«УТВЕРЖДАЮ»
Директор Федерального
государственного учреждения
«Федеральный исследовательский
центр «Информатика и управление»
Российской академии наук»,

_____ М.А. Посыпкин

«07» 05 2026 г.

ЗАКЛЮЧЕНИЕ

Федерального государственного учреждения «Федеральный исследовательский центр «Информатика и управление» Российской академии наук»

Диссертация Ватолина Алексея Сергеевича на тему: «Обучение и оценивание мультиязычных нейросетевых моделей семантического векторного представления научных текстов» выполнена в отделе №27 (математическое моделирование гетерогенных систем) Федерального исследовательского центра «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН).

Ватолин Алексей Сергеевич 1997 года рождения, гражданин России, в 2021 году окончил Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский авиационный институт (национальный исследовательский университет)» по направлению 02.04.02 «Фундаментальная информатика и информационные технологии».

В период подготовки диссертации соискатель Ватолин Алексей Сергеевич обучался в аспирантуре Федерального исследовательского центра «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН) по программе подготовки научных и научно-педагогических кадров в системе послевузовского профессионального образования по научной специальности 05.13.17 «Теоретические основы информатики», дата окончания 22 сентября 2025 г. Тема диссертационного исследования была утверждена на заседании ученого совета, протокол № 1 от 21 января 2022 года.

Справка о периоде обучения с результатами сдачи кандидатских экзаменов по специальности 05.13.17 «Теоретические основы информатики» выдана в 2025 г. в ФИЦ ИУ РАН. Справка о сдаче экзамена по специальности 2.3.8 «Информатика и информационные процессы» выдана в 2026 г. в ФИЦ ИУ РАН.

Научный руководитель д.ф.-м.н., Абгарян Каринэ Карленовна – работает в ФИЦ ИУ РАН в должности главного научного сотрудника отдела № 27.

По итогам обсуждения принято следующее заключение.

Актуальность темы

Актуальность диссертационного исследования обусловлена постоянным ростом объемов научной информации, что предъявляет повышенные требования к качеству и точности информационного поиска. Традиционные подходы, основанные на сопоставлении ключевых слов, не всегда позволяют в полной мере учесть смысловые связи между терминами и концепциями. В этих условиях особую значимость приобретают методы семантического поиска, основанные на нейросетевых моделях, в частности на архитектуре «трансформер», поскольку они способны анализировать контекст и улавливать смысловую

близость текстов, а не только формальное совпадение слов.

При этом для русскоязычного научного домена наблюдается нехватка как специализированных наборов данных для оценки качества таких моделей, так и эффективных легковесных моделей, доступных для широкого применения. Научные тексты обладают специфическими характеристиками: высокой информационной плотностью, особой терминологией и структурой, что делает универсальные оценочные наборы и модели не всегда подходящими для их анализа.

Таким образом, существует потребность в разработке и исследовании специализированных инструментов — моделей и наборов данных для их оценки, ориентированных на русскоязычные тексты. Данная работа направлена на решение этих задач, что и определяет её актуальность.

Обоснованность научных положений

Обоснованность и достоверность научных положений, выносимых на защиту, обеспечивается применением апробированных на практике методов и подходов. В основе разработанных моделей лежит широко используемая архитектура «трансформер», а их обучение проводилось с помощью распространенных методов, таких как маскированное языковое моделирование и контрастивное дообучение. Качество моделей оценивалось с использованием общепринятых в области информационного поиска и обработки естественного языка мер качества (Accuracy, F1-мера, NDCG@k, MRR@k, коэффициент корреляции Кенделла).

Личное участие соискателя ученой степени в получении результатов, изложенных в диссертации

Результаты, включенные в диссертацию, получены автором лично.

Степень достоверности результатов, проведенных соискателем ученой степени исследований

Достоверность результатов подтверждается публикациями в изданиях из перечня ВАК, докладами на конференциях, практическим внедрением разработанной модели на портале elibrary.ru, а также открытой публикацией исходного кода и данных, что обеспечивает воспроизводимость экспериментов.

Научная новизна работы

Научная новизна исследования заключается в сравнительном анализе двух стратегий контрастивного дообучения. Первая стратегия, основанная на использовании легкодоступных пар «заголовок-аннотация», доказывает возможность получения качественных семантических представлений без доступа к графу цитирований. Вторая стратегия использует данные о цитированиях в качестве более сильного семантического сигнала. В ходе исследования было показано, что цитирования позволяют достичь более высокого качества итоговой модели, в то время как пары «заголовок-аннотация» являются эффективной и менее затратной альтернативой.

Кроме того, новизна работы состоит в создании специализированных инструментов для оценки качества подобных моделей. Разработан комплекс данных и задач RuSciBench, обеспечивающий стандартизированную оценку на материале российского научного домена. Также предложена полуавтоматическая методика создания наборов данных для верификации научных фактов, сочетающая генерацию утверждений с помощью больших языковых моделей и экспертную валидацию, на основе которой был создан первый русскоязычный ресурс в этой области — RuSciFact.

Теоретическая значимость

Изучены и систематизированы закономерности в производительности современных мультиязычных моделей при их применении к совокупности научных текстов на разных

языках. На основе созданного двуязычного инструмента стандартизованной количественной оценки (бенчмарка) с парной структурой данных выявлен и количественно оценен систематический разрыв в качестве работы моделей на русском и английском языках, что ставит под сомнение гипотезу об их языковой инвариантности в специализированных областях.

Изучен и апробирован гибридный подход к формированию лингвистических ресурсов, сочетающий генерацию данных с помощью больших языковых моделей (LLM) и последующую экспертную верификацию. Данный подход вносит вклад в теорию создания специализированных наборов данных, демонстрируя способ преодоления дефицита ресурсов для ручной разметки в узких предметных областях.

Практическая значимость результатов, проведенных соискателем ученой степени исследований

Ключевым результатом, подтверждающим практическую ценность работы, является внедрение разработанной легковесной модели SciRus-tiny в информационно-поисковую систему крупнейшего российского научного портала eLibrary.ru, оператором которого является ООО «Научная электронная библиотека». На основе этой модели реализован новый режим «Нейропоиск», который позволяет пользователям находить тематически близкие научные публикации, используя в качестве запроса аннотацию или произвольный фрагмент текста. Применение этого режима значительно расширяет возможности традиционного поиска по ключевым словам, повышая релевантность и полноту получаемых результатов. Внедрение «Нейропоиска» упрощает для исследователей процесс анализа больших объемов научной информации и способствует более эффективному ориентированию в современном научном пространстве.

Апробация работы

Результаты работы докладывались и обсуждались на следующих международных научных конференциях:

1. Сравнительный анализ современных мультиязычных моделей для векторизации текста на русском языке. Международная научно-практическая конференция «Информационные технологии, искусственный интеллект, большие данные: актуальные тенденции, перспективные исследования», 2024 год.
2. ruSciFact: Open Benchmark for Verifying Scientific Facts in Russian. Международная конференция по компьютерной лингвистике и интеллектуальным технологиям «Диалог», 2025 год.
3. Structured Sentiment Analysis with Large Language Models: A Winning Solution for RuOpinionNE-2024. Международная конференция по компьютерной лингвистике и интеллектуальным технологиям «Диалог», 2025 год.

Полнота изложения материалов диссертации в публикациях

Результаты диссертационного исследования опубликованы в 5 работах общим объемом 106 п.л.; личный вклад автора составляет 57 п.л.

1. K. Enevoldsen, I. Chung, I. Kerboua, M. Kardos, A. Mathur, D. Stap, J. Gala, W. Siblini, D. Krzemiński, G. Indra Winata, S. Sturua, S. Utpala, M. Ciancone, M. Schaeffer, G. Sequeira, D. Misra, S. Dhakal, J. Rystrøm, R. Solomatin, A. Vatolin [и др.], MMTEB: Massive Multilingual Text Embedding Benchmark // ICLR 2025.
2. А. Ватолин, Н. Герасименко, А. Янина, К. Воронцов, RuSciBench: открытый бенчмарк для оценки семантических векторных представлений научных текстов на русском и английском языках // Доклады Российской академии наук. Математика, информатика, процессы управления. 2024. Том 520. № 2. С. 284-294.

3. Н. Герасименко, А. Ватолин, А. Янина, К. Воронцов, SciRus: легкий и мощный мультиязычный энкодер для научных текстов // Доклады Российской академии наук. Математика, информатика, процессы управления. 2024. Том 520. № 2. С. 193-202.
4. A. Vatolin, N. Gerasimenko, N. Loukachevitch, A. Ianina, K. Vorontsov, ruSciFact: Open Benchmark for Verifying Scientific Facts in Russian // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2025" April 23 - 25, 2025. V.23. pp. 435-456.
5. A. Vatolin, Structured Sentiment Analysis with Large Language Models: A Winning Solution for RuOpinionNE-2024 // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2025" April 23 - 25, 2025. V.23. pp. 416-434.

Публикации полностью соответствуют теме диссертационного исследования и раскрывают её основные положения.

Ценность научных работ соискателя ученой степени

Ценность научных работ соискателя заключается в том, что:

1. Разработаны практически значимые, легковесные нейросетевые модели (SciRus-tiny и SciRus-small), которые при значительно меньшем размере демонстрируют качество, сопоставимое с более крупными мультиязычными аналогами. Это делает передовые методы семантического поиска доступными для внедрения в реальные информационные системы с ограниченными вычислительными ресурсами, что доказано их успешным применением на портале eLibrary.ru.
2. Создан первый комплексный инструмент для стандартизированной оценки качества моделей, работающих с научными текстами на русском и английском языках. Это позволяет научному сообществу проводить воспроизводимые сравнительные исследования и выбирать наиболее эффективные решения для задач классификации, регрессии и информационного поиска, что способствует развитию и повышению качества алгоритмов в данной области. Актуальность и значимость бенчмарка подтверждается его интеграцией в международную систему оценки MTEB (Massive Text Embedding Benchmark).
3. Решена проблема отсутствия средств для оценки качества моделей верификации научных фактов на русском языке. Разработана полуавтоматическая методика, которая, в отличие от трудоемкого ручного сбора, позволяет эффективно создавать наборы данных, снижая затраты экспертного времени. На основе этой методики создан и опубликован RuSciFact - открытый набор данных и задач для проверки научных утверждений на русском языке, что открывает новое направление для исследований и разработок в области автоматической проверки фактов.

Диссертация Ватолина Алексея Сергеевича на тему: «Обучение и оценивание мультиязычных нейросетевых моделей семантического векторного представления научных текстов» – это законченная научно-квалификационная работа, которая соответствует требованиям пунктов 9, 10, 14 Положения о присуждении ученых степеней, утвержденного постановлением Правительства Российской Федерации от 24 сентября 2013 г. № 842, а также Паспорту научной специальности 2.3.8 — «Информатика и информационные процессы» (технические науки), в частности, по следующим пунктам:

4. Разработка методов и технологий цифровой обработки аудиовизуальной информации с целью обнаружения закономерностей в данных, включая обработку текстовых и иных изображений, видео контента. Разработка методов и моделей распознавания, понимания и синтеза речи, принципов и методов извлечения требуемой информации из текстов.
7. Разработка методов обработки, группировки и аннотирования информации, в

том числе, извлеченной из сети интернет, для систем поддержки принятия решений, интеллектуального поиска, анализа.

12. Разработка технологий извлечения и анализа информации в больших базах данных, в том числе, с использованием концепции многомерного представления (OLAP) и интеллектуального анализа данных (Data Mining) статического и в реальном масштабе времени, реализация моделей баз знаний.

Диссертация Ватолина Алексея Сергеевича на тему: «Обучение и оценивание мультиязычных нейросетевых моделей семантического векторного представления научных текстов» рекомендуется к защите на соискание ученой степени кандидата технических наук по специальности 2.3.8 «Информатика и информационные процессы».

Заключение принято на заседании отдела №27 (математическое моделирование гетерогенных систем) ФИЦ ИУ РАН «06» мая 2026 г., протокол № 1.

Присутствовало на заседании 14 человек.

Результаты голосования: «за» – 14 человек, «против» – нет, «воздержалось» – нет.

Председательствующий на заседании:

Старший научный сотрудник ФИЦ ИУ РАН,
доктор физико-математических наук — Морозов А.Ю.

Секретарь заседания:

Младший научный сотрудник ФИЦ ИУ РАН — Сеченых П.А.