Применение больших языковых моделей для автоматизации учета научных достижений обучающихся образовательных организаций

Грошев Артур Андреевич
студент, Программная инженерия и интеллектуальные системы, Балтийский государственный технический университет «ВОЕНМЕХ» им. Д.Ф. Устинова, wiezerion@gmail.com,

Фирсова Елизавета Дмитриевна
студент, Программная инженерия и интеллектуальные системы, Балтийский государственный технический университет «ВОЕНМЕХ» им. Д.Ф. Устинова, lizafirsova2006@yandex.ru,

Наурусова Гульнара Ахмановна
Заместитель декана факультета Робототехника и инновационная инженерия, Балтийский государственный технический университет «ВОЕНМЕХ» им. Д.Ф. Устинова, gan.vuz@gmail.com

В статье рассматривается проблема автоматизации учета научных достижений студентов образовательных организаций. Выполнен анализ существующих решений, предназначенных для хранения сведений о научных достижениях и формирования отчетности, выявлены их основные ограничения при использовании в локальной инфраструктуре вузов. Предложен подход к автоматизации подготовки отчетных документов, основанный на совместном использовании централизованного хранения данных, методов компьютерного зрения и технологий искусственного интеллекта. Рассмотрены особенности применения больших языковых моделей для обработки подтверждающих документов, выполнен сравнительный анализ современных моделей и обоснован выбор Mistral 7B для интеграции в информационную систему. В статье приведены примеры реализации интерфейса для различных категорий пользователей: студентов, руководителей СНО и администрации. Представленная архитектура обеспечивает масштабируемость и высокую производительность. Разработанная платформа позволяет сократить время подготовки отчетных документов и упростить контроль за активностью студентов. Результаты нагрузочного тестирования разработанного модуля подтверждают возможность его практического применения при автоматизации подготовки отчетности в образовательной организации.

Ключевые слова: цифровизация образования, автоматизация учёта результатов научной деятельности, студенческое научное общество, информационная система, искусственный интеллект, большие языковые модели, OCR, компьютерное зрение, автоматизированное формирование отчётности.

Введение

В современных условиях цифровой трансформации высшего образования возрастает роль научно-исследовательской деятельности студентов, аспирантов и молодых учёных. Согласно данным аналитического доклада Российской академии наук, подготовленного для президента страны, в 2023 году доля исследователей в секторе высшего образования увеличилась по сравнению с 2000 годом почти на 14%, а за период с 2010 по 2023 год количество научных публикаций российских авторов, представленных в международных библиографических системах, увеличилось в 2,7 раза [1].

Объем информации, которую нужно регистрировать, хранить и анализировать растет быстрее, чем совершенствуются инструменты для этой работы. Во многих вузах учет научных достижений студентов по-прежнему ведется разрозненно, как показывает сложившая практика, учет результатов научной деятельности студентов имеет ряд системных недостатков: данные рассредоточены между различными подразделениями вуза, отсутствуют единые стандарты фиксации достижений и не налажены процессы оперативного обновления информации [2]. Отсутствие единой системы не позволяет объективно оценивать вклад каждого студента, своевременно формировать отчетность и принимать обоснованные управленческие решения.

Как определено Постановлением Правительства Российской Федерации, разработка профессиональных стандартов осуществляется с учетом приоритетных направлений развития экономики, что напрямую связывает требования к компетенциям специалистов с актуальными задачами цифровизации и требует создания инструментов, позволяющих объективно оценивать и демонстрировать сформированные компетенции [3, п. 8].

Одним из перспективных направлений решения этой задачи является применение технологий искусственного интеллекта, в частности больших языковых моделей, для автоматизации обработки подтверждающих документов и формирования отчетности.

Таким образом, актуальность исследования обусловлена необходимостью экспериментальной проверки применимости больших языковых моделей для автоматизации обработки подтверждающих документов и формирования отчётности о научных достижениях студентов, а также выявления условий, при которых такое применение обеспечивает приемлемую достоверность и производительность. Научная новизна работы заключается в проведении сравнительного анализа эффективность больших языковых моделей в задаче извлечения информации из подтверждающих документов о научных достижениях студентов. В отличие от существующих исследований, в работе оценивается не только качество извлечения данных, но и влияние стратегий промптинга на достоверность результатов, а также проводится нагрузочное тестирование разработанного модуля в условиях, приближенных к реальной эксплуатации.

Целью работы является исследование применимости больших языковых моделей для автоматизации обработки подтверждающих документов и формирования отчётности о научных достижениях студентов.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Провести анализ существующих программных решений для хранения и управления информацией о результатах научной и образовательной деятельности участников.

2. Исследовать особенности процессов сбора, обработки и представления данных о деятельности участников научно-творческих объединений.

3. Определить основные требования и функциональные возможности информационных систем, предназначенных для автоматизации учёта научных результатов.

4. Рассмотреть перспективы применения современных технологий, включая методы искусственного интеллекта, для повышения эффективности обработки данных и формирования отчётных материалов.

Анализ существующих подходов и систем

Для определения целесообразности применения больших языковых моделей в задаче автоматизации учета научных достижений студентов необходимо рассмотреть какие процесс сегодня выполняются вручную, какие программные решения используются и почему они не обеспечивают требуемого уровня автоматизации.

Учет результатов деятельности обучающихся образовательных организаций представляет собой комплексный процесс, включающий сбор информации, проверку достоверности данных классификацию результатов и формирование отчетных материалов. В зависимости от структуры организации в процессе могут учувствовать несколько категорий пользователей: студенты, научные руководители или кураторы студенческих научных обществ, а также административные сотрудники. Обучающиеся предоставляют сведения о результатах своей деятельности и подтверждающие документы. Ответственные лица осуществляют проверку представленных данных, их распределение по категориям и подготовку итоговой отчетности.

В состав учитываемой информации могут входить следующие категории:
– участие в научных конференциях, семинарах и других мероприятиях;
– сведения о публикационной активности;
– результаты участия в конкурсах, олимпиадах и грантовых программах;
– участие в исследовательских проектах;
– выполнение научно-исследовательских работ;
– сведения об объектах интеллектуальной собственности;
– иные индивидуальные достижения научного и творческого характера.

При отсутствии единой информационной системы данные категории могут храниться в различных документах и файлах, что затрудняет их повторное использование и проведение комплексного анализа.

Для автоматизации учёта результатов деятельности участников могут использоваться различные программные средства. В качестве аналогов можно выделить платформы электронных портфолио (Mahara), федеральные системы учёта достижений (Talent.Krushok), а также модули систем управления обучением на базе Moodle.

Mahara - является открытой, гибкой платформой для создания электронных портфолио [2]. Система позволяет создавать персональные страницы с текстовыми материалами, изображениями и файлами, настраивать права доступа для преподавателей и кураторов. Поддерживается групповая работа и экспорт материалов в форматы PDF и HTML, возможна интеграция с Moodle. Однако функциональность платформы ориентирована на представление индивидуальных достижений, средства формирования отчётности по утверждённым кафедральным шаблонам и автоматическое создание документов формата DOCX не предусмотрены.

Talent.Krushok – российская платформа, более ориентированная на управление талантами и компетенциями. Она предоставляет инструменты для формирования цифрового следа студента, включая образовательные внеучебные достижения [2]. Однако функциональность ориентирована на учёт достижений в рамках федеральной системы. Поддержка отдельных форм научной отчётности кафедры, таких как учёт НИР или публикаций в соответствии с внутренними нормативными требованиями образовательной организации, не предусмотрена. Данные хранятся в системе вне инфраструктуры вуза, что создаёт риски при формировании внутренней отчётности.

Moodle – открытая система управления обучением, широко используемая в образовательных организациях для организации онлайн-курсов, проверки заданий и учёта успеваемости. Система поддерживает загрузку файлов, оценку работ, ведение журналов успеваемости и формирование базовых отчётов. Возможна интеграция с внешними инструментами через плагины и API, включая модули портфолио и автоматизации отчётности. Поддерживается разграничение прав доступа для студентов, преподавателей и администраторов. Однако функциональность Moodle ориентирована в первую очередь на учебный процесс, а не на учёт научных достижений СНО. Специализированная поддержка учёта НИР, публикаций, участия в конференциях и формирование отчётности по кафедральным шаблонам в стандартной поставке не предусмотрена.

Таким образом, анализ существующих подходов показывает, что учет научных достижений студентов в большинстве вузов по-прежнему осуществляется с использованием ручных методов или разрозненных инструментов, а существующие автоматизированные системы не обеспечивают полного цикла сбора, обработки и формирования отчетности.

В связи с этим возникает вопрос о возможности применения современных технологий искусственного интеллекта, в частности больших языковых моделей, для автоматизации обработки подтверждающих документов. В отличие от традиционных подходов, большие языковые модели способны анализировать неструктурированный текст, выделять значимые сущности и формировать структурированный вывод. Однако остаётся открытым вопрос о том, насколько эффективно и достоверно современные модели справляются с извлечением информации из документов в условиях локальной инфраструктуры образовательной организации.

Большие языковые модели как инструмент обработки документов

Развитие технологий обработки естественного языка в последние годы связано с появлением архитектуры трансформеров, которая изменила подход к работе с текстовыми данными. Архитектурной основой всех рассматриваемых моделей служит трансформер Vaswani et al. (2017), а его ключевое нововведение – механизм самовнимания (self-attention): каждый токен одновременно обращается ко всем остальным и получает взвешенный агрегат их представлений, устанавливая прямые связи между элементами, разделёнными сотнями слов [4]. В отличие от рекуррентных сетей, трансформер не теряет контекст начала документа к моменту обработки его конца, что делает его особенно эффективным для анализа длинных текстов.

По сравнению с традиционными подходами, основанными на жёстких правилах и регулярных выражениях, большие языковые модели обладают рядом преимуществ. Они способны анализировать тексты различной структуры, не требуя определенного формата входных документов; выделять значимые сущности без предварительного описания всех возможных шаблонов; обрабатывать документы с опечатками и нестандартными формулировками; формировать структурированный вывод в формате JSON, пригодный для автоматической загрузки в базу данных.

В последние годы всё большее распространение получают открытые большие языковые модели, доступные для локального развёртывания и дообучения под специализированные задачи. Это особенно важно для образовательных организаций, где вычислительные ресурсы могут быть ограничены, а использование облачных сервисов сопряжено с рисками передачи персональных данных третьим лицам.

Таким образом, для практического применения больших языковых моделей в задаче автоматизации учёта научных достижений требуется экспериментальная проверка их эффективности: какие модели обеспечивают наилучшее соотношение качества и производительности, какие стратегии формулировки запросов дают наиболее достоверные результаты, и насколько решение устойчиво при увеличении нагрузки.

Экспериментального исследование применимости больших языковых моделей для обработки документов

Для оценки эффективности применения больших языковых моделей в задаче автоматизации обработки подтверждающих документов было проведено исследование на базе разработанного веб-приложения для учета научных достижений студентов. В качестве исходных данных использовались подтверждающие документы участников студенческого научного общества: дипломы, сертификаты, грамоты, программы конференций и другие материалы, содержащие информацию о мероприятиях, авторах и результатах участия.

Первоначально предполагалось использовать модель искусственного интеллекта для полной автоматизации обработки подтверждающих документов. После загрузки документа система должна была определять тип достижения, извлекать сведения о мероприятии, его статусе, уровне проведения, результатах участия, составе авторов и других реквизитах, необходимых для последующего заполнения базы данных.

Подобный подход позволял существенно сократить объём ручного ввода информации и практически полностью автоматизировать регистрацию достижений участников.

Однако в ходе предварительных экспериментов было установлено, что использование только большой языковой модели не обеспечивает требуемой достоверности результатов. Несмотря на высокое качество обработки текста, модель в отдельных случаях неверно определяла тип достижения, допускала ошибки при извлечении отдельных реквизитов либо формировала сведения, отсутствующие в исходном документе. Такие ошибки недопустимы при подготовке официальной отчётности образовательной организации.

Дополнительным источником ошибок является этап распознавания текста. Перед передачей документа языковой модели выполняется его обработка средствами компьютерного зрения и OCR. Ошибки распознавания отдельных символов, слов или фрагментов документа приводят к искажению входных данных, что снижает качество последующего анализа искусственным интеллектом. Таким образом, итоговая точность обработки определяется не только возможностями языковой модели, но и качеством предварительного распознавания документа.

По результатам исследования было принято решение изменить область применения искусственного интеллекта. Вместо полной автоматизации заполнения карточки достижения модель используется для извлечения только наиболее трудоёмких текстовых реквизитов, включая название мероприятия, сведения о соавторах и результат участия. Остальные поля пользователь заполняет самостоятельно либо выбирает из заранее определённых справочников.

После определения области применения искусственного интеллекта возникла необходимость выбора языковой модели, обеспечивающей приемлемое качество извлечения текстовой информации при возможности локального развёртывания информационной системы. В связи с этим был проведён сравнительный анализ современных больших языковых моделей.

Сравнительный анализ проводился на четырёх моделях: LLaMA 3-8B, Mistral 7B, Gemma 7B и Mixtral 8×7B. Оценка производилась по скорости генерации, времени отклика и объёму формируемого ответа. По результатам эксперимента для дальнейшей работы была выбрана модель Mistral 7B, которая демонстрировала наилучшее соотношение скорости генерации, вычислительной нагрузки и устойчивости следования заданным структурным форматам [5].

Выбранная модель Mistral 7B была интегрирована в разрабатываемое веб-приложение для автоматизации учёта результатов деятельности участников научно-творческих объединений. В разработанной системе модель используется для извлечения названия мероприятия, сведений о соавторах и результата участия из подтверждающих документов. Полученные данные отображаются пользователю для проверки и при необходимости могут быть скорректированы перед сохранением в базе данных. Такой подход позволяет снизить вероятность ошибок при сохранении официальной информации и сохранить контроль пользователя над корректностью извлечённых данных.

Для оценки возможности практического использования разработанной системы было проведено нагрузочное тестирование модуля автоматического извлечения данных из документов. Результаты представлены таблице 1.

Тестирование проводилось на локальном сервере, характеристики которого представлены на рисунке 1.

Таблица 1
Результаты нагрузочного тестирования

Количество пользователей	Среднее время ответа	Запросы в секунду
1	6356,48	0,16
5	10637,44	0,3
10	17924,45	0,4
25	25619,3	0,5
40	30521,79	0,43
50	47693,9	0,4



Рисунок 1 – Системные характеристики устройства

На рисунке 2 представлен график зависимости пропускной способности сервера от количества одновременно работающих пользователей.

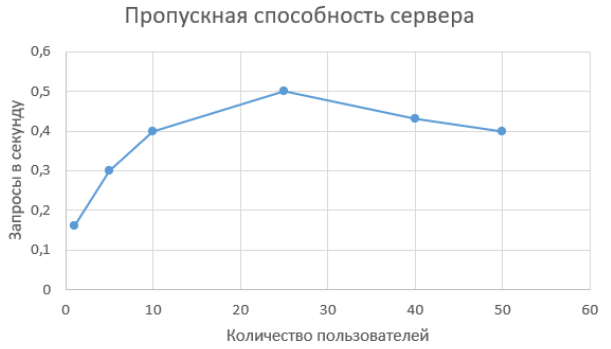


Рисунок 2 –Зависимость пропускной способности сервера от количества пользователей

На рисунке 3 представлен график зависимости среднего времени ответа модуля от количества одновременно работающих пользователей.

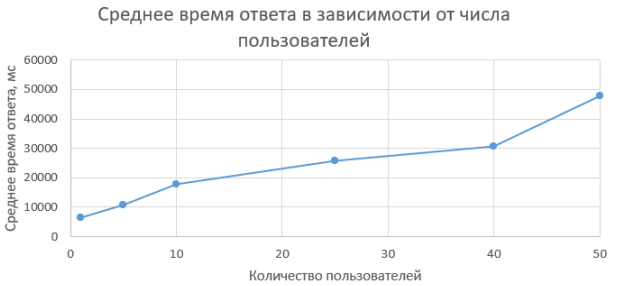


Рисунок 3 – График зависимости количества пользователей от среднего времени ответа

Анализ результатов нагрузочного тестирования показывает, что с увеличением количества одновременно работающих пользователей наблюдается закономерный рост среднего времени обработки запросов. При увеличении нагрузки с 1 до 40 пользователей среднее время ответа возрастает с 6,4 до 30,5 с, что обусловлено высокой вычислительной сложностью операций OCR и обработки текста языковой моделью. Максимальная пропускная способность достигается при нагрузке около 25–40 одновременно работающих пользователей и составляет около 0,5 запроса в секунду. Дальнейшее увеличение количества пользователей приводит к росту времени ожидания и снижению пропускной способности сервера вследствие исчерпания доступных вычислительных ресурсов. Полученные результаты свидетельствуют о возможности использования разработанного решения в условиях небольших подразделений образовательной организации, где отсутствует необходимость одновременной обработки большого количества документов. При увеличении нагрузки целесообразно применение очередей обработки запросов, масштабирование вычислительных ресурсов либо выделение отдельного сервера для выполнения операций OCR и анализа документов средствами искусственного интеллекта.

Заключение

Проведен анализ существующих программных решений для хранения и управления информацией о результатах научной деятельности студентов. Рассмотрены платформы электронных портфолио Рассмотрены платформы электронных портфолио (Mahara), федеральные системы учёта достижений (Talent.Krushok) и системы управления обучением (Moodle). Выявлено, что ни одно из решений не обеспечивает полного цикла автоматизации учёта научных достижений: Mahara ориентирована на индивидуальное использование и не поддерживает формирование отчётности по внутренним шаблонам вуза; Talent.Krushok привязана к федеральным мероприятиям и не учитывает специфику кафедральной отчётности; Moodle нацелена на учебный процесс, а не на учёт научной деятельности обучающихся.

Исследованы особенности процессов сбора, обработки и представления данных о деятельности участников научно-творческих объединений. Установлено, что в состав учитываемой информации входят публикации, участие в конференциях, конкурсах, грантах, исследовательских проектах, объекты интеллектуальной собственности и иные индивидуальные достижения. При отсутствии единой информационной системы эти данные хранятся в различных документах и файлах, что приводит к дублированию, ошибкам и значительным временным затратам при формировании отчётности.

Определены основные требования к информационной системе для автоматизации учёта научных результатов: централизованное хранение данных, разграничение прав доступа между студентами, руководителями СНО и администрацией, поддержка загрузки и обработки подтверждающих документов, автоматическое формирование отчётности по утверждённым шаблонам.

Рассмотрены перспективы применения технологий искусственного интеллекта, в частности больших языковых моделей, для повышения эффективности обработки данных и формирования отчётных материалов. В ходе экспериментального исследования установлено, что использование LLM для полной автоматизации обработки подтверждающих документов не обеспечивает требуемой достоверности из-за ошибок OCR и особенностей работы языковых моделей. Наиболее эффективным признан подход, при котором модель извлекает только наиболее трудоёмкие текстовые реквизиты, а окончательная проверка выполняется пользователем. Сравнительный анализ четырёх моделей показал, что Mistral 7B обеспечивает наилучшее соотношение скорости, вычислительной нагрузки и устойчивости к заданным форматам. Результаты нагрузочного тестирования подтвердили работоспособность решения при одновременной работе до 40 пользователей.

Таким образом, предложенный подход к автоматизации учёта научных достижений студентов, основанный на совместном использовании централизованного хранения данных, методов компьютерного зрения и больших языковых моделей, позволяет сократить время подготовки отчётных документов, упростить контроль за активностью студентов и снизить вероятность ошибок. Дальнейшие исследования предполагают масштабирование системы на большее количество пользователей, дообучение моделей на специфических данных и разработку методов автоматической верификации извлечённой информации.

Литература

1. Веденеева Наталья. Доклад РАН: финансирование науки падает, количество научных работников растет – «МК». — URL: https://scientificrussia.ru/articles/doklad-ran-finansirovanie-nauki-padaet-kolicestvo-naucnyh-rabotnikov-rastet (дата об ращения: 08.08.2026).
2. Фирсова, Е. Д. Разработка цифровой среды для мониторинга достижений участников студенческого научного общества вуза / Е. Д. Фирсова, А. А. Грошев, Г. А. Наурусова // "VI Конференция Молодежного научно-инженерного центра" для школьников, студентов, аспирантов и молодых учёных : Материалы конференции. В 2-х томах, Санкт-Петербург, 08–09 декабря 2025 года. – Санкт-Петербург: Балтийский государственный технический университет "ВОЕНМЕХ" им. Д.Ф. Устинова, 2026. – С. 150-155.
3. Постановление Правительства Российской Федерации от 22 января 2013 г. № 23 «О Правилах разработки, утверждения и применения профессиональных стандартов». - URL: https://www.garant.ru/products/ipo/prime/doc/70204190/ (дата обращения: 15.07.2026).
4. Тлегенов, Р. Н. Применение больших языковых моделей (LLM) для автоматической классификации и индексации неструктурированных документов // Universum: технические науки : электрон. научн. журн. – 2026. - № 5(146). – DOI: 10.32743/UniTech.2026.146.5.22660. – URL: https://cyberleninka.ru/article/n/primenenie-bolshih-yazykovyh-modeley-llm-dlya-avtomaticheskoy-klassifikatsii-i-indeksatsii-nestrukturirovannyh-dokumentov/viewer (дата обращения: 31.07.2026).
5. Грошев, А. А. Создание модуля отчетности для учета результатов участников сно вуза на базе ИИ-технологий / А. А. Грошев, Г. А. Наурусова, Е. Д. Фирсова // Как понять машинный разум: между кодом и культурой : Материалы Международного научного форума, Санкт-Петербург, Мариинский дворец, 09 декабря 2025 года. – Санкт-Петербург: ООО "Скифия-принт", 2026. – С. 208-216.
6. Грошев, А. А. Разработка веб приложения для учёта деятельности студенческого научного общества вуза / А. А. Грошев, Г. А. Наурусова // Информационные технологии в высокотехнологичных производствах (ВТП) : Сборник тезисов докладов IV Всероссийской молодежной научной конференции, Санкт-Петербург, 12–13 марта 2026 года. – Санкт-Петербург:

Балтийский государственный технический университет "ВОЕНМЕХ" им. Д.Ф. Устинова, 2026. – С. 91-92.

The use of large language models to automate the accounting of scientific achievements of students in educational organizations

Groshev A.A., Firsova E.D., Naurusova G.A.
Baltic State Technical University "VOENMEH" named after D.F. Ustinov

JEL classification: C01, C02, C1, C4, C5, C6, C8

The article discusses the problem of automating the accounting of scientific achievements of students in educational institutions. The analysis of existing solutions designed to store information about scientific achievements and generate reports has been carried out, and their main limitations have been identified when used in the local infrastructure of universities. An approach to automating the preparation of accounting documents based on the combined use of centralized data storage, computer vision methods and artificial intelligence technologies is proposed. The features of using large language models for processing supporting documents are considered, a comparative analysis of modern models is performed, and the choice of Mistral 7B for integration into the information system is justified. The article provides examples of interface implementation for various categories of users: students, heads of secondary schools and administration. The presented architecture provides scalability and high performance. The developed platform makes it possible to reduce the preparation time of accounting documents and simplify the monitoring of student activity. The results of the load testing of the developed module confirm the possibility of its practical application in automating the preparation of reports in an educational organization.

Keywords: digitalization of education, automation of accounting for scientific results, student scientific society, information system, artificial intelligence, large language models, OCR, computer vision, automated reporting.

References

1. Vedeneeva Natalia. RAS report: science funding is falling, the number of researchers is growing – MK. — URL: https://scientificrussia.ru/articles/doklad-ran-finansirovanie-nauki-padaet-kolicestvo-naucnyh-rabotnikov-rastet (date of notification: 08/08/2026).
2. Firsova, E. D. Development of a digital environment for monitoring the achievements of university student scientific society participants / E. D. Firsova, A. A. Groshev, G. A. Naurusova // "VI Conference of the Youth Scientific and Engineering Center" for schoolchildren, students, postgraduates and young scientists : Conference materials. In 2 volumes, St. Petersburg, December 08-09, 2025. Saint Petersburg: D.F. Ustinov Baltic State Technical University "VOENMEH", 2026, pp. 150-155.
3. Decree of the Government of the Russian Federation No. 23 dated January 22, 2013 "On the Rules for the Development, Approval and Application of Professional Standards". - URL: https://www.garant.ru/products/ipo/prime/doc/70204190 / (date of access: 07/15/2026).
4. Tlegenov, R. N. Application of large language models (LLM) for automatic classification and indexing of unstructured documents // Universum: technical sciences : electron. scientific journal. – 2026. - № 5(146). – DOI: 10.32743/UniTech.2026.146.5.22660. – URL: https://cyberleninka.ru/article/n/primenenie-bolshih-yazykovyh-modeley-llm-dlya-avtomaticheskoy-klassifikatsii-i-indeksatsii-nestrukturirovannyh-dokumentov/ viewer (date of request: 07/31/2026).
5. Groshev A. A., Naurusova G. A., Firsova E. D. Creating a reporting module to take into account the results of the participants of the Higher School of Economics on the basis of AI technologies // How to understand the machine mind: between code and culture : Proceedings of the International Scientific Forum, St. Petersburg, Mariinsky Palace, December 09, 2025. Saint Petersburg: Scythia-Print LLC, 2026, pp. 208-216.
6. Groshev, A. A. Development of a web application for accounting for the activities of the student scientific society of the university / A. A. Groshev, G. A. Naurusova // Information technologies in high-tech industries : Collection of abstracts of the IV All-Russian Youth Scientific Conference, St. Petersburg, March 12-13, 2026. Saint Petersburg: D.F. Ustinov Baltic State Technical University "VOENMEH", 2026, pp. 91-92.