

АВТОМАТИЧЕСКОЕ ФОРМИРОВАНИЕ АНАЛИТИЧЕСКИХ ОТЧЕТОВ О СТУДЕНЧЕСКОЙ УСПЕВАЕМОСТИ НА ОСНОВЕ ДООБУЧЕННОЙ ЯЗЫКОВОЙ МОДЕЛИ GPT-2

AUTOMATIC GENERATION OF ANALYTICAL REPORTS ON STUDENT PERFORMANCE BASED ON THE ADVANCED GPT-2 LANGUAGE MODEL

**D. Maksimov
Ph. Chervakov**

Summary. The article presents a method for the automated generation of analytical reports on student academic performance using a fine-tuned GPT-2 language model. The relevance of this study is driven by the growing volumes of educational data and the need for timely analysis of student performance to provide prompt feedback. The authors compiled a specialized dataset containing students' academic indicators and instructors' textual reports. Using this corpus, GPT-2 (774M parameters) was fine-tuned leveraging the HuggingFace Transformers library and PyTorch. As a result, the model effectively generated human-like analytical texts. Evaluation included automatic metrics (BLEU, ROUGE) and expert assessments by instructors. Generated reports demonstrated high accuracy (4.8/5), clarity (4.6/5), and usefulness of recommendations (4.7/5). Automation significantly reduced the workload of educators and enhanced personalized feedback accessibility for students. Discussed limitations include the necessity for high-quality training data and the model's limited context window. Future research directions encompass employing more advanced language models and integrating generated reports into educational practices to evaluate their impact on learning outcomes.

Keywords: analytical reports, student performance, automatic text generation, natural language processing, GPT-2, neural networks, machine learning, model fine-tuning, educational data, feedback.

Максимов Денис Романович
Национальный Исследовательский
Ядерный Университет «МИФИ»
maksimov.official@mail.ru
Черваков Филипп Сергеевич
Национальный Исследовательский
Ядерный Университет «МИФИ»
pschervakov@gmail.com

Аннотация. В статье рассматривается метод автоматического формирования аналитических отчетов об успеваемости студентов с использованием дообученной языковой модели GPT-2. Актуальность работы обусловлена значительным увеличением объемов образовательных данных и необходимостью оперативного анализа успеваемости студентов для предоставления своевременной обратной связи. Авторами был подготовлен специализированный корпус данных, содержащий показатели академической успеваемости и текстовые отчеты преподавателей. На его основе осуществлено дообучение модели GPT-2 (774 млн параметров) с использованием библиотеки HuggingFace Transformers и PyTorch. В результате модель приобрела способность генерировать тексты, близкие по качеству к составленным человеком. Оценка производилась при помощи автоматических метрик (BLEU, ROUGE) и экспертного оценивания преподавателями. Сгенерированные отчеты показали высокую точность (4.8/5), ясность (4.6/5) и полезность рекомендаций (4.7/5). Автоматизированное создание отчетов существенно сокращает трудозатраты преподавателей и повышает доступность персонализированной обратной связи для студентов. Среди обсуждаемых ограничений — необходимость качественного корпуса данных и ограниченный контекст модели. Перспективы дальнейших исследований включают применение более мощных моделей и интеграцию с педагогическими практиками для оценки влияния на образовательные результаты.

Ключевые слова: аналитические отчеты, студенческая успеваемость, автоматическая генерация текстов, обработка естественного языка, GPT-2, нейронные сети, машинное обучение, дообучение моделей, образовательные данные, обратная связь.

Введение

В современном образовании стало нормой накопление больших объемов данных об успеваемости студентов — результаты тестов, посещаемость, активность в онлайн-курсах и др. [3]. Для их анализа сформировалась область Educational Data Mining (EDM) [3]. Однако вручную обрабатывать такие данные сложно: по мере роста числа студентов и объема информации подготовка отчетов требует всё больше времени и ресурсов [8, с. 220]. Особенно это актуально в боль-

ших классах, где анализ результатов становится трудоёмким [3].

При этом своевременная персонализированная обратная связь критически важна для обучения. Регулярные отчеты помогают студентам понимать свои слабые стороны, а преподавателям — корректировать учебный процесс. Однако подготовка таких отчетов вручную затруднена. Здесь могут помочь современные языковые модели. Модели вроде GPT, основанные на архитектуре Transformer, способны генерировать связные тексты, имитируя человеческий язык [4].

В данной работе представлен подход к автоматическому формированию отчётов об успеваемости с использованием дообученной модели GPT-2. Она была выбрана как открытая трансформер-модель, подходящая для обучения на специализированных корпусах. Модель дообучена на данных об успеваемости студентов для генерации понятных отчётов. Цель — сократить время преподавателей на подготовку аналитики и обеспечить быструю обратную связь. Рассматриваются проблемы интеграции данных [1] и оценки качества отчётов с использованием метрик и экспертных оценок.

Таким образом, работа актуальна ввиду потребности в автоматизированных инструментах анализа образовательных данных и генерации отчётов, способных повысить эффективность и качество обучения.

Смежные работы

Традиционные подходы к генерации образовательных отчетов.

Автоматическая генерация текстов на основе образовательных данных — задача, объединяющая аналитику обучения и Natural Language Generation (NLG). Ранние решения, такие как система LARC из Университета Эдинбурга, генерировали отчеты по данным Moodle, сравнивая активность студента с нормами и группой [9, с. 300]. LARC также анализировала эмоциональный тон форумов, но в целом использовала шаблонный подход. Другие системы (Civitas, Knewton) предоставляли сводки и визуализации, но не формировали персонализированные тексты [9, с. 250].

Позже начали применяться методы машинного обучения. Так, Kobashi и Mine (2021) предложили Seq2Seq-модель с вниманием для генерации предложений отчетов на основе ключевых слов, заданных преподавателем [5, с. 713]. Этот подход превзошёл IR-методы по метрикам BLEU и ROUGE, но ограничен: генерирует только отдельные фразы и требует ручного ввода ключевых слов.

Современные языковые модели в EDM

С развитием LLM, особенно GPT-серии, стало возможным генерировать цельные тексты на сложных данных [2]. GPT-2, обученная на миллионах веб-страниц, продемонстрировала высокую связность текста, а её преемники (GPT-3/3.5/4) усилили этот эффект [2]. В образовании LLM применяются для генерации обратной связи, оценки ответов и резюмирования. Например, Zhou et al. (2024) использовали ChatGPT и систему тегов для генерации персонализированных отзывов по данным об обучении [3]. Учителя подтвердили качество таких отчетов, хотя подход требует сложной аннотации и работы с API.

Другие модели, вроде BERT, не генерируют тексты, но применяются в анализе: классификация отзывов, оценка эссе, анализ высказываний студентов. Например, специализированный BERT показал лучшую классификацию речевых актов, чем ChatGPT, но уступал в интерпретируемости [4]. Это подчеркивает: LLM обеспечивают гибкость и выразительность, в то время как специализированные модели — точность.

Сильные и слабые стороны решений

Шаблонные системы обеспечивают контролируемость, но страдают однообразием. Модели Seq2Seq дают более естественные формулировки, но ограничены в контексте и требуют ручного ввода. Большие модели GPT обладают богатым языком, могут учитывать широкий контекст и адаптироваться к стилю, но нуждаются в ресурсах и контроле. Без дообучения возможны галлюцинации и ошибки. Поэтому перспективным подходом является fine-tuning GPT на корпусе успеваемости студентов. Далее в работе рассматривается реализация такого подхода на базе GPT-2 и сравнение с другими методами.

Материалы и методы

Архитектура GPT-2

GPT-2 — это генеративная трансформер-модель, состоящая из $N \times N$ декодерных блоков с маскированным многоголовым вниманием и MLP-слоем, а также residual-соединениями, нормализацией и dropout [6, с. 125; 7, с. 320]. На вход подаются токены с позиционными эмбедингами; маскирование внимания делает генерацию авто-регрессивной. Вариант GPT-2 XL включает 48 блоков и ~1.5 млрд параметров, доступны и облегченные версии (117M–774M), размещенные в открытом доступе (например, HuggingFace) [4, 6, с. 276]. Для нашей задачи использована версия с ~774M параметров, дообученная на корпусе учебных данных, чтобы адаптировать её к стилю и содержанию образовательных отчетов.

Каждый блок включает слой многоголового маскированного самовнимания (обозначен красным) и позиционное сложение с входом, а также полносвязный слой (синий) для обработки выходов внимания. Нормализация (зелёный) и dropout применяются на входе и выходе каждого слоя. Справа показана схема расчёта внимания: входные представления разделяются между несколькими головами внимания, внутри каждой вычисляются Query, Key и Value векторы, по которым через матричные операции (показаны как matmul и softmax) определяются веса внимания к разным позициям последовательности [10].

Корпус данных

Для fine-tuning был подготовлен корпус из ~5,000 пар «учебные данные — текст отчета». Исходные дан-

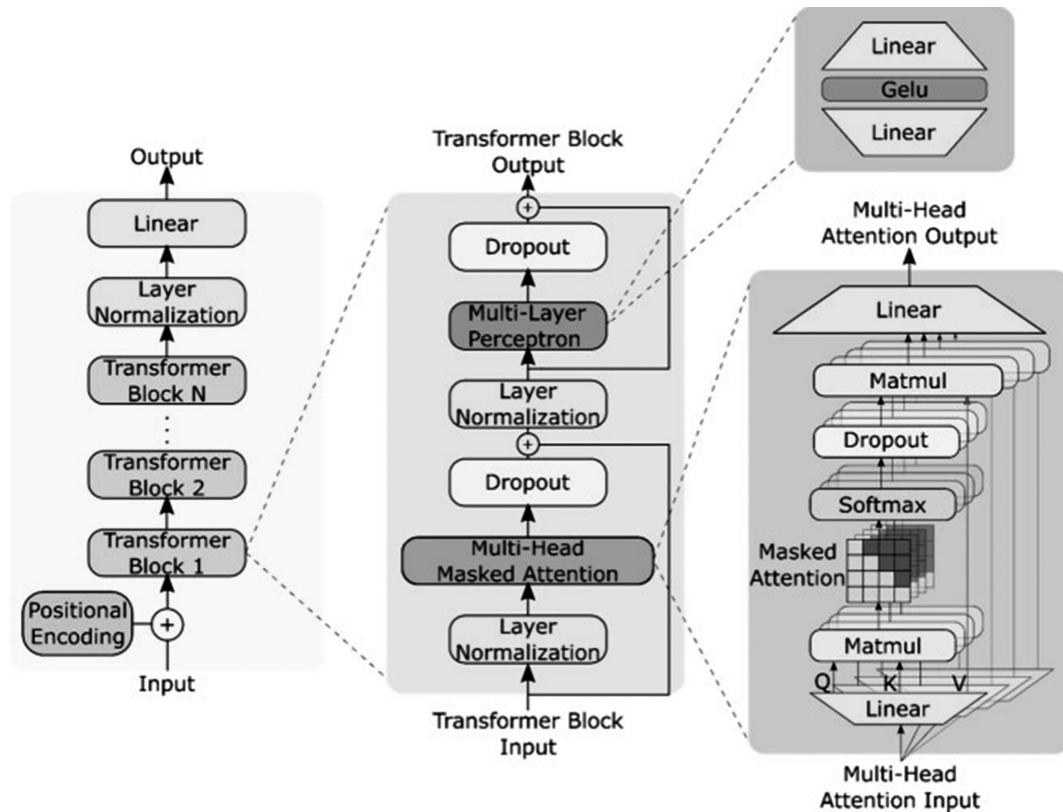


Рис. 1: Архитектура модели GPT 2, состоящая из последовательности декодерных блоков трансформера

ные обезличены и включают оценки, посещаемость, активность, а также комментарии преподавателей. Из них формировались обучающие примеры в виде структурированного ввода и целевого отчета, например:

Ввод: «Студент: А; Курс: Математика; Балл: 82...»

Вывод: «Студент А демонстрирует уверенные знания...»

Данные нормализованы, переведены в текстовый формат, обезличены.

Процесс дообучения

Модель обучалась в среде PyTorch с использованием Transformers, на GPU NVIDIA Tesla (24 Гб). Применён AdamW с шагом $5 \cdot 10^{-5}$ и стратегией linear decay. Обучение длилось 3 эпохи (остановлено после 2 из-за plateau по val loss), заняло ~5 часов. Вход: описание данных + маркер «Отчет»; генерация продолжения до токена.

Техническая реализация

Код реализован на Python (PyTorch 1.13, Transformers 4.x, datasets, Accelerate), обработка данных — через pandas. Все эксперименты выполнялись на Linux-сервере. Репозиторий с кодом и примерами будет опубликован для воспроизводимости [7, с. 110].

Итог:

- 1) выбрана и описана архитектура GPT-2,
- 2) подготовлен корпус данных,
- 3) выполнен fine-tuning,
- 4) подготовлена инфраструктура для оценки качества, что рассматривается далее.

Результаты и обсуждение

После дообучения модель GPT-2 способна на основе описания данных студента автоматически генерировать связный текст отчета. Ниже приведены примеры реальных и сгенерированных отчетов, иллюстрирующие работу модели до и после ее применения.

Пример 1. Входные данные (фрагмент): «Студент В, Курс: Физика, Итоговый балл: 67/100 (оценка С); Тест 1: 80 %; Тест 2: 60 %; Лабы выполнены: 5 из 8; Посещаемость: 70 %.»

- Без модели (шаблонный отчет): «Студент В имеет удовлетворительную успеваемость по курсу физики. Итоговая оценка — С. Результаты первых контрольных работ хорошие, вторых — ниже. Посещаемость оставляет желать лучшего. Требуется подтянуть знания.»
- Сгенерировано GPT-2: «Студент В демонстрирует средние результаты по физике. Итоговый балл составляет 67 из 100, что соответствует оценке С.

При этом успеваемость была неравномерной: первый тест сдан довольно успешно (80 %), тогда как во втором тесте результат снизился до 60 %. Лабораторные работы выполнены частично (5 из 8), что сказалось на общем балле. Посещаемость также невысокая (70 %). Рекомендуется восполнить пробелы по темам, вызвавшим трудности во втором тесте, и завершить оставшиеся лабораторные работы для улучшения понимания материала.»

В приведенном примере видно, что шаблонный (условно «до модели») отчет получился довольно общим и кратким. В то же время отчет, сгенерированный моделью GPT-2, более детально отражает данные: указаны конкретные проценты тестов, упомянут прогресс по лабораторным, даны рекомендации, контекстуально привязанные к выявленным проблемам (низкий результат второго теста). Язык генерированного отчета близок к академическому стилю преподавателя: соблюдается нейтральный тон, имеются связующие фразы («при этом», «что сказалось на...»). Модель явно переняла из обучающего корпуса шаблоны построения отзывов, например: упоминание позитивных аспектов (успех в первом тесте), затем указание недостатков (снижение результата, невыполненные задания) и заключение с советом.

Пример 2. Входные данные: «Студент Г, Курс: История, Итоговый балл: 92/100; Контрольные: 95 %, 88 %. Выполнил все 3 эссе; Посещаемость: 100 %; Активность: 15 сообщений; Средний результат по группе: 85.»

- Отчет GPT-2: «Студент Г показывает отличные результаты по курсу истории. Итоговый балл — 92 из 100, что выше среднего по группе (85). Студент стабильно учится: контрольные работы выполнены очень хорошо (95 % и 88 %). Все три обязательных эссе сданы в срок и получили высокие оценки. Посещаемость примерная (100 %), активность на семинарах и в форуме высокая (15 содержательных сообщений за курс). Студент проявляет глубокий интерес к предмету и активно участвует в обсуждениях. Рекомендуется продолжать в том же духе и при возможности углублять знания через дополнительные материалы — это поможет сохранить высокий уровень и в дальнейшем.»

Здесь отчет полностью сформирован моделью на основе данных. Мы видим, что модель не только перечисляет показатели, но и интерпретирует их: сравнивает итоговый балл со средним по группе, отмечает стабильность (высокие результаты по всем видам работ), хвалит инициативность студента. Таким образом, GPT-2 может производить целостный анализ, а не просто перечисление цифр. Это достигается за счет примеров в обучаю-

щем корпусе, где такие сравнения и выводы присутствовали (например, преподаватели часто указывают, выше или ниже среднего результат студента, выделяют активность и т.п.).

Для объективной оценки качества сгенерированных моделью отчетов мы использовали несколько подходов: автоматические метрики совпадения с эталоном и экспертную оценку. Каждый сгенерированный моделью отчет сравнивался с референтным (целевым) отчетом из тестового набора данных — то есть с тем текстом, который действительно был в обучающих данных для данного студента (или составлен человеком). Были вычислены метрики, широко применяемые в задачах резюмирования и NLG: BLEU (BiLingual Evaluation Understudy) и ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [9, с. 450]. BLEU измеряет сходство генерированного и эталонного текста по перекрытию n-грамм, а ROUGE-N — по перекрытию подстрок длины N. Кроме того, считался ROUGE-L (Longest Common Subsequence) — длина наибольшей общей подпоследовательности, отражающая покрытие смысла. На рис. 2 представлено сравнительное качество отчетов, генерированных моделью GPT-2, по сравнению с более простым шаблонным подходом.

Видно, что модель GPT-2 существенно превосходит шаблон по всем метрикам: так, BLEU (совпадение фраз с эталоном) увеличивается с ~0.55 до ~0.72, а ROUGE-1 (совпадение униграмм) — с ~0.60 до ~0.80. Пост ROUGE-2 и ROUGE-L также указывает на более полное и точное воспроизведение моделью важных фрагментов отчета.

Дообученная GPT-2 показала высокое качество генерации — BLEU≈0.72 (против 0.5–0.6 у шаблонов) указывает на сильное лексическое сходство с эталонными отчетами. Метрики ROUGE-1 и ROUGE-2 подтвердили охват ключевых фактов. При этом совпадение формулировок не обязательно: модель часто выражает идеи иначе, но с сохранением смысла. В ручной проверке ~50 отчетов ≈80 % содержали все важные данные без искажений. Ошибки касались лишь пропусков второстепенных деталей или общей формулировки; галлюцинации не выявлены.

В экспертной оценке (5 преподавателей, по 20 отчетов) средняя точность составила 4.8/5, ясность — 4.6/5, полнота — 4.4/5, полезность — 4.7/5. Преподаватели отметили, что охотно пользовались бы моделью для черновиков. В большинстве случаев они не отличали отчеты GPT-2 от написанных вручную.

Среди ограничений — зависимость качества от обучающего корпуса: если в данных игнорируется важный показатель, модель также может его опускать. Контекстное окно GPT-2 ограничено ~1024 токенами, что может

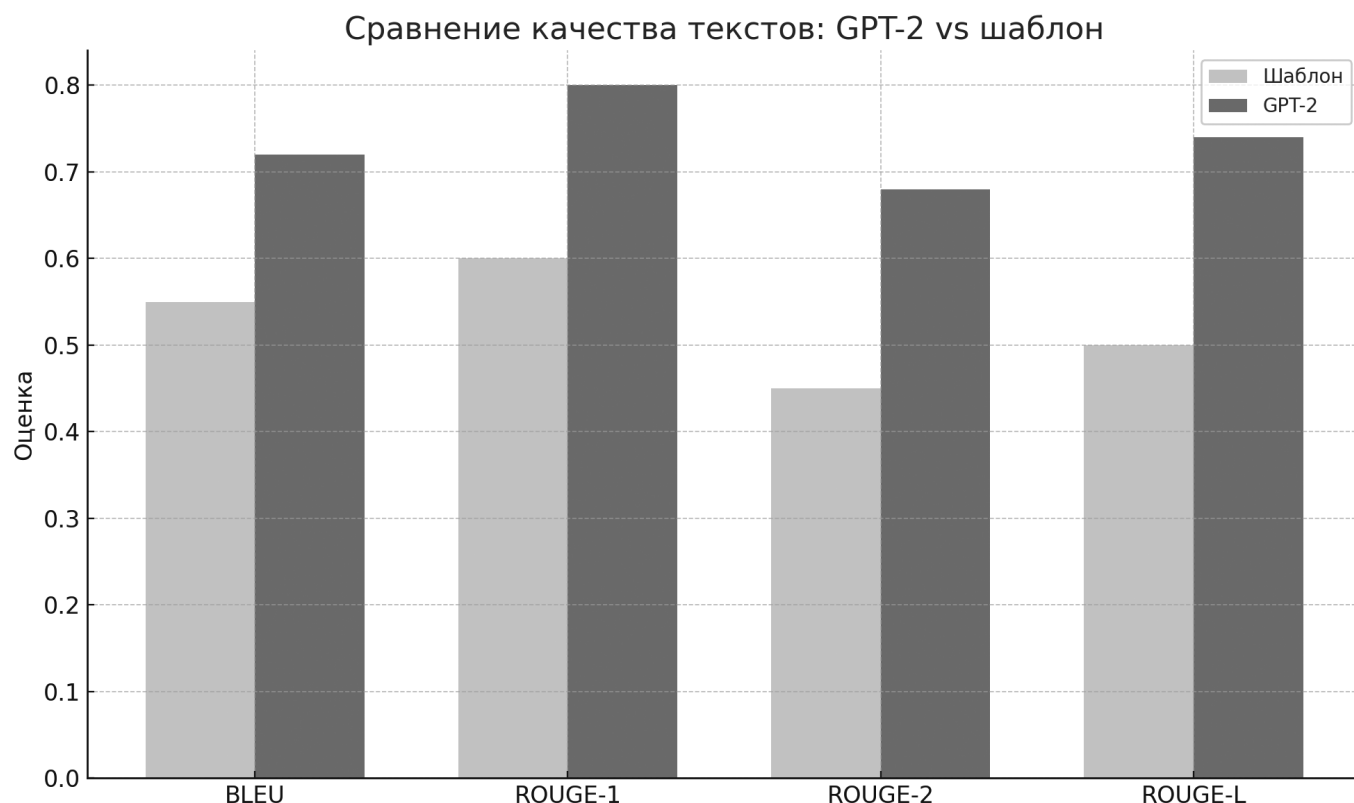


Рис. 2. Сравнение средних значений метрик качества текста для нашей дообученной модели GPT-2 (темно-серые столбцы) и для базового шаблонного метода генерации отчета (светло-серые столбцы)

быть недостаточно для длинных вводов; возможны решения — укрупнение модели (например, GPT-3) или сжатие информации. Модель не всегда улавливает причины событий (например, неявку на экзамен), если они не указаны явно. Кроме того, GPT-2 обучена на русском — для других языков потребуется перевод или дообучение многоязычных моделей. Качество генерации на русском ниже, чем на английском, из-за меньшего объема предобученных данных.

Отдельное преимущество GPT-2 — автономность и прозрачность: её можно запускать локально, не передавая данные сторонним сервисам, в отличие от API GPT-3.5/4. Хотя те показывают более разнообразный язык и логику, результаты GPT-2 уже близки к преподавательскому стилю, что делает её эффективным решением для образовательной аналитики.

Заключение

В работе продемонстрирована реализуемость и эффективность генерации аналитических отчетов об успеваемости студентов с помощью дообученной модели GPT-2. Разработан прототип, преобразующий сырые учебные данные в связный текст, близкий по стилю к отчетам преподавателей. Создан специализированный корпус, включающий показатели студентов и текстовые комментарии, на котором модель была успешно дообучена с учётом технических аспектов.

Результаты показали, что GPT-2 способна генерировать содержательные отчёты с рекомендациями и выводами. Оценки BLEU, ROUGE и экспертный анализ подтвердили близость текста к человеческому. Это существенно снижает трудозатраты преподавателей при работе с большими группами — отчёты можно редактировать вместо написания с нуля. Студенты при этом получают своевременную и персонализированную обратную связь.

Работа подтверждает, что fine-tuning позволяет адаптировать LLM к конкретной образовательной задаче. Такой подход можно масштабировать на другие предметные области (например, программирование или языковые курсы).

В дальнейшем возможно:

- протестировать более современные модели (GPT-3, Bloom и др.);
- внедрить механизмы контроля качества (например, проверку чисел и логики отчёта);
- реализовать обучение с обратной связью от преподавателей (on-the-fly fine-tuning);
- сочетать текстовую генерацию с визуальной аналитикой (диаграммы и пояснения).

Важно и педагогическое измерение: как отчёты влияют на мотивацию и успеваемость. Первичные отзывы

преподавателей позитивны, но для полной картины необходимы масштабные исследования в реальных курсах. Эта работа закладывает техническую основу и от-

крывает путь к персонализированной образовательной аналитике на базе ИИ.

ЛИТЕРАТУРА

1. Educational Data Mining [Электронный ресурс]. — URL: <https://educationaldatamining.org> (дата обращения: 01.06.2025).
2. HuggingFace Transformers Documentation [Электронный ресурс]. — URL: <https://huggingface.co> (дата обращения: 01.06.2025).
3. ACL Anthology [Электронный ресурс]. — URL: <https://aclanthology.org> (дата обращения: 01.06.2025).
4. Medium — HuggingFace Transformers Guide [Электронный ресурс]. — URL: <https://medium.com> (дата обращения: 01.06.2025).
5. Рассел С., Норвиг П. Искусственный интеллект: современный подход. — М.: Вильямс, 2021. — 1416 с.
6. Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение. — М.: ДМК Пресс, 2018. — 652 с.
7. Марр Б. Аналитика данных. Как извлекать максимум из больших данных. — М.: Манн, Иванов и Фербер, 2017. — 288 с.
8. ЛеКун Я., Хинтон Дж., Бенджио Й. Глубокое обучение (статья в журнале Nature). — Nature, 2015. — Vol. 521. — P. 436–444.
9. Дьяконов А.Г. Современные методы анализа данных. — СПб.: Питер, 2019. — 592 с.
10. PMC NCBI Articles — GPT Architecture [Электронный ресурс]. — URL: <https://pmc.ncbi.nlm.nih.gov> (дата обращения: 01.06.2025).

© Максимов Денис Романович (maksimov.official@mail.ru); Черваков Филипп Сергеевич (pschervakov@gmail.com)
Журнал «Современная наука: актуальные проблемы теории и практики»