных документов для получения поисковой выдачи рассматривается как сценарий генерации поискового запроса оцениваемой моделью так и использование заранее заготовленного (экспертного) запроса, а на этапе планирования оценивается как прямой подход генерации плана по тематически сгруппированным фрагментам источников так и вариант предварительного выделения ключевых моментов этих фрагментов.

Результаты экспериментов показали, что несмотря на хорошее знание ан-

гоязычной Википедии, современные LLM испытывают основные трудности с воспроизведением логики повествования статьи: даже при идеальном отборе источников лучшие открытые модели оказываются не способны полностью повторить работу эксперта Рувики. При этом с точки зрения качества генерации текста секций даже младшие модели довольно близко воспроизводят содержание. Так согласно метрикам ROUGE-L и BLEU, оценивающим лексическое и структурное сходство с эталоном, адаптированная модель RuadaptQwen3-32B показала наилучшие результаты, превосходящие даже модели DeepSeek V3 и Qwen3-235B-A22B.

3. Разработка и исследование эффективных методов сжатия моделей

В дополнение к планам на первый год выполнения проекта были начаты исследования по разработке эффективных методов сжатия языковых моделей.

В результате исследований был разработан метод по итерационному сжатию (пруннингу) LLM с вырезанием целых слоев, где между каждым шагом производится дополнительный шаг дообучения (дистилляции). В отличие от существующих подходов (таких как ShortGPT), которые оценивают важность слоев трансформера статически и единожды перед удалением, предложенный метод базируется на динамической переоценке значимости блоков на каждой итерации.

Для восстановления функциональности модели после структурного вмешательства был добавлен шаг дистилляции с использованием комбинированной функции потерь, объединяющая дивергенцию Кульбака-Лейблера (KLDivLoss) и среднеквадратичную ошибку (MSELoss) для синхронизации скрытых состояний модели-учителя и модели-ученика. Также было предложено рассчитывать значимость блоков не на базе опосредованной метрики, измеряющей влияние каждого слоя на модификацию внутренних представлений, а на базе представительного набора конкретных задач.

Метод был экспериментально провалидирован на модели Qwen2.5-3B, и было показано, что предложенный итерационный подход существенно превосходит базовую версию алгоритма (ShortGPT). Подход позволил сократить количество слоев трансформера с 36 до 28 (уменьшив число параметров до 2.47 млрд), при этом падение итоговой метрики качества составило менее 10%.

ЗАКЛЮЧЕНИЕ

В ходе выполнения работ первого этапа проекта были проведены исследования в трех взаимосвязанных направлениях: создание эффективных технологий языковой адаптации, разработка инструментов и наборов данных для оценки и диагностики языковых моделей на русском языке и поиск методов структурного сжатия LLM.

1. Модифицирована и провалидирована методология адаптации больших языковых моделей на русский язык с заменой токенизации. В отличие от ранних версий, метод Ruadapt реализует четырехступенчатую схему: построение морфологически оптимизированного словаря методом Unigram, двухступенчатая адаптация базовой модели, проекция эмбедингов методом LEP (Learned Embedding Propagation) в инструктивные версии модели и финальная калибровка. Методология выложена в открытый доступ на github: <https://github.com/RefalMachine/ruadapt>

2. С использованием разработанной методологии Ruadapt были адаптированы 4 языковые модели (4, 8 и 32 миллиардов параметров) семейства Qwen3. Модели выложены в открытый доступ на платформе huggingface: <https://huggingface.co/collections/RefalMachine/ruadaptqwen3>

3. Разработаны новые методики и наборы данных для оценки больших языковых моделей:

3.1. В части оценки культурного соответствия проведено исследование эмоционального выравнивания моделей с носителями русского языка. Исследование выявило свойственную современным LLM тенденцию – склонность к гиперболизации негативных эмоций. Другим научным результатом стала демонстрация того, что предложенная методика адаптации модели на русский язык сближает “эмоциональную картину мира” модели и человека-носителя языка.

3.2. В области синтаксического анализа было проведено исследование по оценке различных LLM на наборах данных SynTagRus, GSD, PUD, Taiga, Poetry. Было показано, что языковые модели до сих пор существенно уступают специализированным моделям в задаче синтаксического анализа, а итоговое качество в основном коррелирует с размером модели.

3.3. Был создан новый набор данных RuBookSum для абстрактного реферирования (summarization) художественной литературы, который позволил протестировать работу моделей с длинным контекстом и сложной подачей информации. Также была предложена алгоритмическая оптимизация алгоритма иерархического абстрактного реферирования, которая обеспечила ускорение генерации итогового реферата до 300%. Код проекта RuBookSum выложен на github: <https://github.com/Nejimaki-Tori/WikiBench>. Набор данных выложен на платформе huggingface https://huggingface.co/datasets/NejimakiTori/literature_sum.

3.4. Был создан новый бенчмарк RuWikiBench, моделирующий процесс написания энциклопедических статей, который позволил оценивать поведение моделей на трех основных этапах: отбор источников, составление плана (оглавления), генерация секций. Было показано, что даже при идеальном отборе источников языковые модели испытывают трудности с воспроизведением экспертной логики создания оглавления, а с точки зрения качества генерации секций, адаптированная модель RuadaptQwen3-32B показала наилучшие результаты, превосходящие модели DeepSeek V3 и Qwen3-

235B-A22B. Код бенчмарка выложен на github: <https://github.com/Nejimaki-Tori/WikiBench>

4. Был разработан новый метод сжатия моделей Iterative Layer-wise Distillation, который позволяет снизить потери в качестве при подходе, ко-гда из модели полностью вырезаются некоторые слои. Предложенный ал-горитм, основанный на динамической переоценке важности слоев и итера-ционной дистилляции, позволил сократить количество слоев с 36 до 28 у модели Qwen2.5 3B при сохранении более 90% итогового качества. Код ме-тода выложен на github:

https://github.com/kaengreg/layer-wise_distillation

Были подготовлены рукописи и опубликованы и приняты в печать 6 на-учных ста-тей:

Тихомиров М. М., Чернышев Д. И. Ruadapt: вычислительно эффективная языковая адаптация больших языковых моделей //Доклады Российской ака-демии наук. Математика, информатика, процессы управления. – 2025. – Т.

527. – С. 290-300. (WoS, Scopus)

Григорьев Д. А., Чернышев Д. И. RuWikiBench: оценка больших языковых мо-делей посредством воспроизведения энциклопедических статей //Доклады Российской академии наук. Математика, информатика, процессы управле-ния. – 2025. – Т. 527. – С. 171-181. (WoS, Scopus)

Iaroshenko P. V., Louckachevitch N. V. Large Language Models versus Native Speakers in Emotional Assessment of Russian Words //Supercomputing Frontiers and Innovations. – 2025. – Т. 12. – №. 3. – С. 20–30. (Scopus)

Shamaeva E. D., Tikhomirov M. M., Loukachevitch N. V. One-Shot Prompting for Russian Dependency Parsing //Supercomputing Frontiers and Innovations. – 2025. – Т. 12. – №. 3. – С. 108–120. (Scopus)

Grigoriev D. A., Khudiakov D. V., Chernyshev D. I. RuBookSum: Dataset for Russian Literature Abstractive Summarization //Supercomputing Frontiers and Innovations. – 2025. – Т. 12. – №. 3. – С. 76–89-76–89.(Scopus)

Grigory Kovalev, Mikhail Tikhomirov. Iterative Layer-wise Distillation for Efficient Compression of Large Language Models // Pattern Recognition and Image Analysis 2025 (Принята к публикации, Scopus)

ПРИЛОЖЕНИЕ А
Объем финансирования темы в 2025 году Таблица А.1

Источник финанси- рования	Объем (руб.)	
	Получено	Освоено собственными силами
грант РНФ	7 000 000,00000	