

ЗАКЛЮЧЕНИЕ ДИССЕРТАЦИОННОГО СОВЕТА 24.1.224.03,
СОЗДАННОГО НА БАЗЕ ФЕДЕРАЛЬНОГО ГОСУДАРСТВЕННОГО
УЧРЕЖДЕНИЯ
«ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР
«ИНФОРМАТИКА И УПРАВЛЕНИЕ»
РОССИЙСКОЙ АКАДЕМИИ НАУК»,
ПО ДИССЕРТАЦИИ НА СОИСКАНИЕ УЧЕНОЙ СТЕПЕНИ
КАНДИДАТА ТЕХНИЧЕСКИХ НАУК

Аттестационное дело № _____

Решение диссертационного совета 24.1.224.03 от 18.12.2025 г., №13

О присуждении Ватолину Алексею Сергеевичу, гражданину Российской Федерации, учёной степени кандидата технических наук.

Диссертация «Обучение и оценивание мультязычных нейросетевых моделей семантического векторного представления научных текстов» по специальности 2.3.8 – «Информатика и информационные процессы» принята к защите 18 сентября 2025 г., протокол №8, диссертационным советом 24.1.224.03 на базе Федерального государственного учреждения «Федеральный исследовательский центр «Информатика и управление» Российской Академии Наук» (ФИЦ ИУ РАН), 119333, г. Москва, ул. Вавилова, д. 44/2, созданным на основании приказа Министерства науки и высшего образования Российской Федерации № 1732/нк от 13.12.2022.

Соискатель Ватолин Алексей Сергеевич, дата рождения 18 июня 1997 года, 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». С 2021 по 2025 г. обучался в очной аспирантуре ФИЦ ИУ РАН. В настоящее время работает в ООО «Яндекс.Технологии» в должности разработчика программного обеспечения.

Диссертация выполнена в отделе №27 ФИЦ ИУ РАН.

Научный руководитель – Абгарян Каринэ Карленовна, доктор физико-математических наук, доцент, главный научный сотрудник отдела №27 ФИЦ ИУ РАН.

Официальные оппоненты:

Котельников Евгений Вячеславович, доктор технических наук, доцент, профессор Школы вычислительных социальных наук Автономной некоммерческой

образовательной организации высшего образования «Европейский университет в Санкт-Петербурге»;

Кузнецов Сергей Олегович, доктор физико-математических наук, профессор департамента анализа данных и искусственного интеллекта Национального исследовательского университета «Высшая школа экономики»,

дали **положительные** отзывы на диссертацию.

Ведущая организация – Федеральное государственное бюджетное учреждение науки Институт проблем управления им. В. А. Трапезникова Российской академии наук (ИПУ РАН) в своём положительном отзыве, подписанном Губановым Дмитрием Алексеевичем, доктором технических наук, заведующим лабораторией № 25, и утвержденном директором института, академиком РАН Новиковым Дмитрием Александровичем, указала, что работа посвящена актуальной задаче создания инструментов автоматической обработки научных текстов на русском и английском языках.

В отзыве отмечено, что Ватолиным А.С. предложено комплексное решение проблемы дефицита эффективных моделей для русскоязычного научного домена. Ключевыми результатами названы разработка семейства компактных двуязычных моделей SciRus, создание открытых инструментариев для их оценки (RuSciBench и RuSciFact) и успешное внедрение разработок в инфраструктуру научной электронной библиотеки eLibrary.ru. Подчеркнута практическая ценность снижения требований к вычислительным ресурсам, что делает внедрение таких моделей доступным для широкого круга научных организаций.

В отзыве ведущей организации сформулирован ряд замечаний, не носящих принципиального характера и не снижающих общей положительной оценки диссертации. Отмечено, что диссертация полностью соответствует требованиям, установленным «Положением о присуждении ученых степеней», предъявляемым к диссертациям на соискание учёной степени кандидата технических наук по специальности 2.3.8 – «Информатика и информационные процессы», а её автор, Ватолин А.С., заслуживает присуждения ему учёной степени кандидата технических наук по данной специальности.

Основные результаты диссертации опубликованы в 5 работах в рецензируемых изданиях из перечня ВАК и приравненных наукометрических баз.

Недостоверные сведения об опубликованных соискателем работах в диссертации отсутствуют. Диссертация соответствует п. 14 Положения о присуждении учёных степеней. Автор подробно указал личный вклад в опубликованные в соавторстве работы.

Работы Ватолина А.С. по теме диссертации:

1. Герасименко Н., **Ватолин А.**, Янина А., Воронцов К. SciRus: легкий и мощный мультязычный энкодер для научных текстов // Доклады Российской академии наук. Математика, информатика, процессы управления. 2024. Том 520. № 2. С. 193-202 (**RSCI, Scopus**)
2. **Ватолин А.**, Герасименко Н., Янина А., Воронцов К. 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 (**RSCI, Scopus**)
3. **Vatolin A.**, Gerasimenko N., Loukachevitch N., Ianina A., Vorontsov K. ruSciFact: Open Benchmark for Verifying Scientific Facts in Russian // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue". 2025. V. 23. pp. S435–S459. DOI: 10.28995/2075-7182-2025-23-435-459 (**Scopus**)
4. **Vatolin A.** Structured Sentiment Analysis with Large Language Models: A Winning Solution for RuOpinionNE-2024 // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue". 2025. V. 23. pp. S416–S434. DOI: 10.28995/2075-7182-2025-23-416-434 (**Scopus**)
5. Enevoldsen K., Chung I., Kerboua I., Kardos M., Mathur A., Stap D., Gala J., Siblini W., Krzemiński D., Winata G. I., Sturua S., Utpala S., Ciancone M., Schaeffer M., Sequeira G., Misra D., Dhakal S., Rystrom J., Solomatin R., Çağatan Ö., Kundu A., Bernstorff M., Xiao S., Sukhlecha A., Auras D., Plüster B., Harries J. P., Magne L., Mohr I., Hendriksen M., Zhu D., Gisserot-Boukhlef H., Aarsen T., Kostkan J., Wojtasik K., Lee T., Šuppa M., Zhang C., Rocca R., Hamdy M., Michail A., Yang J., Faysse M., **Vatolin A.**, Thakur N., Dey M., Vasani D., Chitale P., Tedeschi S., Tai N., Snegirev A., Günther M., Xia M., Shi W., Lù X. H., Clive J., Krishnakumar G., Maksimova A., Wehrli S., Tikhonova M., Panchal H., Abramov A., Ostendorff M., Liu Z., Clematide S., Miranda L. J., Fenogenova A., Song G., Safi R. B., Li W.-D., Borghini A., Cassano F., Su H., Lin J., Yen H., Hansen L., Hooker S., Xiao C., Adlakha V., Weller O., Reddy S., Muennighoff N. MMTEB: Massive Multilingual Text Embedding Benchmark // International Conference on Representation Learning (ICLR). 2025. pp. 101715–101771. <https://openreview.net/forum?id=zl3pfz4VCV> (**Scopus**)

На автореферат поступили отзывы:

1. Куравского Льва Семеновича, д.т.н., профессора, декана факультета «Информационные технологии» ФГБОУ ВО «Московский государственный психолого-педагогический университет»;
2. Степаняна Ивана Викторовича, д.б.н., к.т.н., ведущего научного сотрудника лаборатории исследования биомеханических систем Института машиноведения им. А.А. Благонравова РАН;
3. Тихомирова Михаила Михайловича, к.ф.-м.н., старшего научного сотрудника НИВЦ МГУ.

Все отзывы положительные, замечания носят рекомендательный характер, не являются принципиальными и не влияют на главные теоретические и практические результаты диссертационной работы.

Также имеется акт о внедрении полученный от ООО НАУЧНАЯ ЭЛЕКТРОННАЯ БИБЛИОТЕКА.

Выбор официальных оппонентов и ведущей организации обосновывается их достижениями в области искусственного интеллекта, обработки естественного языка и анализа данных, что подтверждается их публикациями в авторитетных рецензируемых изданиях.

Диссертационный совет отмечает, что на основании выполненных соискателем исследований:

разработаны компактные двуязычные (русско-английские) нейросетевые модели SciRus-tiny и SciRus-small для семантического векторного представления научных текстов, которые при значительно меньшем числе параметров демонстрируют качество, сопоставимое с лучшими мировыми аналогами;

разработан первый мультизадачный русско-английский набор данных RuSciBench для оценки качества моделей векторного представления научных текстов, включающий задачи классификации, регрессии и информационного поиска;

предложена полуавтоматическая методика формирования наборов данных для задачи верификации научных фактов на русском языке, сочетающая генерацию утверждений с помощью больших языковых моделей (LLM) и экспертную валидацию, на основе которой создан тестовый набор данных RuSciFact;

показана эффективность стратегии контрастивного дообучения моделей на парах «заголовок-аннотация», позволяющая достигать высоких метрик качества без использования графа цитирований.

Теоретическая значимость работы заключается в изучении и систематизации закономерностей производительности мультиязычных моделей при их применении к научному домену на русском и английском языках. На основе созданного двуязычного бенчмарка выявлен и количественно оценен

систематический разрыв в качестве работы моделей на русском и английском языках, что уточняет границы применимости современных трансформерных моделей в специализированных доменах.

Практическая значимость исследования определяется созданием и внедрением модели SciRus-tiny в промышленную эксплуатацию на портале Научной электронной библиотеки eLibrary.ru в качестве основы сервиса семантического поиска «Нейропоиск», что подтверждено соответствующим актом о внедрении. Разработанные бенчмарки RuSciBench и RuSciFact выложены в открытый доступ и интегрированы в международный рейтинг MTEB (Massive Text Embedding Benchmark), что обеспечивает стандартизацию оценки русскоязычных моделей.

Оценка достоверности результатов исследования выявила их воспроизводимость в различных условиях функционирования представленного решения. Результаты подтверждены экспериментальной проверкой на общепризнанных (SciDocs) и разработанных автором (RuSciBench, RuSciFact) бенчмарках, сравнением с результатами state-of-the-art моделей (SPECTER, SciNCL, E5). Достоверность полученных результатов также подтверждена апробацией на международных научных конференциях и публикациями в рецензируемых журналах.

Личный вклад соискателя в работах с соавторами заключается в следующем: в работе [1] проведено предобучение малоразмерной версии модели (SciRus-tiny) с помощью маскированного языкового моделирования, дообучение обеих версий модели (SciRus-tiny и SciRus-small) на парах заголовков-аннотация, а также на парах «цитирующая статья – цитируемая статья», валидация моделей на наборе данных SciDocs. В [2] автором собраны датасеты для классификации по ГРНТИ, по типу публикации, для поиска цитирований, для регрессии по количеству цитат, для поиска английского перевода по тексту на русском языке. Также реализация исходного кода инструментария для оценки, валидация моделей, интеграция бенчмарка в международный бенчмарк MTEB. В работе [3] вклад соискателя является определяющим. В публикации [5] в рамках работы над созданием и расширением международного многоязычного бенчмарка MMTEB соискателем осуществлено добавление новых задач на различных языках, включая задачи, разработанные им самостоятельно. Также выполнен значительный вклад в обеспечение качества и корректности данных во всех задачах, вошедших в итоговый набор бенчмарка.

В ходе защиты диссертации были высказаны следующие замечания:

1. В работе представлена методика разработки моделей SciRus с использованием собственного предобучения, однако не обсуждаются альтернативы, такие как дообучение существующих базовых моделей.

2. Валидация данных RuSciFact выполняется с привлечением экспертной разметки, однако в работе не приведены показатели согласованности разметчиков (например, Cohen's kappa, или Krippendorff's alpha).

3. Одна из двух задач бенчмарка RuSciFact, а именно задача поиска, вероятно достаточно проста для современных моделей векторизаторов, так как лучшие из представленных моделей показывают метрики качества близкие к 100%.

Соискатель Ватолин А.С. ответил на заданные вопросы и согласился с замечаниями, указанными ему в ходе заседания.

На заседании 18 декабря 2025 г. диссертационный совет принял решение присудить Ватолину Алексею Сергеевичу учёную степень кандидата технических наук за существенный вклад в развитие методов компьютерной лингвистики и искусственного интеллекта для анализа научной информации.

При проведении тайного голосования диссертационный совет в количестве 23 человек, из них 6 докторов наук по специальности 2.3.8 – «Информатика и информационные процессы», участвовавших в заседании, из 29 человек, входящих в состав совета, дополнительно введены на разовую защиту 0 человек, проголосовали: за – 23, против – 0, недействительных бюллетеней 0.

Председатель

диссертационного совета 24.1.224.03, д.ф.-м.н.

Воронцов К.В.

Учёный секретарь

диссертационного совета 24.1.224.03, к.т.н.

Рейер И.А.

18 декабря 2025 г.

