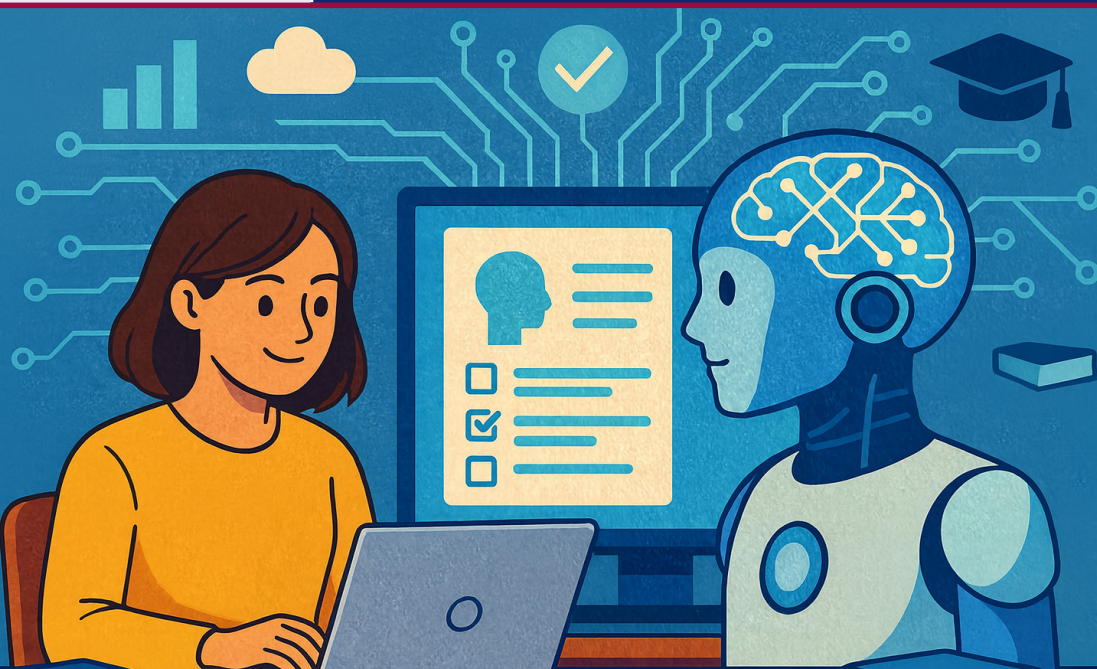
ISSN 2500-0608

НОВЫЕ ПОДХОДЫ К ОЦЕНИВАНИЮ: ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ КАК ДРАЙВЕР ИЗМЕНЕНИЙ В ОБРАЗОВАНИИ

Под научной редакцией Е.Ю. Кардановой

Современная аналитика образования
№ 5 (88)
2025



ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

ИНСТИТУТ ОБРАЗОВАНИЯ

НОВЫЕ ПОДХОДЫ К ОЦЕНИВАНИЮ: ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ КАК ДРАЙВЕР ИЗМЕНЕНИЙ В ОБРАЗОВАНИИ

Серия
Современная аналитика образования

№ 5 (88)
2025

УДК 371.39:004.8
ББК 32.813
Н 76

Н 76 **Новые** подходы к оцениванию: искусственный интеллект как драйвер изменений в образовании / Е. Ю. Карданова (научная редакция), С. В. Тарасов, А. Е. Иванова, Э. М. Юсупова, Д. А. Грачева, К. В. Тарасова, И. С. Денисов, Д. П. Талов, А. С. Струкова ; Национальный исследовательский университет «Высшая школа экономики», Институт образования. — М.: НИУ ВШЭ, 2025. —88 с. — 100 экз. — (Современная аналитика образования. № 5 (88)).

Искусственный интеллект (ИИ) — удобный инструмент, открывающий новые возможности в сфере образовательного оценивания. Авторы рассматривают применение ИИ на всех ключевых этапах оценивания: автоматическая разработка заданий, проверка работ, предоставление персонализированной обратной связи и анализ результатов. Особое внимание уделено этическим вопросам, связанным с прозрачностью, предвзятостью, конфиденциальностью и ограничениями использования ИИ в оценивании.

В выпуске приводятся практические примеры и современные исследования, что делает его полезным ресурсом для специалистов в области образования, психометрики и цифровых технологий.

Публикация подготовлена в рамках гранта, предоставленного Российским научным фондом (№ 25-18-00751).

Содержание

Глоссарий и используемые сокращения

Искусственный интеллект (Artificial Intelligence, далее — **ИИ**) — совокупность технологий, способных выполнять задачи, обычно требующие человеческого интеллекта, такие как распознавание образов, обучение и принятие решений.

Машинное обучение (Machine Learning, далее — **ML**) — раздел ИИ, направленный на создание алгоритмов и статистических моделей, способных выявлять закономерности и паттерны в данных, находить связи, делать прогнозы и так далее, с минимальным вмешательством человека.

Обработка естественного языка (Natural Language Processing, далее — **NLP**) — технология машинного обучения, позволяющая компьютерным системам распознавать, понимать, генерировать и обрабатывать человеческую речь в устном и письменном виде.

Обучение (Training) — процесс адаптации модели машинного обучения к решению конкретной задачи на основе анализа примеров или выборки данных. Включает в себя различные методы, такие как *предварительное обучение (Pre-training)*, *тонкая настройка или дообучение (Fine-tuning)*, *обучение с учителем (Supervised Learning)*, *обучение без учителя (Unsupervised Learning)* и *обучение с подкреплением (Reinforcement Learning)*.

Нейронная сеть (Neural Network, далее — **NN**) — архитектура машинного обучения, имитирующая работу нервных клеток в человеческом мозге. Выделяют особые типы нейронных сетей, специализирующихся на распознавании, генерации и переводе естественного языка — *большие языковые модели* (Large Language Models, далее — **LLM**), а также на распознавании изображений — *сверточные нейронные сети* (Convolutional Neural Networks, далее — **CNN**).

Промптинг (prompt engineering) — процесс взаимодействия с ИИ для достижения желаемого результата. Взаимодействие происходит с помощью предъявления определенных инструкций или запросов (*промптов*), на основе которых ИИ генерирует ответ.

Введение

За последние десятилетия искусственный интеллект претерпел значительные изменения, и современный его уровень — это результат нескольких последовательных прорывов. Начиналось все с простых алгоритмических моделей, таких как линейная или логистическая регрессия. Сейчас же ИИ у большинства людей ассоциируются с ChatGPT или GigaChat, которые являются большими языковыми моделями, основанными на моделях трансформеров[1].

Области применения ИИ с каждым годом расширяются. Его применяют в промышленном производстве, медицине, экономике, образовании и даже в искусстве. В условиях роста объемов данных, вычислительных возможностей компьютерных технологий и развития архитектур моделей ИИ оказывает существенное влияние на целый ряд областей в образовании, таких как адаптивное обучение, рекомендательные системы, процессы оценивания, разработка заданий, — и это только часть списка.

Предлагаемый материал посвящен рассмотрению роли ИИ в оценивании и тому, как современные технологии помогают решать ключевые задачи в данной области. Мы разделили оценивание на этапы, относительно которых рассмотрели возможности использования ИИ. Каждый этап оценивания анализируется с точки зрения, во-первых, возможностей ИИ с практическими примерами из современных исследований, а во-вторых, вызовов, связанных с его применением в образовании и диагностике.

В первой главе основное внимание уделяется разработке тестовых заданий, включая генерацию стимульного материала и самих заданий с использованием ИИ. Мы анализируем, как большие языковые модели и другие методы машинного обучения могут автоматизировать процесс создания тестов, обеспечивая соответствие заданий образовательным целям и психометрическим требованиям.

Во второй главе рассматривается автоматическая оценка различных типов ответов: текстовых, графических и аудиовизуальных. ИИ

[1] Под моделями трансформеров понимается архитектура NN, которая позволяет обрабатывать длинные последовательности данных с помощью параллельных вычислений, тем самым позволяя NN понимать смысл и контекст.

предлагает решения для такой оценки, снижая человеческий фактор в оценивании и обеспечивая более объективную оценку как структурированных, так и креативных ответов.

Третья глава посвящена предоставлению обратной связи. Здесь мы обсуждаем, как ИИ помогает адаптировать обратную связь под потребности учащихся, обеспечивая своевременное и детализированное сопровождение обучения.

В четвертой главе освещены вопросы анализа данных тестирования, включая как количественные, так и качественные методы. ИИ помогает обрабатывать большие массивы данных, выявлять аномалии и предлагать новые подходы к интерпретации результатов тестирования.

Наконец, в пятой главе рассматриваются этические аспекты использования ИИ в оценивании. Мы затрагиваем такие важные вопросы, как прозрачность алгоритмов, предвзятость, конфиденциальность данных, а также влияние ИИ на учебные процессы и взаимодействие между учащимися и преподавателями. Этические аспекты играют ключевую роль в успешном и ответственном внедрении ИИ, и важность их понимания подчеркивается на каждом этапе работы с данными и оценивания.

1. Разработка заданий

Автоматизация работы по генерации заданий и стимульного материала к ним обусловлена необходимостью разработки большого количества заданий в рамках крупных мониторинговых исследований и снижения затрат на их создание, а также желанием разработчиков расширить рамки привычных инструментов оценивания, дополнив их мультимодальными стимулами, такими как изображения или аудио. Разработки в этом направлении велись уже давно. Однако ранние модели машинного и глубокого обучения не могли эффективно работать с длинными последовательностями слов и сохранять контекст, что ограничивало их способность генерировать осмысленные тексты и задания. Появление моделей-трансформеров кардинально изменило технологии обработки естественного языка, позволив сохранять контекст между длинными последовательностями слов. Это стало настоящим прорывом в области автоматической генерации заданий и стимульного материала. В данном разделе мы представим примеры применения ИИ для автоматической генерации заданий и стимульного материала к заданиям и рассмотрим некоторые связанные с этим проблемы.

1.1. Генерация стимульного материала

Технологии генерации текста значительно продвинулись за последние годы, особенно благодаря появлению глубокого обучения и больших языковых моделей (LLM). Некоторые из наиболее важных моделей, которые сыграли ключевую роль в развитии современных методов генерации текстов, включают модели Word2Vec[2], GloVe, BERT и GPTx.

[2] Word2Vec — алгоритм для векторного представления слов на естественном языке (с открытым кодом) с использованием нейронных сетей; был опубликован группой исследователей Google в 2013 г. [Mikolov et al., 2013]. GloVe — еще один алгоритм векторизации, разработанный группой ученых из Стенфорда примерно в то же время, также с открытым кодом; основан на факторизации матрицы частот совместной встречаемости слов [Pennington, et al., 2014]. BERT — трансформер-модель, обучаемая методом двунаправленного (bidirectional) предсказания пропущенных слов (Masked Language Model, MLM); впервые опубликована исследователями из группы Google AI в 2018 г. [Devlin et al., 2018]; код доступен в репозитории компании. GPTx — серия моделей генеративных трансформеров

Слово «большие» в названии LLM отсылает нас к тем огромным наборам данных, на которых они обучены, и множеству параметров, которые они содержат. Так, по оценкам экспертов, модель GPT-4 содержит около 1.8 триллионов параметров, что почти в 10 раз больше ее предшественницы GPT-3[3]. А чтобы обучить последнюю, потребовалось 570 гигабайтов очищенных текстовых данных, что составляет примерно 300 миллиардов токенов [Brown et al., 2020]. Пожалуй, именно этот эффект масштаба позволяет моделям понимать тонкие нюансы и особенности человеческого языка, необходимого для создания текстов. Современные LLM созданы для того, чтобы понимать и создавать текст, который будет не только логичным и релевантным в нужном контексте, но и похожим на созданный человеком [Lancaster, 2023].

И все же некоторые авторы отмечают, что, несмотря на огромное количество исследований в области ИИ и его применения для разных целей, в том числе в образовании, конкретные темы, такие как генерация текстов для целей оценивания, все еще имеют ряд белых пятен [Becker et al., 2024; von Davier, 2018]. И несмотря на наличие вполне успешных примеров использования ИИ и конкретно LLM для таких целей, все еще слишком малочисленны реальные кейсы применения этих моделей для создания текстов, специально разработанных, скажем, для внутриклассного оценивания или крупномасштабных мониторингов на конкретном уровне образования, например, определенного класса школы.

В недавнем исследовании компания Duolingo English Test (DET) использовала GPT-3 для генерации информационных и художественных текстов для студентов вузов [Attali et al., 2022]. Она также применяла GPT-3 для создания вопросов с множественным выбором и ответных опций к ним. Исследователи из Duolingo использовали подход «человек в цикле ИИ» (human in the loop) на протяжении всего процесса разработки теста для обеспечения безопасности тестов, генерации тестовых заданий и автоматической оценки ответов тестируемых [Burstein et al.,

(GPT-1, GPT-2, GPT-3, GPT-4), основанных на авторегрессионном обучении; первая версия (GPT-1) была представлена исследователями из компании OpenAI в 2018 г. [Radford et al., 2018], код первой версии также открыт.

[3] OpenAi официально не публикует сведений о точном наборе и количестве параметров GPT-4. Но оценки различных экспертов можно найти в сети. См., например, https://explodingtopics.com/blog/gpt-parameters?utm_source=chatgpt.com.

2021]. Под «человеком в цикле ИИ» понимается процесс, включающий оценку сгенерированных текстов исследователями и разработчиками тестов, которые тщательно проверяют новые типы заданий, например, оценивая справедливость и возможные искажения в автоматически сгенерированных текстах и заданиях к ним.

Схожими результатами о попытках генерации текстов делится проект Gwern, где авторы активно экспериментируют с GPT-3 для написания художественных и научно-популярных текстов [Gwern, 2023]. Автор проекта показывает, что GPT-3 смог создать связные и креативные истории, но при генерации длинных текстов результаты иногда становились повторяющимися или менее оригинальными. В генерации же научно-популярных текстов GPT-3 показал хорошие результаты в создании кратких обобщений научных статей, а также полностью уникальных текстов по заданным инструкциям. Однако ввиду большой вероятности ошибок в техническом или специализированном контенте в работе с моделью требовалась вмешательство человека по принципу «человек в цикле ИИ».

Упомянутые исследования в основном фокусировались на уровне высшего образования или, более широко, на оценке взрослых; примером другой интересной работы является исследование Bezirhan и von Davier (2023), ориентированное на школьный уровень, где авторы применили LLM для генерации текстов для оценки чтения учеников 4 класса, участвующих в международном исследовании PIRLS. В этой работе подробно и точно описан процесс генерации текста для целей оценивания, и здесь мы позволим себе привести его краткое изложение, описывающее, как разворачивалась работа исследователей и разработчиков текстов.

Авторы составили набор из 24 текстов, использованных в исследовании PIRLS в период с 2001 по 2016 год, для использования их в промптах при генерации текстов с помощью LLM. Далее они иллюстрировали свою работу на основе двух информационных текстов («Антарктида: земля льда» и «Муравьи») и одного художественного текста («Храбрая Шарлотта»). Авторы пытались понять, можно ли создать тексты уровня PIRLS на базе LLM. Для этого они использовали модель text-davinci-002 от OpenAI, а также Python для отправки API-запросов.

Очевидно, что дизайн промптов имеет ключевое значение для больших языковых моделей, таких как GPT, поскольку он напрямую

влияет на качество и релевантность генерируемых текстов. В своем исследовании авторы испробовали несколько версий промптов: без обучения, с примером, с четкой инструкцией; а также включили информацию о возрасте целевой аудитории, чтобы обеспечить соответствие текста возрасту учеников 4 класса. Вместе с промптами исследователи также манипулировали параметром температуры[4], генерируя истории с температурами от 0.5 до 0.9, чтобы получить разные варианты текстов на выходе.

В итоге на базе указанных выше промптов и параметров были созданы информационные и художественные тексты для PIRLS, по десять повторений для каждого набора настроек. После генерации текстов механизм отбора включал в себя вычисление сложности текста с помощью онлайн-анализатора сложности текста [Cathoven, 2023], чтобы измерить уровень сложности чтения, включая частоту слов, длину предложений и другие параметры[5]. После вычисления показателей сложности текста — как для оригинальных текстов PIRLS, так и для сгенерированных — для дальнейшей работы отбирались только те сгенерированные тексты, которые попали в диапазон одного стандартного отклонения от исходных текстов по показателю сложности. Наконец, после такого «технического» отбора человек-редактор финально проверял полученные и отобранные тексты с целью исправления в них грамматических и фактических ошибок. Итоговые тексты, а также итоговые промпты приведены в указанном исследовании и доступны для ознакомления [Bezirhan, von Davier, 2023].

Разумеется, тексты не единственный вид стимульного материала, который может производиться с применением ИИ. Один из самых позитивных моментов использования ИИ для создания стимульного материала для тестов и опросов — это возможности мультимодальности. Генерация мультимодальных стимулов, таких как изображения или аудио вместо или вместе с текстами, позволяет интегрировать в кон-

[4] Температура — параметр модели ИИ, который позволяет управлять случайностью и разнообразием генерируемого материала. При низких значениях температуры ИИ выбирает слова с более высокой вероятностью, что приводит к более предсказуемому и как правило консервативному результату. При более высокой температуре ИИ рассматривает более широкий диапазон слов, что ведет к более случайному и более креативному результату.

[5] В случае использования для России альтернативой может быть, например, текстометр [Лапошина, Лебедева, 2021].

тент тестов сразу несколько способов коммуникации с тестируемым [Bulut et al., 2024]. Даже традиционные технологии мультимодального оценивания способствуют внедрению принципов универсального дизайна благодаря множеству способов взаимодействия, представления и (само)выражения [Rao, 2015]; ИИ же обладает колоссальными возможностями для этого. Использование аудио, видео, графики позволяет сделать оценивание не только разнообразным и увлекательным, но и доступным для всех учащихся, снижая барьеры и расширяя возможности для каждого тестируемого продемонстрировать свои истинные способности.

Использование генеративного ИИ для создания мультимодальных стимульных материалов может успешно дополнять более «простые» традиционные текстовые задания. Например, в тесте на понимание языка ИИ может сгенерировать короткий рассказ в виде аудиозаписи, сопровождаемой визуальными материалами, изображающими ключевые сцены или концепции. Такой подход не только способствует лучшему пониманию, но и активизирует различные когнитивные навыки, предлагая более всестороннюю оценку способностей учащегося по сравнению с традиционными тестами [Bulut et al., 2024]. Возможность учащихся взаимодействовать с контентом более осмысленными способами способствует более глубокому измерению их критического мышления, аналитических и интерпретационных навыков [Sharma, Giannakos, 2020].

В сфере образования изучается использование ИИ в приложениях дополненной и виртуальной реальности (AR и VR), которые позволяют имитировать ситуацию реальной деятельности и оценивать действия человека. Такие приложения использовались и ранее, но ИИ разрешает сделать виртуальную среду более сложной и реалистичной за счет быстрой генерации необходимого стимульного материала [Al Balushi et al., 2024]. Недавние исследования показывают, что возможности ИИ для создания адаптивных и погружающих сред для образовательной оценки значительно усиливают процесс тестирования и обучения [Al-Ansi et al., 2023; Rock Paper Reality, 2024; Strivr Labs, 2024]. Например, погружение в обучение с использованием AR и VR усиливается благодаря генеративным ИИ-инструментам, которые позволяют динамически создавать контент, такой как интерактивные сценарии и реакции персонажей на действия пользователя. Этот подход особенно полезен для оценки сложных навыков, таких как со-

трудничество, решение проблем и адаптивность, поскольку ИИ может в режиме реального времени адаптировать VR-среду в зависимости от действий участника, создавая более реалистичное и глубокое оценочное взаимодействие [Al-Ansi et al., 2023].

Еще одной областью применения является оценка социальных навыков, где VR-среды, управляемые ИИ, симулируют командные взаимодействия с виртуальными агентами. Эти агенты могут изменять свое поведение в зависимости от действий участника, позволяя точно оценить как когнитивные, так и эмоциональные реакции в рамках коллективных заданий. Такая гибкость не только обеспечивает получение надежных данных, но и позволяет фиксировать тонкие нюансы человеческих взаимодействий, которые трудно наблюдать в традиционных условиях [Burgues et al., 2024].

Кроме того, VR, дополненный генеративными возможностями ИИ, поддерживает контекстное обучение, позволяя участникам исследовать, взаимодействовать и ошибаться в безопасной, но реалистичной среде [Nocera et al., 2023]. Этот подход эффективен для оценки технических и профессиональных навыков, когда погружающие сценарии могут имитировать реальные профессиональные задачи. Совмещение VR и ИИ для создания интерактивности становится перспективным инструментом для повышения точности оценки [Estejab, Bayramzadeh, 2025] и поддержания инклюзивности в образовании [Salas-Pilco et al.,2022].

Источников, которые бы описывали применение и оценку валидности заданий, созданных на основе ИИ с использованием AR и VR, пока не удалось найти. Но есть несколько кейсов применения AR- и VR-технологий, которые, как представляется, интересно рассмотреть. Один из них — это приложение, которое оценивает уровень исполнительных функций, в частности, способности к многозадачности, через действия человека в виртуально созданном пространстве квартиры. Это приложение призвано решить проблему ранней диагностики деменции и болезни Альцгеймера: часто пожилые люди успешно справляются с заданиями на память и внимание, но испытывают трудности с решением бытовых задач. Общая инструкция звучит следующим образом: «Вы живете в гостях у своего друга. Пока он на работе, вы выполняете некоторую работу по дому, а затем вечером вы планируете пойти на спектакль». Испытуемый получает несколько заданий: разложить продукты, купить билеты на вечерний спектакль,

покормить рыбок и т.д. Важно удерживать в памяти некоторые обстоятельства, например, что дверь в спальню должна быть закрыта, чтобы собака не залезла на кровать. В выполнение задач врываются непредвиденные обстоятельства: внезапно начинается гроза, и ветер роняет вазу с водой, звонит телефон и т.д. Главное задание, которое требуется выполнить и которое оценивается, — это сортировка продуктов. Оценивается, сколько попыток было сделано для выполнения задания (например, по заданию нужно положить определенный продукт на определенную полку; оценивается, было ли это сделано с первого раза или испытуемый сначала сделал неверно, но исправил ошибку) и общее время выполнения задания. Результаты, полученные в приложении, значимо коррелировали с традиционными заданиями, оценивающими исполнительные функции, так что авторы приходят к выводу, что приложение может достоверно оценивать функциональные способности при старении [Banville et al., 2018].

В других подобных приложениях испытуемым предлагают парковать машину или навести порядок в химической лаборатории [Davison et al., 2018] и др. Такие приложения повышают вовлеченность в выполнение заданий, мотивацию, интерес и удовольствие от активности [Al Balushi et al., 2024], снижают риск предвзятости [Davis, 2019]. Однако необходимы дальнейшие исследования, направленные на подтверждение валидности такого рода заданий, поскольку пока не накоплено достаточное количество свидетельств [Kirkham et al., 2024]. Существенным ограничением таких приложений является высокая, по сравнению с традиционными опросниками, стоимость необходимого оборудования.

1.2. Генерация заданий

Автоматическая генерация заданий — стремительно развивающееся направление в тестировании, появившееся в начале двухтысячных и ставшее ответом на необходимость разработки большого количества заданий в рамках крупных мониторинговых исследований, а также снижения затрат на их создание [Gierl, Lai, 2015]. Первоначально новые задания создавались подбором числовых значений переменных из заданного множества для создания «клонов» простых алгебраических задач. Позже одним из самых распространенных способов генерации заданий стало формирование моделей зада-

ний с радикальными и сопроводительными элементами (radicals and incidentals) [Irvine, Kyllonen, 2013]. Предполагается, что радикальные элементы влияют на психометрические свойства заданий, а сопроводительные — не влияют, поэтому радикальные элементы заданий остаются неизменными, а сопроводительные можно настраивать для получения разных заданий с похожими психометрическими характеристиками. Сопроводительные элементы чаще всего включают в себя детали контекста задач, варианты ответа из заранее составленной базы данных и численные показатели в допустимом пределе. Это позволяет получать новые задания путем клонирования уже существующих, с доказанными психометрическими свойствами.

Генерация заданий с помощью моделей заданий может быть эффективной в плане разработки психометрически валидных заданий, однако их разработка требует значительных усилий, поскольку и модель задания, и другие компоненты должны быть написаны вручную. Кроме того, генерируемые задания будут ограничены по контенту и даже трудности, поскольку по сути являются клонами исходного задания и ограничены сложностью написания модели задания [Attali et al., 2022]. Эти недостатки мотивировали исследователей на работы по использованию ИИ в качестве генератора заданий. Von Davier (2018) был одним из первых, кто исследовал это и использовал рекуррентные нейронные сети для генерации заданий для оценки личности.

Однако, ранние модели машинного и глубинного (глубокого) обучения не могли эффективно обрабатывать длинные последовательности слов и сохранять контекст, что ограничивало их способность создавать осмысленные тексты. Модели трансформеров кардинально изменили технологии обработки естественного языка (natural language processing, NLP), дав возможность сохранять контекст между длинными последовательностями слов и тем самым совершив революцию в области NLP. Благодаря моделям трансформеров, стало возможным обучить такие большие языковые модели как GPT, Bert, GigaChat, YandexGPT и другие.

С появлением больших языковых моделей, которые генерируют длинные, связные и информационно насыщенные тексты [Devlin, 2018; Radford et al., 2019; Zhang, Li, 2021], исследования по оценке эффективности LLM для задачи генерации заданий получили новый импульс [см., например, Tan et al., 2024; Bulut, Yildirim-Erbasli, Gorgun, 2024]. Так, в исследовании Küchemann et al. (2023) использовалась

модель ChatGPT 3.5 для сравнения сгенерированных ею заданий по физике для учащихся старшей школы (10 класс) с заданиями из учебника. Было показано, что модель способна генерировать аналоги заданий из учебников физики, в том числе адекватных по трудности; однако при этом была выявлена неоднородность качества сгенерированных моделью заданий (по критериям ясности, правильности, релевантности и другим). Также исследователи отметили, что применение LLM имеет ряд преимуществ по сравнению с классической разработкой заданий, например, за счет предложения более разнообразного контекста.

В другом исследовании, в области математики, сравнивались задания из учебника по математике для колледжей с заданиями, созданными ChatGPT [Bhandari et al., 2023]. Целью исследования была оценка способности заданий, созданных моделью, дифференцировать студентов по уровням способности, а также оценка психометрических характеристик сгенерированных заданий по сравнению с заданиями из учебника. Исследователи провели тестирование выборки студентов, выполнявших задания и из учебника, и сгенерированные моделью, и использовали современную теорию тестирования (Item Response Theory, IRT; [Van der Linden, 2017]) для анализа качества заданий. Результаты показали, что задания, созданные ChatGPT, были сопоставимы с заданиями из учебника как по психометрическим характеристикам, так и по способности дифференцировать студентов.

Эти и многие другие исследования демонстрируют потенциал ИИ и больших языковых моделей, в частности, для автоматической генерации заданий для оценки способностей учащихся по различным учебным областям [см., например, исследования Laverghetta Jr, Licato, 2023; Bezirhan, von Davier, 2023; Rangapur, Rangapur, 2024; Minaee et al., 2024]. Однако в большинстве исследований генерация заданий проводилась без опоры на теорию измерений. В традиционной разработке заданий людьми или автогенерации заданий на основе моделей (без использования LLM) создание заданий начинается с четкого определения того, что должно быть измерено (т. е. определения конструкта, который будет измеряться с учетом учебных целей и ожидаемых результатов обучения), почему это надо измерять (т. е. каковы цели оценивания) и как это надо измерять (т. е. определение дизайна инструмента и формата заданий). В настоящее время есть только несколько исследований, использующих LLM для автогенера-

ции, которые рассматривают в какой-то мере эти важные аспекты [Tan et al., 2024].

Например, одной из задач исследования Doughty и коллег (2024) являлась оценка того, насколько хорошо сгенерированные моделью GPT-4 задания соответствуют заранее определенным целям обучения. Исследователи оценивали качество 1100 сгенерированных моделью заданий по программированию с выбором одного правильного ответа из нескольких предложенных в сравнении с качеством традиционных заданий. Особое внимание уделялось таким факторам, как ясность, единственный правильный ответ, качество дистракторов, синтаксическая/логическая правильность кода, а также соответствие целям обучения. Результаты исследования показали, что модель GPT-4 способна производить задания, которые сформулированы на понятном языке, имеют единственный правильный ответ и качественные дистракторы. Более того, авторы статьи отметили более сильную связь с целями обучения у заданий, сгенерированных моделью, чем у заданий, созданных человеком, что подчеркивает потенциал LLM для создания инструментов оценивания, которые напрямую нацелены на предполагаемые результаты обучения.

Также в обзоре [Tan et al., 2024] отмечается, что в большинстве исследований авторы не пытаются создавать задания, нацеленные на когнитивные процессы более высокого уровня, указанные в таксономии Блума, такие как применение, анализ, оценка, создание. Генерация заданий преимущественно фокусируется на нижних уровнях таксономии — уровнях запоминания и понимания, что не всегда соответствует целям и задачам оценивания и может быть недостаточно для оценки результатов и прогресса в обучении. Одной из первых работ, демонстрирующих возможности больших языковых моделей для генерации заданий более высоких уровней таксономии Блума, является работа «Automatic generation of physics items with Large Language Models (LLMs)» [Omopekunola, Kardanova, 2024], в которой изучаются способности двух моделей ChatGPT (GPT-4) и Gemini в генерации заданий по физике для старшей школы, соответствующих уровню применения таксономии Блума. Авторы использовали различные техники промптинга для генерации заданий на разных когнитивных уровнях. Сгенерированные задания оценивались экспертно с использованием ряда критериев, которые включали ясность формулировки задания, правильное использование языка, отсутствие вводящего в заблужде-

ние контента, подходящую трудность, единственность правильного ответа, а также соответствие предполагаемому уровню таксономии Блума. Результаты исследования показали, что и ChatGPT, и Gemini способны создавать качественные задания по физике, однако их эффективность отличалась в зависимости от используемых методов промптинга. В частности, авторы выделяют варианты промптинга, давшего наилучшие результаты для обеих моделей, и показывают примеры качественных во всех аспектах заданий, которые соответствуют уровню применения таксономии Блума.

Следует отметить, что разработка промптов (запросов к модели) для LLM играет решающую роль в повышении эффективности и качества генерируемых заданий [Brown et al., 2020]. Сам по себе промптинг заключается в построении четкого текстового запроса, содержащего цели и пожелания разработчика [Marvin et al., 2023]. Точность запроса можно регулировать с помощью таких техник, как использование однозначных выражений и соответствующего тона, уточнение ожидаемого формата результата, дополнение примерами, контекстом и важными деталями [Bozkurt, Sharma, 2023; Liu et al., 2023].

Промпты могут принимать форму прямых вопросов или инструкций, и конкретная формулировка промпта формирует реакцию — ответ, генерируемый LLM [Brown et al., 2020]. Самые распространенные вариации промптинга применительно к задаче автогенерации заданий обычно классифицируют следующим образом:

1) Zero-shot prompting, One-shot prompting, Few-shot prompting — варианты промптинга, когда модели дается промпт либо вообще без примеров или контекста — и модель опирается только на те знания, что содержатся в пре-трейне [Zhong et al., 2023; Miao et al., 2024], либо с одним или несколькими примерами с целью объяснить ей, что требуется сделать и как должен выглядеть результат [Song et al., 2023; Polat et al., 2024; Gao et al., 2020; Agarwal et al., 2024];

2) Instructional prompting — вариант промптинга, который предполагает предоставление LLM четких инструкций в виде простого текста [Mishra et al., 2021], что позволяет включить теоретические основы обучения (таксономии, цели обучения и так далее) непосредственно в запрос разработчика, явно выделяя их для языковой модели.

Отметим, что промптинг в процессе генерации заданий с помощью LLM имеет очень важное значение, так как он определяет последующие действия модели и их результат — сгенерированные зада-

ния, которые соответствуют определенным характеристикам и целям оценки, имеют определенный формат и содержание. Помимо этого, различные методы промптинга для LLM [Marvin et al., 2023] дают возможность встроить образовательные теории в процесс генерации (к примеру, в таксономию учебных целей).

1.3. Выводы по главе

Таким образом, на сегодня опубликовано уже много исследований по автоматической генерации заданий с помощью больших языковых моделей в самых разных областях. В обзоре [Tan et al., 2024] обобщен опыт по использованию LLM для автогенерации заданий. Авторы отмечают, что:

а) наиболее часто используемыми моделями для генерации заданий являются различные версии моделей T5, BERT, GPT;

б) LLM способны генерировать задания различных форм — как с выбором одного или нескольких правильных ответов, так и открытые задания с кратким регламентируемым ответом (например, задания на заполнение пропусков), а также задания со свободно конструируемым ответом;

в) наибольшее число исследований демонстрируют результаты генерации заданий в двух областях: знание языка и общие знания; остальные области, такие как математика, физика, биология, химия, программирование и другие, представлены в исследованиях реже;

г) авторы нашли опубликованные работы о генерации заданий с использованием ИИ на двенадцати языках, однако с большим доминированием английского.

Подводя итог, авторы обзора приходят к выводу, что LLM являются гибким и эффективным решением для создания различных типов заданий на разных языках и в разных предметных областях.

Необходимо отметить, что большинство из опубликованных исследований в области автогенерации не используют теорию измерений в образовании, не проводят апробацию заданий, не оценивают их психометрическое качество. Основной упор делается на генерацию большого числа заданий, при этом качество этих заданий не анализируется. Психометрика как наука об измерениях в социальных науках задает стандарты, которым должны удовлетворять задания и тест в целом, чтобы обеспечить справедливое оценивание учащихся

с минимальной ошибкой измерения. Анализ может делаться в рамках классической теории тестирования или в рамках современной теории тестирования, при этом оцениваются как характеристики самих заданий (например, их трудность и дискриминативность), так и характеристики всего теста как измерительного инструмента (надежность и валидность). Пренебрежение этапами апробации заданий и анализом их психометрического качества может привести к ошибочным выводам о способностях учащихся и даже к их несправедливому оцениванию, что может негативно повлиять на их образовательные траектории.

Кроме того, важно отметить необходимость человеческого участия в процессе генерации заданий с применением LLM, ведь даже на текущем уровне развития ИИ подвержен ошибкам и расхождениям с желаемым результатом. В идеале команда проекта по генерации заданий с использованием ИИ должна включать специалистов из разных областей: эксперты в предметной области определяют содержание и оценивают качество формулировок заданий, обеспечивают связь разрабатываемых заданий с образовательными результатами; специалисты по измерениям оценивают психометрическое качество сгенерированных заданий и проверяют надежность измерения и валидность интерпретации его результатов; специалисты по NLP отвечают за технические аспекты автогенерации, обеспечивая эффективное использование LLM в соответствии с их характеристиками и особенностями.

2. Автоматическая оценка

Автоматическая оценка (автоскоринг) ответов на задания, требующих короткого конструируемого или развернутого текстового ответа (например, эссе), графического (например, рисунок), видео- или аудио- (например, запись речи школьника) ответа, является одной из важных задач в образовании. В традиционной практике проверка открытых ответов требует привлечения экспертов, что накладывает ряд ограничений, связанных не только с временны́ми и финансовыми затратами, но также и с проблемой субъективности оценки. Каждая оценка эксперта может быть подвержена искажениям, связанным с восприятием отдельно взятого эксперта (склонность усреднять все оценки или, напротив, ставить только максимальные или только минимальные баллы любым ответам, эффекты гало, неосознанного сравнения работ респондентов, усталости и т. д.) [Haley et al., 2007]. Разработка технологий автоскоринга таких заданий была призвана преодолеть эти проблемы. ИИ предлагает возможности для более объективной (в том смысле, что ко всем ответам будет применен единый «оценщик» в лице ИИ) и оперативной оценки, что особенно актуально в условиях массового обучения и при необходимости обработки большого количества ответов.

Исследователи уже более пятидесяти лет разрабатывают различные методы автоскоринга. Более ранние методы автоскоринга текстовых ответов основываются на предсказании баллов с помощью регрессионных уравнений для простых признаков (например, количества слов в тексте) [Page, 1967]; позже разработали и стали применять латентный семантический анализ [Landauer, Foltz, Laham, 1998]. Однако эти методы позволяли оценивать только текстовые ответы, то есть практически отсутствовала возможность автоскоринга графических, видео- и аудиоответов до появления нейросетей. С ростом вычислительных возможностей компьютеров и развитием ИИ удалось повысить точность оценивания текстовых ответов, появилась возможность оценивать графические, видео- и аудиоответы. В настоящее время мировое сообщество активно использует методы машинного обучения для автоскоринга заданий по школьным предметам естественно-научного профиля (биология, физика и химия) [Zhai et al., 2020], по математике [Erickson et al., 2020; Botelho et al.,

2023], программированию [Mehta et al., 2023], медицине [Masikisiki et al., 2023] и т. д.

Наиболее часто используемые в автоскоринге методы ИИ делятся на метрические (k Nearest Neighbors, kNN), линейные (регрессия, SVM) и ансамблевые (Random Forest, boosting, J48) [Brohi et. al., 2019], также используются нейросети и LLM. Перечисленные классы моделей машинного обучения и нейронные сети, а иногда и LLM, предполагают наличие обучающей и тестовой выборок — размеченных наборов данных, включающих ответы респондентов и баллы за них. На основе этого набора данных алгоритмы ИИ и нейросети находят определенные паттерны или закономерности, которые характерны для каждого из баллов. Дополнительно при обучении моделей может быть использована коллатеральная информация, например, время выполнения задания или различные лог-данные, собираемые в цифровой среде при компьютерном тестировании [Mathias, Bhattacharyya, 2018], а также дополнительные корпуса текстов [Zhang et. al., 2019], которые могут помочь модели уловить более тонкие закономерности. Далее на тестовой выборке проверяется качество предсказания полученного прототипа и в зависимости от результата принимается решение о готовности или неготовности модели к использованию. С ростом производительности LLM потребность в дообучении или тонкой настройке этих моделей решению задачи автоскоринга снижается, и зачастую исследователи применяют готовые LLM и подбирают различные промпты без примеров или с примерами [Jiang, Bosch, 2024], либо используют технологии обучения без примеров или с примерами [Liu et al., 2023].

Для оценки качества моделей, то есть меры корректности и валидности предсказания баллов моделями, применяются разные метрики, которые чаще всего основаны на оценивании меры согласованности предсказанных моделью баллов с эталонными (экспертными) баллами. Наиболее часто используемые метрики — точность (accuracy) и сбалансированная точность (balanced accuracy), которые показывают долю верно предсказанных баллов, или каппа Коэна (Cohen’s kappa) и средневзвешенная каппа Коэна, которые демонстрируют меру согласованности баллов, начисленных ИИ, и экспертных баллов. Иногда встречаются работы, где используют корреляцию Пирсона между данными баллами или коэффициент детерминации в регрессионной модели между баллами, нормализованными по методу минимакса.

Метрики точности меняются от 0% до 100%, где 0% означает, что модель не предсказала верно ни один балл, то есть предсказанные и экспертные баллы не совпадают, а 100% — что модель верно предсказала все баллы, то есть предсказанные и экспертные баллы находятся в полном соответствии. Каппа Коэна, средневзвешенная каппа Коэна и коэффициент корреляции Пирсона меняются от −1 до +1, где −1 означает совершенную несогласованность между предсказанными и экспертными баллами, 0 — отсутствие согласованности (оценки предсказаны хаотично), +1 — совершенную согласованность. Коэффициент детерминации изменяется от 0 до +1, где 0 показывает, что баллы, полученные ИИ, никак не предсказывают экспертные баллы, а +1 — что между баллами наблюдается линейная функциональная зависимость. Соответственно, чем выше значения перечисленных метрик, тем выше качество модели, то есть выше согласованность оценок ИИ и экспертных.

Таким образом, ИИ значительно облегчает задачу автоскоринга открытых ответов, причем в разных предметных областях, экономя при этом и временны́е, и финансовые ресурсы. Далее в данном разделе будут описаны возможности ИИ для оценки открытых ответов, их потенциальные преимущества и ограничения. Раздел состоит из трех частей, выделенных на основе формата ответа: текстовые, графические и аудио-/видеоответы.

2.1. Текстовые ответы

Для автоскоринга открытых текстовых ответов используется обработка естественного языка (Natural Language Processing, NLP) — раздел ИИ, созданный для работы с текстовыми данными. Основная идея этой технологии заключается в том, чтобы подобрать соответствующий метод для перевода текста из символьного (буквенного) представления в числовое. Достичь этого можно разными путями, начиная от простых методов, например, подготовки таблицы (матрицы), описывающей относительную частоту слов в ответах (Term Frequency Inverse Document Frequency, TF-IDF), и до векторных представлений слов (эмбеддингов) [Jurafsky, Martin, 2024]. Затем полученное числовое представление ответов используется для обучения моделей ИИ задаче автоматического оценивания открытых ответов.

При автоматической оценке текстовых ответов различают два их типа: 1) задания с кратким ответом, которые нередко рассматрива-

ются как полузакрытые ответы и 2) задания с развернутым открытым ответом (например, эссе, сочинения, решение задачи). Отличия коротких открытых ответов от развернутых заключаются, во-первых, в длине ответа, которая может варьироваться от одной фразы до одного абзаца, во-вторых, в критериях оценивания. В коротких открытых ответах критерии более сфокусированы на содержании, а не на связности изложения, как, например, в эссе [Higgins et al., 2004]. Несмотря на определенные различия между этими двумя видами открытых текстовых ответов, задача автоскоринга каждого из них непростая, что подчеркивается исследователями в области автоматического оценивания [Zhang et al., 2019]. Далее будут представлены результаты исследований по разработке моделей ИИ или использованию LLM для оценивания коротких открытых ответов, эссе и ответов по программированию и математике.

Проверка коротких открытых ответов. Чаще всего задания с короткими ответами встречаются в тестах по читательской грамотности, а также во многих тестах по языковым, естественно-научным и социально-гуманитарным предметам. Обзор результатов исследований по разработке моделей автоматического оценивания показывает, что точность предсказания варьируется от 65% до 99% [Çınar et al., 2020; Ariely et al., 2023; Raz et al., 2024; Dood et al., 2024].

Используемые методы ИИ также сильно варьируются: от алгоритмических методов и до LLM. Однако трудно выделить определенные методы, которые демонстрируют очень высокое качество автоскоринга, — так или иначе разработчики моделей применяют различные подходы для повышения качества модели. Например, Zhang и коллеги (2019) использовали модель для автоматического оценивания коротких ответов на тесты по смысловому чтению с применением нейросетей с долгосрочной кратковременной памятью (long-short-term memory recurrent neural network, LSTM RNN) в сочетании с увеличением обучающего корпуса путем объединения корпусов слов, демонстрирующих общие фоновые знания, связанные с предметной областью, и слов, демонстрирующих предметно-специфические знания. Таким образом, исследователи добились повышения качества предсказания, измеряемого через каппу Коэна, с 0.496 до 0.624.

С ростом производительности LLM растет число исследований, где делается попытка оценивать открытые ответы про помощи промпт-инжиниринга. Так, в работе Jiang и Bosch (2024) оценивались

возможности LLM в оценивании коротких ответов по английскому языку, биологии и предметам из естественных наук. Используя разные промпты, в частности, добавляя в них примеры правильных ответов, исследователям удалось добиться успехов в отдельных вопросах (средневзвешенная каппа > 0.7), а в среднем средневзвешенная каппа в зависимости от промпта варьировалась от 0.610 до 0.677.

Проверка эссе. В отличие от заданий с короткими ответами, которые требуют ответа точного и содержательного, эссе направлено на проверку умения тестируемых описывать и раскрывать свои идеи и концепции в письменной форме, руководствуясь лишь собственными мыслями. Соответственно, эссе предполагают длинный ответ, наличие четкой структуры и использование определенного стиля изложения. Еще одной отличительной чертой эссе является определенная свобода и субъективность изложения. Помимо особенностей самого эссе, отдельно стоит отметить многоаспектность рубрик их оценивания, многие из которых учитывают проверку грамматики, структуры, степени раскрытия темы т.д. Но, ввиду большого числа этих аспектов, исследователи как правило ограничивают набор признаков эссе, используемых ими при обучении моделей. Эти признаки условно делят на три категории: статистические (количество слов и предложений в эссе, средняя длина предложений и т.д.), стилевые (структура предложений, грамматика, пунктуация и т.д.) и контекстные (последовательность изложения, корректность, согласованность и т.д.) [Ramesh, Sanampudi, 2022]. Так, Darwish и Mohamed (2020) при разработке модели ИИ для предсказания балла по эссе использовали стилевые признаки, такие как выбор слов и структура предложений. Точность предсказания баллов моделью, которая учитывает указанные признаки, варьировалась от 77% до 83% в зависимости от набора данных, что выше, чем у моделей, которые не использовали дополнительные признаки [Darwish, Mohamed, 2020]. В исследовании Mathias и Bhattacharyya (2018) список стилевых признаков (частота слов, стиль изложения) был дополнен статистическими (число слов, длина предложений, пунктуация), и на полученном наборе признаков был обучен один из ансамблевых алгоритмов машинного обучения; при этом были достигнуты умеренные показатели качества предсказания (каппа Коэна варьировалась от 0.54 до 0.76). Примерно такого же качества предсказания с коэффициентом каппы Коэна, равным 0.764, добились Dong и коллеги (2017), которые использовали эмбендинги и

обучили сверточную нейронную сеть (CNN) в сочетании с LSTM RNN. Похожие результаты о сопоставимости качества традиционных методов и нейронных сетей были обнаружены в исследовании Yao и Jiao (2023). Более того, исследователи продемонстрировали, что тщательная предварительная обработка данных и добавление признаков, таких как индексы читабельности или сложности эссе, способны повысить качество скоринга и традиционных методов машинного обучения, и нейронных сетей.

Современные LLM позволяют снизить концентрацию внимания исследователей на поиск и выделение разных признаков эссе, которые способны повысить качество предсказания, поскольку могут работать с необработанными данными («сырыми ответами»). На сегодня LLM являются одним из многообещающих инструментов, их стремительно развивают и совершенствуют. Так, в начале написания работы исследования свидетельствовали о том, что LLM уступают по качеству оценивания специально обученным моделям ИИ. Например, в исследовании Masikisiki и коллег (2023) самый высокий показатель коэффициента согласованности каппа Коэна между оценками LLM и человеческими составил 0.53, что говорит об умеренной согласованности. Kostic и коллеги (2024) также подчеркивают проблемы точной оценки сложных текстов с помощью LLM в соответствии с заданными критериями и указывают на необходимость постоянных исследований и разработок в области LLM для повышения точности, надежности и согласованности автоматизированных систем оценки эссе в образовательных контекстах. В начале 2025 года известные LLM, такие как ChatGPT-4o, стали демонстрировать высокое качество производительности на бенчмарке EssayJudge, который содержит более 1000 эссе по 125 различным темам. EssayJudge [Su et al., 2025] — это первый мультимодальный бенчмарк для оценки возможностей LLM оценивать эссе по лексическим критериям, грамматике и пунктуации, а также оценивать общую структуру, аргументацию и связность эссе. Эксперименты авторов бенчмарка показывают согласованность оценок людей и ChatGPT-4o по лексическим критериям, а вот по корректности оценки по критерию «аргументация» LLM еще сильно уступают [там же].

Проверка ответов по программированию и математике. Рост производительности LLM позволяет оценивать решения задач по программированию и математике. Условно можно считать эту область

молодой, и пока исследователи только ищут пути для получения высокого качества оценивания ответов. Результаты на сегодня достаточно противоречивые. Так, Morris и его коллеги (2024) добились, благодаря грамотной предобработке данных, включающих решения заданий по математике, высокой согласованности между баллами ИИ и экспертными оценками (средневзвешенная Каппа равна 0.945). В программировании несколько иная картина: например, Mehta с коллегами (2023) обнаружили умеренные корреляции между оценками LLM и оценками экспертов по двум из трех типов заданий по программированию, а по последнему корреляции и вовсе были низкими (коэффициент корреляции Пирсона составил 0.14).

Таким образом, ИИ позволяет решать задачу автоскоринга коротких ответов и эссе по разным предметным областям. Качество оценивания моделями ИИ варьируется в зависимости от набора используемых признаков их предварительной обработки и моделей ИИ. Более того, обученные задаче автоскоринга модели показывают более высокое качество, чем LLM без дообучения. Хотя большие языковые модели могут обеспечить дополнительную перспективу в автоматической оценке, они еще не готовы к независимой автоматизированной оценке без человеческого контроля [Schneider et al., 2023].

2.2. Графические ответы

Современные технологии позволяют автоматизировать оценку графических ответов, таких как рисунки, схемы и диаграммы, выполняемые в рамках учебных и диагностических заданий. Более того, они помогают анализировать креативные и когнитивные аспекты ответов. Далее приведем области применения ИИ, в которых он может использоваться для оценки графических ответов.

В *образовательных тестах*, где учащимся нужно рисовать графики или диаграммы (например, по математике или биологии), ИИ может анализировать правильность выполнения задания. Хотя автоматическое оценивание еще не заявлялось в крупных международных исследованиях как официальная процедура, но уже есть ряд исследований, показывающих возможности ИИ на примере оценки графических решений заданий из TIMSS-2019. Использование сверточных нейронных сетей (CNN) для данной задачи демонстрирует значительный потенциал для повышения точности и результативности автоматизи-

рованных оценок. Исследования показывают, что CNN могут классифицировать ответы на основе изображений с высокой точностью, превосходя возможности людей в скорости оценки без потери качества. Модели CNN достигли точности предсказания баллов в 97.53% в ранних работах [von Davier et al., 2023], а в более поздних уже выявлено, что с дихотомическими элементами задания CNN справляются в более чем 99% случаях [Tyack et al., 2024], причем было обнаружено множество ответов, которые люди-оценщики неправильно классифицировали, что свидетельствует о более высокой надежности автоскоринга графических ответов с помощью ИИ. Таким образом, можно сделать вывод, что внедрение CNN может снизить нагрузку и затраты, связанные с оценщиками-людьми, существенно не потеряв в точности оценки.

В *клинической практике* могут использоваться задания, где испытуемым предлагается нарисовать определенные объекты, как, например, в тесте рисования часов для диагностики когнитивных нарушений (деменции, болезни Альцгеймера). ИИ может унифицировать оценку рисунков, анализируя такие параметры, как симметрия, пропорции, детализация, что помогает в объективной диагностике когнитивных способностей. С помощью методов CNN точность прогнозирования состояния когнитивных способностей респондента из трех возможных (нормальные, умеренно нарушенные или сильно нарушенные) на основе одного теста рисования часов составила 71% [Youn et al., 2021]. Более продвинутые архитектуры нейросетей, такие как Attentive Pairwise Interaction (API-net), для этих же тестов демонстрируют точность в 78% [Raksasat et al., 2023]. Добавление в нейронные сети CNN слоев самовнимания (self-attention) позволяет повысить точность оценивания ответов на тесты по выявлению когнитивных нарушений до 81% [Ruengchaijatuporn et al., 2022]. Следует отметить, что не стоит принимать решений с высокими ставками на основе только оценки ИИ, но текущие исследования уже показывают, что использование ИИ в клинической практике может помочь в выявлении некоторых когнитивных нарушений и тем самым ускорить оказание соответствующей помощи.

Психологические тесты, такие как проективные методики или тесты на креативность, где испытуемым предлагается создать рисунок, могут быть также автоматизированы с помощью ИИ. Алгоритмы могут анализировать ключевые особенности рисунков, которые коррелиру-

ют с психологическими характеристиками, — такие как эмоциональная выраженность, использование пространства, стиль рисования и другие аспекты, помогающие выявить уровень тревожности, депрессии, креативности или другие черт личности. ИИ может быть полезен в определении унифицированных и воспроизводимых систем оценки, что особенно важно в проективной психодиагностике.

На текущий момент уже существует множество исследований по применению CNN в оценке различных тестов на креативность и творческое мышление; при этом чаще всего встречается автоскоринг «рисуночного теста творческого мышления» Урбана (TCT-DP). Например, в международном исследовании Cropley и Marrone (2022) изучалась точность классификации на три уровня развития творческого мышления (низкий, средний и высокий), которая составили 94.2%. Этот инструмент также рассматривался в российском исследовании Панфиловой и коллег (2024), в рамках которого предсказывались баллы на шкале от 1 до 63, полученной при суммировании баллов по 14 оценочным категориям, что делает задачу для ИИ более сложной, чем при классификации на несколько групп. Были обучены различные модели CNN, и наилучший результат с коэффициентом детерминации 0.76 был получен при использовании модели MobileNet V2, которая также была использована в исследовании Cropley и Marrone (2022). Другие часто встречающиеся модели CNN в работах — это AlexNet, DenseNet, ResNet, VGGNet.

В Институте образования НИУ ВШЭ также были предприняты попытки использования CNN: для инструмента 4K, оценивающего в том числе креативность, было проведено исследование, в котором сравнивались оценки на основе латентного классового анализа и модели Densenet201 с целью бинарной классификации по двум факторам, а именно разделение изображений на недетальные и детальные, неоригинальные и оригинальные. Точность составила около 88% по обоим факторам [Угланова и др., 2021].

В этом направлении, как и в тестах клинической практики, пока трудно достигнуть высокой точности, так как существует множество индивидуальных особенностей созданных тестируемыми решений по сравнению с образовательными тестами, где проверяется в основном верность выполнения. Однако использование алгоритмов ИИ открывает новые горизонты в оценивании графических заданий, что в перспективе делает их незаменимым инструментом для разработки

высокоточных и надежных систем оценивания в различных контекстах, от школьных тестов до медицинских обследований. Но следует отметить, что, несмотря на многообещающие результаты, переход к автоматизированным системам оценки должен учитывать нюансы человеческого суждения при оценке, особенно в тестированиях с высокими ставками, гарантируя, что технология будет дополнять, а не полностью заменять проводящих оценку людей.

2.3. Аудио-/видеоответы

Оценка аудио- и видеоматериалов с помощью ИИ активно используется в таких сферах, как образование, здравоохранение и рекрутинг. Современное развитие ИИ позволяет автоматически оценивать устные ответы и видеозаписи выступлений. Аудио- и видеоматериалы имеют свои особенности, которые требуют более сложных моделей ИИ для обработки таких данных. Вот *основные отличия оценки аудио- и видео- от текстов и изображений:*

• Временна́я шкала: аудио- и видеоматериалы требуют анализа во времени, поскольку это динамические данные. Тексты и изображения, напротив, предоставляют статическую информацию, что упрощает их обработку.

• Мультимодальность: в отличие от текста или изображения, которые содержат один вид информации, видео- и аудиоматериалы объединяют несколько источников данных. Например, в аудиозаписи устного ответа можно анализировать как смысл сказанного, так и интонации и произношение, а в видео добавляется невербальная информация — мимика и жесты.

• Качество данных: аудио- и видеоматериалы сильно зависят от качества записи (шумы, освещение, разрешение), что может усложнить работу ИИ. Тексты и изображения, напротив, часто более «чистые» и легко структурируются.

Таким образом, для анализа такого типа данных используются мультимодальные модели, которые могут обрабатывать разные типы данных. Мультимодальные модели ИИ — это модели, способные обрабатывать и интегрировать данные из различных модальностей (например, изображения, звук, видео). Они объединяют информацию из разных источников для более глубокого понимания контекста и улучшения качества решений.

Аудио- и видеоответы можно применять в разнообразных заданиях: чтение вслух, устные ответы на вопросы, монологические высказывания и участие в диалогах. Такой тип оценивания особенно популярен в образовании при тестировании языковых навыков. Поэтому подавляющее большинство ИИ-моделей в языковых тестированиях направлены на предсказание уровня по общеевропейской шкале языковых компетенций (The Common European Framework of Reference for Languages, CEFR).

Качество оценивания аудиоответов с применением мультимодальных моделей стремительно повышается. В исследованиях 2021 года согласованность моделей по сравнению с экспертами была умеренной: взвешенная каппа Коэна равнялась 0.6 и 0.52 соответственно для MMAF (Мультимодальная модель с применением внимания) [Grover et al., 2020] и для предобученной языковой модели BERT [Singla et al., 2021]. В исследовании, проведенном в 2024 году исследователями из Тайваня, были использованы новые стратегии моделирования, которые позволили дообучить BERT для оценки устных ответов с высокой точностью — 93% [Lo et al., 2024]. Методология оценки аудиоответов уровня владения языками совершенствуется, и исследователи находят пути, как оценивать при помощи ИИ без дообучающих данных. Так, в исследовании Gupta и коллег (2024) применялась модель XLS-R (большая языковая модель для обучения кросс-языковому представлению аудио), предсказания которой коррелировали с оценками экспертов на среднем уровне r = 0.62 — 0.65.

В клинической практике ИИ можно использовать для диагностики с помощью анализа голоса пациента таких заболеваний как болезнь Альцгеймера [Agbavor, Liang, 2022], болезнь Паркинсона, деменция и депрессия [Idrisoglu et al., 2023]. Модели ИИ, такие как метод опорных векторов (SVM) и искусственные нейронные сети (ANN), демонстрируют точность при диагностике перечисленных заболеваний от 79% до 99%, что открывает новые возможности для диагностики и мониторинга заболеваний.

Оценка видеоматериалов также широко используется, особенно для анализа публичных выступлений и видеоинтервью. Например, асинхронные видеоинтервью (AVI) позволяют кандидатам записывать свои ответы на заранее подготовленные вопросы, а ИИ анализирует их вербальные и невербальные характеристики. Основное отличие видео от аудио — это наличие визуальных данных. Однако визуальные данные не всегда гарантируют наличие большей информации. Напри-

мер, в исследовании, посвященном предсказанию интеллектуальных способностей по видеоинтервью, было установлено, что невербальные признаки, такие как выражение лица, наименее информативны [Hickman et al., 2024]. Это же подтверждается в исследовании Hsiao и коллег (2017), которые выяснили, что голосовые сигналы являются более значимыми для оценки публичных выступлений, чем визуальные данные.

Оценивание видеоматериалов пока менее распространено в образовании, но активно применяется в психологии и в HR. Исследования в этих областях показали, что модели машинного обучения могут предсказывать интеллект на основе вербальных данных с корреляцией с экспертными оценками до r = 0.67 [там же]. Кроме того, в работе Koutsoumpis и коллег (2024) было установлено, что асинхронные видеоинтервью (AVI) обладают потенциалом для предсказания личностных черт, таких как экстраверсия (r = 0.62), однако предсказание других характеристик, таких как сознательность, требует доработки. В России также применяется оценка персонала на основе видеоинтервью. Например, в компании «Экопси» для оценки кандидатов применялась модель ИИ, которая показала точность в 74% [Экопси, 2024]. Таким образом, можно сказать, что оценка видеоответов с помощью ИИ имеет высокий потенциал для применения, однако на данном этапе результаты предсказания недостаточно высоки, чтобы использовать ИИ для оценивания с высокими ставками.

Оценивание на основе аудио- и видеоматериалов — стремительно развивающаяся область, которая уже сейчас находит практическое применение. Внедрение мультимодальности в LLM позволяет обрабатывать видео- и аудиозаписи, что открывает новые перспективы в оценивании. В частности, появляются перспективы для разработки заданий, требующих устного ответа и более близких к реальным жизненным ситуациям. Это важно для предметов, проверяющих языковые, коммуникативные навыки, навыки коллаборации, актерские навыки. Более того, появляется возможность улучшить процедуру оценивания учащихся с ОВЗ с нарушениями опорно-двигательного аппарата, зрения и т.д. за счет увеличения заданий с устным ответом.

2.4. Выводы по главе

Автоматизация процесса оценивания заданий открытого типа способствует повышению объективности результатов, ускорению

процесса оценки, снижению нагрузки на экспертов и оптимизации распределения финансовых ресурсов, что особенно актуально в условиях массового тестирования и обработки значительного объема ответов. Современные достижения в области ИИ позволяют автоматически оценивать не только текстовые, но и графические, аудио- и видеоответы.

Качество предсказаний, генерируемых моделями ИИ, может значительно варьироваться, при этом более высокие результаты демонстрируют модели, специально обученные задаче автоскоринга. Однако полностью исключить эксперта из процедуры оценивания не представляется возможным, поскольку даже системы, основанные на LLM, не готовы к самостоятельной автоматизированной оценке без вмешательства человека. Одним из возможных путей повышения качества предсказания представляется интеграция ИИ с психометрическими моделями.

Несмотря на очевидные преимущества использования моделей ИИ для автоскоринга, следует учитывать и ограничения данного подхода. Поскольку многие методы ИИ требуют наличия обучающей выборки для построения модели, необходимо тщательно проверять исходные данные, оцененные экспертами. Это включает проверку согласованности оценок экспертов, их последовательности в оценках, отсутствия предвзятости, излишней строгости. Модель, обученная на данных, которые не проходили такой проверки, будет воспроизводить названные изъяны и оценивать задания соответствующим образом.

3. Обратная связь

Обратная связь — это информация, которую обучающийся получает о прогрессе обучения, уровне освоения знаний, глубине понимания предмета изучения. Основная цель практики оценивания и предоставления обратной связи — способствовать переходу от действия в зоне ближайшего развития к самостоятельному действию. Эффективная обратная связь должна отвечать нескольким основным требованиям [Haughney et al., 2020; Wisniewski et al.,2020]:

- быть своевременной (чем меньше времени проходит от выполнения действия до обратной связи, тем лучше);
- быть развернутой и детальной (чем больше качественной и точной информации предоставляет обратная связь, тем она эффективнее);
- быть персонализированной (адаптированной для возраста, уровня мотивации, компетентности, учитывающей личностные особенности, ситуативный контекст и т.д.).

На практике далеко не всегда обратная связь соответствует этим требованиям: в школе ученики чаще всего получают оценки без комментариев или обобщенные комментарии по результатам работы всего класса [Panhoon, Wongwanich, 2014]. При том что сами ученики хотели бы получать более подробные и персонализированные комментарии [Азбель и др., 2021], процесс подготовки обратной связи по заданиям из высших категорий таксономии образовательных результатов Блума (оценивание, создание) [Pereira et al., 2016] весьма трудоемок, преподаватели не всегда обладают навыком давать эффективную обратную связь [Азбель и др., 2022; Nieminen, Carless, 2023].

Исследователи признают, что интеграция ИИ может помочь повысить качество обратной связи за счет возможности справедливо и эффективно анализировать данные о процессе обучения и регулярно информировать учителей, студентов и родителей. Это подчеркивают и недавние рекомендации ЮНЕСКО [Miao et al., 2021]. В данном разделе мы рассмотрим примеры исследований, где технологии ИИ используются для генерации обратной связи в оценивании (в заданиях с выбором ответа и открытыми ответами), обозначим возможности ИИ для генерации персонализированной обратной связи и рассмотрим преимущества и недостатки обратной связи, сгенерированной ИИ, по

сравнению с традиционной обратной связью, которую готовит и сообщает учащемуся преподаватель.

3.1. Обратная связь ИИ в заданиях с выбором ответа и открытыми ответами

До недавнего времени автоматизация систем обратной связи строилась на шаблонах, разработанных экспертами. В частности, участники, выбирая вариант ответа, получали заранее подготовленные фрагменты обратной связи. Однако такие системы плохо адаптируются к изменениям, например, в структуре курса, и часто не учитывают разнообразные потребности в обучении людей, или, скажем, разные темпы их обучения. К тому же создание обширного банка заданий требует значительных усилий от специалистов по разработке таких фрагментов. В этом контексте технологии искусственного интеллекта предоставляют более эффективные решения.

В работе Norrthon и Schörling (2023) рассматривается успешное применение ИИ для генерации обратной связи к каждому варианту ответа в тестах по программированию для студентов. Авторы использовали детализированные промпты, которые позволяли не только генерировать обратную связь, но и улучшать экспертно составленные фрагменты обратной связи, а также оптимизировать формулировки заданий и вариантов ответов. Пример промпта из этой статьи для генерации обратной связи для заданий с выбором ответа представлен в тексте в рамке[6].

Авторы статьи указывают, что при разработке промптов важно четко определить задачу для ИИ и установить критерии для оценки качества обратной связи:

• Начало промпта «Представь, что ты эксперт в языке программирования…» — необходимо, чтобы LLM приняла на себя роль эксперта и подстроила свои ответы под эту роль.

• Промпт содержит информацию о том, как будет выглядеть задание («Ты получишь вопрос с несколькими вариантами ответа…») и варианты ответа («Все варианты будут содержать букву»).

• Приводится описание основной задачи: создать обратную связь по каждому варианту ответа, а также детализация задачи, например

[6] Дан перевод на русский язык. Оригинал промптов на английском языке можно найти в дополнительных материалах статьи.

Представь, что ты эксперт в языке программирования Java и теме, к которой относится следующий вопрос. Ты получишь вопрос с несколькими вариантами ответа. Все варианты будут содержать букву, затем двоеточие (:) и сам вариант ответа. Например, ты получишь «A: вариант 1, B: вариант 2», и это два разных варианта.

Твоя задача — использовать всю информацию, которую ты получаешь, и свои знания в этой области, чтобы создать обратную связь по каждому варианту ответа. Обратная связь должна указывать, является ли ответ правильным или неправильным, вместе с объяснением, почему.

Обратная связь должна соответствовать следующим критериям:

- она конструктивна;
- она уникальна и предоставляется для каждого варианта ответа;
- для вариантов ответа она не раскрывает правильный ответ.

Обратная связь должна быть представлена в следующем формате:

A: [правильный или неправильный в зависимости от того, правильный ответ или нет]; объяснение, почему.

B: то же самое для варианта B, и так далее для всех вариантов.

Обратная связь не должна содержать другой информации.

Вопрос выглядит следующим образом: альтернативы ответов: A: B: C: D: Правильный ответ:

Пожалуйста, создай обратную связь для предложенных вариантов ответа. Помни, что только один вариант ответа должен быть правильным. Обдумай свой ответ, прежде чем напечатать, и убедись, что ты используешь всю информацию, которую знаешь по этой теме. Обратная связь должна быть понятна студенту первого курса бакалавриата, который изучает информатику.

«Обратная связь должна содержать указание на правильный или неправильный ответ».

• Установлены критерии качества обратной связи. Это необходимо для того, чтобы LLM ориентировалась на критерии при генерации ответа, а также для дальнейшей оценки качества результатов модели.

• Описан формат ответа LLM: «Обратная связь должна быть представлена в следующем формате…»

• Указана целевая аудитория: «Обратная связь должна быть понятна студенту первого курса бакалавриата».

По результатам исследования и эксперты, и учащиеся положительно оценили обратную связь, сгенерированную ИИ по промптам. Обратная связь была уникальной, предоставлялась для каждого варианта ответа, была сформулирована корректно и объясняла, почему каждый вариант является правильным или неправильным [Norrthon, Schörling, 2023].

Большинство примеров использования ИИ для генерации обратной связи основаны на материалах курсов по программированию [Cavalcanti et al., 2021]. Большие языковые модели добились значительных успехов в проверке корректности кода в контексте обучения программированию [Gabbay, Cohen, 2024; Pankiewicz, Baker, 2023]. LLM анализируют исходный код напрямую, что позволяет предоставлять целенаправленную обратную связь, а также решать вопросы, связанные с качеством кода, синтаксическими ошибками и структурными требованиями [Maier, Klotz, 2022]. Реже встречаются исследования, где обратная связь по заданиям с вариантами ответов разрабатывается для других предметных областей, например, математики [McNichols et al., 2023]. Особое внимание технологиям ИИ для генерации обратной связи уделяют в массовых онлайн-курсах [Gabbay, Cohen, 2024].

Преимущества использования ИИ становятся особенно очевидными в контексте заданий с открытыми ответами. Для таких заданий сложно применять заранее подготовленные фрагменты обратной связи из-за высокой вариативности ответов, что значительно увеличивает время, необходимое для проверки и предоставления детализированной обратной связи в соответствии с установленными критериями качества. Технологии ИИ для генерации обратной связи уже успешно апробированы на заданиях по математике с открытым ответом [Kakarla et al., 2024]. Результаты исследования показывают, что большие языковые модели обладают потенциалом для предостав-

ления обратной связи по открытым заданиям по математике, однако испытывают трудности при обнаружении уникальных ошибок участников, не встречавшихся в данных, на которых была обучена модель [McNichols et al., 2024].

Ведутся исследования по генерации обратной связи на основе эссе и других форматов письменных работ [Wang et al., 2024]. В одном из исследований приведен пример промпта, который был предложен ИИ для оценки эссе учеников младших классов по заданным критериям [Fokides, Peristeraki, 2024] (табл. 2).

В рамке показан пример промпта, используемый для получения обратной связи на задания с открытыми ответами с помощью ИИ[7].

«Я хочу, чтобы ты выступил в роли греческого учителя начальной школы. Ниже приведен текст, написанный одним из учеников. Поскольку текст написан на греческом языке, используй все имеющиеся знания о греческом языке. Составь список грамматических, синтаксических и орфографических ошибок, которые ты обнаружил. (Помни, что в греческом языке неправильная постановка знака ударения или отсутствие знака ударения в слове, в котором он должен быть, считается орфографической ошибкой.) Не описывай одну и ту же ошибку дважды. Не группируй похожие ошибки, перечисляй их отдельно. Кроме того, во второй части я хочу, чтобы ты написал комментарии о сильных сторонах текста, то есть о положительных элементах, которые ты нашел в грамматике, орфографии, синтаксисе, выразительности, структуре и содержании. В третьей части я хочу, чтобы ты записал свои комментарии к тексту о слабых сторонах, которые ты нашел в грамматике, орфографии, синтаксисе, выразительности, структуре и содержании. Затем напиши итоговые комментарии в качестве обратной связи ученику, взяв за основу найденные ошибки, а также сделанные замечания о сильных и слабых сторонах текста. Исходя из серьезности найденных ошибок, а также на основании сделанных комментариев и предоставленного отзыва оцени текст по шкале от 1 до 10. Не обосновывай свою оценку. Не переписывай текст с исправлениями. Текст, который я хочу, чтобы ты проверил, следующий: “…”».

[7] Перевод с английского.

Авторы приводят пояснение к промпту.

• Первая команда «Я хочу, чтобы ты выступил в роли греческого учителя начальной школы» требует LLM вести себя в соответствии с ролью учителя начальных классов. Такие директивы необходимы, чтобы ChatGPT продемонстрировал опыт в той или иной профессиональной области.

• Инструкция: «Поскольку текст на греческом языке, используй все имеющиеся знания о греческом языке» позволяет ChatGPT задействовать все лингвистические возможности.

• Чтобы избежать избыточности, в текст был включен пункт «Не повторяйте одну и ту же ошибку дважды», так как ошибки часто повторяются.

• Условие «Не группируй похожие ошибки, перечисляй их отдельно» учитывало тенденцию ChatGPT объединять ошибки в общие группы грамматических или синтаксических ошибок и не давать указаний по конкретным ошибкам.

• Команда «Затем напиши итоговые комментарии в качестве обратной связи с учеником, используя в качестве основы…» предоставила ChatGPT основу для построения комплексной обратной связи. Обратите внимание, что ему намеренно не было дано указание предоставить конкретный тип/категорию обратной связи, чтобы добиться наибольшей полноты ответа.

• Что касается оценки, то команда «Исходя из серьезности найденных ошибок, … оцени текст по шкале от 1 до 10» определяла критерии, которыми ChatGPT должен руководствоваться при оценке текстов. Можно было бы дать более подробные инструкции (например, определить конкретное количество баллов за каждый тип ошибок или положительные/отрицательные стороны текста). Поскольку в тексте может быть много ошибок, но структура и содержание могут быть хорошими, или наоборот, а преподаватели решили подходить индивидуально к оцениванию каждой работы, было решено не добавлять никаких уточнений.

• Указание «Не обосновывай свою оценку…» было основано на предположении, что обратная связь уже дает достаточно информации о том, почему выставлена такая оценка.

• И, наконец, инструкция «Не переписывай текст с исправлениями…» была включена, чтобы предотвратить создание ChatGPT отредактированных текстов.

Описан случай использования генеративной модели для оценки и предоставления обратной связи по письменным работам у студентов педагогической специальности [Ossa, Willatt, 2023]. Студенты получили задание письменно описать решение нескольких педагогических ситуаций. Для подготовки обратной связи была использована модель «Предварительно обученная генеративная трансформационная педагогическая обратная связь по написанию кейсов» (RECaP-GPT), которая включала активные действия преподавателя, студента и задействовала платформу IAG с API ChatGPT-4 (Application Programming Interface). Промпт, который преподаватели предлагали модели, включал три группы критериев: соблюдение заданной структуры ответа, качество решения кейса, речевое оформление текста (грамотность, стиль, последовательность и т.д.). Время, необходимое модели для оценки одной работы, составило от 45 до 90 секунд, что значительно меньше необходимого времени для работы экспертов. Студенты получали оценку, а также обратную связь о степени выполнения критериев (полностью выполнено/не полностью выполнено) и рекомендации по улучшению работы для повышения балла. Участники отметили, что модель предоставляла точные и обоснованные предложения, которые были полезны для их дальнейшего обучения.

В рамке показан пример фрагмента обратной связи по письменному решению кейса, сгенерированной ИИ[8].

Критерий 1. Структура письменного ответа		
Часть 1. Описание проблемы	Выполнено	
Часть 2. Анализ ситуации и обоснование предложенного решения	Выполнено не полностью	Анализ и обоснование предложенного решения приемлемы, но могли бы быть убедительнее за счет более явного привлечения теорий обучения и развития. Например, можно упомянуть теорию социального научения Бандуры, чтобы обосновать привлечение сверстников к решению проблемы.

[8] Перевод с испанского.

Часть 3. Программа действий по решению ситуации	Выполнено	
Количество слов	Выполнено	
Критерий 2. Качество предложенного решения		
Стратегия действий	Хороший уровень	Стратегия, предложенная студентом, уместна, но могла бы выиграть от более детальной проработки и более четкой связи с теориями обучения и развития.
Обоснование стратегии	Достаточный уровень	Несмотря на упоминание социокультурной теории Выготского, было бы полезно подробнее рассказать о том, как эта теория и, возможно, другие теории обучения и развития конкретно поддерживают предложенную стратегию действий.
Критерий 3. Речевое оформление текста		
Логика	Очень хороший уровень	
Грамотность	Хороший уровень	Хотя большая часть текста написана хорошо, есть несколько орфографических и пунктуационных ошибок, которые могут отвлекать читателя и затруднять понимание.
Стиль	Очень хороший уровень	
Заключение: Студент демонстрирует четкое понимание поставленной проблемы и предлагает логичное и подробное решение ситуации. Тем не менее, ответ мог бы выиграть от более надежного теоретического обоснования и улучшения орфографии. В будущих работах студенту следует:		

- Проверить текст и исправить орфографические и пунктуационные ошибки.
- Четко показать, как предлагаемая программа действий подкрепляется соответствующими теориями обучения и развития.
- Дать более глубокий анализ проблемы, которую необходимо решить, и того, как предлагаемая стратегия действий должна решить эту проблему.

Современные технологии ИИ позволяют генерировать обратную связь как для заданий с вариантами ответа, так и для открытых ответов. Это значительно сокращает время получения обратной связи и одновременно предоставляет обучающемуся детальную информацию о том, почему ответ является правильным или неправильным, а также рекомендации для его улучшения. Кроме того, технологии ИИ открывают новые возможности для персонализации обратной связи.

3.2. Возможности ИИ для персонализации обратной связи

Персонализация обратной связи — одно из условий ее эффективности: информация, которую можно предоставить дошкольникам в детском саду, старшеклассникам, сдающим экзамены, студентам вузов и высокопрофессиональным специалистам, проходящим курс повышения квалификации, очевидно, будет сильно отличаться. Кроме возраста играет роль и ситуативный контекст, например, эмоциональное состояние учащегося, его уровень мотивации, а также более стабильные личностные факторы: наличие особых образовательных потребностей, черты характера и т.д. Персонализация обратной связи также имеет важное значение в инклюзивном образовании, поскольку ученики с особыми образовательными потребностями чаще других выпадают из общего темпа, не справляются с заданиями, которые легко даются большинству учеников.

Основное достоинство обратной связи с применением ИИ заключается в том, что эта технология может способствовать одновременно более широкому (в смысле охвата и обработки используемых данных) и более глубокому пониманию учебного процесса и прогресса обучающихся. Широкие возможности для персонализации открывают чатботы: программы на основе ИИ, которые имитируют устную или пись-

менную речь и могут поддерживать диалог. Пользователь вводит или произносит вопрос, и чат-бот отвечает, предоставляя информацию или выполняя более-менее простую задачу. Например, так устроены голосовые помощники: Алиса, Siri или Xiaoyi — они используют обработку естественного языка и машинное обучение для генерации уникальных ответов. Современные исследования изучают возможности адаптирования чат-ботов для образовательных задач. В образовательных контекстах чат-боты помогают в ситуациях от ответов на вопросы приемной комиссии в вузе (например, «А у вас есть курсы по вычислительным наукам?») до прямой поддержки обучения (например, в рамках диалоговой учебной системы) [Agostini, Picasso, 2024].

Рассмотрим несколько кейсов использования чат-ботов в образовании разных уровней и в разных сферах.

В работе Liang и коллег (2024) исследователи поставили перед собой задачу разработать для курса программирования такую модель, которая генерировала бы обратную связь, полезную и студентам для решения образовательных задач, и преподавателям для высокоточных прогнозов, кто из учащихся имеет риск остаться на второй год или бросить обучение, — и им это удалось. Авторы делятся кодом созданной ими системы в репозитории GitHub[9].

Еще один аналогичный пример — кейс университета Мадрида (UAM), где разработана система персонализированной обратной связи для учебных курсов на платформе edX [Becerra et al., 2024]. Разрабатывая модель, авторы пробовали несколько видов промптов для разных GenAI: ChatGPT демонстрировал более высокую производительность по сравнению с Bard и Llama. Получившийся инструмент называется GePeTo (акроним для «Generative AI-based Personalized Guidance Tool») и интегрирует сразу несколько систем:

1) использует анонимизированные данные множества учащихся курса и анализирует их;

2) подключается к серверу генеративного ИИ (авторы использовали несколько разных моделей и далее называют их просто GenAI) и генерирует персонализированные рекомендации для учащегося, адаптируя их к определенному количеству символов и формату дашборда;

3) проверяет ответы GenAI перед их отображением на дашбордах учащихся.

[9] https://github.com/CoLAMZP/AIED-2024-AutoFeedback.

Модель генерирует индивидуализированный веб-дашборд для каждого студента: информационную панель, которая структурирует и визуализирует информацию о прогрессе в изучении курса (рис. 1). Информацию, сгенерированную моделью, преподаватели используют, чтобы отслеживать ход обучения, а студенты — чтобы видеть свой прогресс и получать персональные рекомендации.

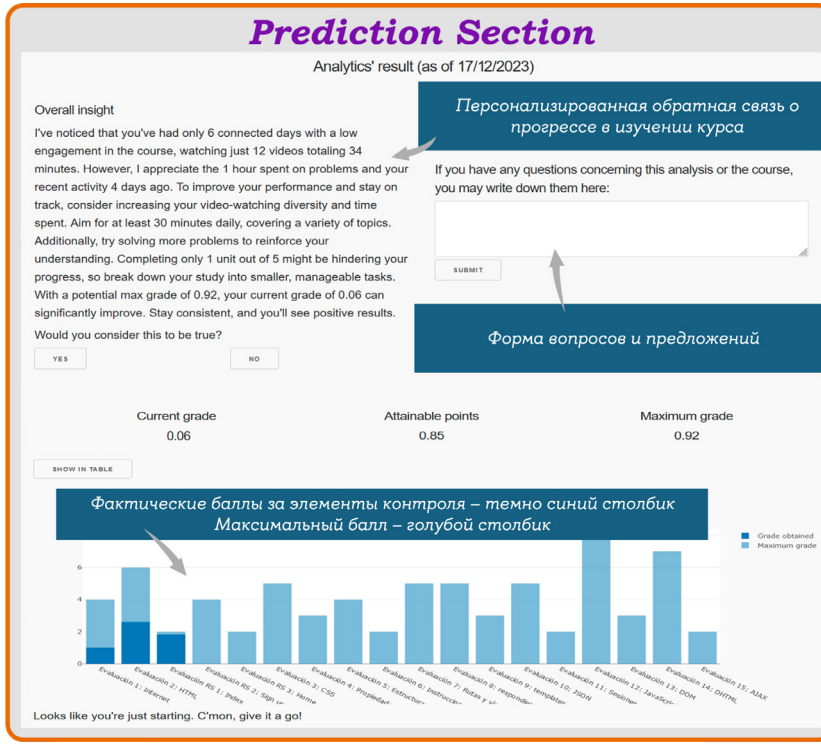


Рис. 1. Пример персональной информации о прогрессе обучения, которую получает учащийся в личном кабинете [Becerra et al., 2024]

Системы, подобные GePeTo, помогают учащимся определить свои точки роста, снизить чувство изоляции и получить поддержку, благодаря чему они не бросают обучение на курсах.

Однако бывают и менее удачные варианты реализации, о которых мы поговорим в разделе, посвященном этическим проблемам.

Использование персонализированной обратной связи, предоставляемой ИИ, эффективно и для творческих занятий. Например, описан опыт использования тренажера по игре на фортепиано на базе ИИ. Тренажер был заранее обучен на материале, содержащем записи упражнений, частых ошибок и замечаний преподавателей по музыке, и продолжал обучаться в процессе занятий студентов. Студенты пользовались подсказками тренажера и получали обратную связь, в которой содержалась подробная информация о том, какие были совершены ошибки, какие упражнения стоит выполнить и как подготовиться к следующему занятию. Группа, которая пользовалась тренажером, достигла более высоких результатов и проявила большую вовлеченность в занятия музыкой [Hua Zhen Lv., 2023].

Описаны кейсы применения чат-ботов в системе школьного образования. Технологии на базе ИИ позволяют не просто давать готовый ответ (правильно / неправильно), но за счет последовательно выстроенных вопросов помогать ученикам самим найти ответ, что особенно эффективно при обучении детей младшего школьного возраста, поскольку такая форма более увлекательна, помогает детям самим активно анализировать получаемую информацию и мотивирует применять ее на практике [Carless, 2016]. Так, в исследовании по использованию такого чат-бота для изучения иностранного языка младшими школьниками, разработанного на базе сервиса Google Dialogflow [Jeon, 2023], чат-бот помогал ученикам начальной школы догадаться, что означает незнакомое английское слово. Бот последовательно делал следующие действия:

1) предлагал еще раз перечитать предложение и попробовать догадаться, что означает слово;

2) выделял фрагменты текста, которые могут помочь ученику понять значение слова;

3) давал подсказку, как связаны выделенные фрагменты текста и неизвестное слово;

4) предлагал предложение с новым контекстом, где легко понять, что за слово было использовано.

Ученики, которые работали с чат-ботом, добились лучших результатов в увеличении словарного запаса.

Также разработан прототип тренажера по математике для учеников начальной школы, который может обеспечивать детям поддержку

в зоне ближайшего развития. Тренажер не дает обратную связь напрямую, но, опираясь на ответы учеников, адаптирует темп и сложность обучения под их уровень и подсказывает, как решить задания, вызывающие затруднения [Dijk, 2021].

Персонализация обратной связи с помощью ИИ может быть полезна в ситуациях, когда получение обратной связи учеником затруднено из-за ограниченных возможностей зрения или слуха. Например, ИИ может превращать аудиосигналы в субтитры; кроме того ведутся исследования, как ИИ может помочь детям с нарушениями зрения [Garg, Sharma, 2020]. Изучается также, как роботы, обученные поддерживать диалог, могут быть полезны для обучения детей с расстройствами аутического спектра (РАС), поскольку есть свидетельства, что дети меньше пугаются роботов и легче вступают с ними во взаимодействие в силу их предсказуемых реакций, — и такие занятия могли бы дополнить занятия с человеком [Rudovic et al., 2018].

3.3. Сравнение традиционной и сгенерированной обратной связи

Существуют разные взгляды на использование ИИ для обратной связи.

В одном из качественных исследований студенты и преподаватели рассказывали, как, по их мнению, воспринимается обратная связь от ИИ по сравнению с обратной связью преподавателя [Otaki, 2023]. Были отмечены следующие преимущества ИИ:

- **беспристрастность и эффективность**: ИИ не устает, не отвлекается и всегда доступен для оценивания и предоставления обратной связи;
- **детальность**: ИИ способен замечать даже мельчайшие ошибки и нюансы, которые человек может пропустить;
- **скорость и доступность**: студенты могут получить мгновенную обратную связь, что помогает им быстрее сориентироваться в своих ошибках, что в том числе помогает быстрее освоить новый материал на первых этапах обучения.

При этом результаты другого исследования показали, что LLM лучше людей выявляют ошибки в текстах на английском языке, в то время как по другим языкам (в этом исследовании — греческому) отличий между экспертами и ИИ не было [Fokides, Peristeraki, 2024]. В одном

из исследований при сравнении обратной связи, предоставленной ИИ и студентами-оценщиками, было отмечено, что обратная связь ИИ имеет более описательный характер [Banihashem et al., 2024]. ИИ, как правило, акцентировал внимание на описании задачи эссе и выполненных действиях, в то время как студенты в своей обратной связи лучше выявляли содержательные проблемы. Однако важно отметить, что ни одно из упомянутых исследований не подразумевало проведение «слепого» сравнения.

Существует мнение, что ИИ имеет ограниченные возможности в вопросах, связанных с эмоциональной поддержкой и мотивацией студентов. Преподаватели и студенты отмечали, что для более сложных аспектов обучения, таких как личностное развитие, саморазвитие и психологическая поддержка, они предпочитают человеческую обратную связь [Otaki, 2023]. Студентам также кажется, что обратная связь от ИИ не такая эмоциональная, и они относятся к ней менее серьезно, чем к комментариям, которые дает преподаватель. Согласно исследованию это связано с рядом причин:

- **эмпатия и поддержка**: преподаватели не просто сообщают информацию, но и выражают поддержку, развивают мотивацию и уверенность студентов, что редко содержится в обратной связи от ИИ;
- **искренность и мотивация**: студенты ценят, когда преподаватель выражает искреннее восхищение их усилиями, что играет важную роль в поддержании мотивации;
- **эффект личного взаимодействия**: личные отношения с преподавателем усиливают восприятие обратной связи; комментарии человека, которого уважают, оказывают больший эффект, чем безличные замечания ИИ.

Однако в ряде исследований отмечается, что обратная связь, предоставляемая ИИ, в некоторых случаях оказывается более поддерживающей, чем обратная связь преподавателя. Так, в исследовании на примере эссе для учащихся начальной школы [Fokides, Peristeraki, 2024] ИИ предоставлял более объемную и поддерживающую обратную связь по сравнению с экспертами, которые в большей степени фокусировались на механическом указании ошибок. Существует исследование, где ИИ обучали предоставлять поддерживающую обратную связь [Lin et al., 2024]. Согласно мнению авторов, начинающие преподаватели не всегда могут давать обратную связь в вежливой и корректной форме, а обученная модель ИИ позволяла корректиро-

вать обратную связь преподавателя и формулировать ее в более поддерживающей манере.

3.4. Выводы по главе

Эффективность обратной связи зависит от ее своевременности, детальности, персонализированности, но подготовка такой обратной связи трудоемка и времязатратна. Современные технологии ИИ помогают решить эту проблему на разных уровнях образования как для учебных дисциплин, так и для творческих занятий. Использование ИИ может быть особенно эффективно, например, в контексте онлайн-обучения, когда в когорте учащихся может быть по 300-400 человек, и традиционная обратная связь просто невозможна в более-менее развернутом формате и в адекватные сроки. ИИ также упрощает персонализацию обратной связи, что полезно и для учащихся, поскольку они получают персонализированные подсказки, и для преподавателей, которым ИИ предоставляет информацию для диагностики трудностей в обучении и прогнозирования рисков. Важным применением ИИ может стать инклюзивное образование.

Обратная связь, генерируемая ИИ, по справедливости, детальности и своевременности не уступает, а иногда и превосходит качество обратной связи, которую дает преподаватель. Однако есть ряд ограничений: ИИ может предоставлять некорректную информацию, в том числе давать неполную или ошибочную обратную связь [Simkute et al., 2024], что может быть связано как с некорректным промптингом, так и с более системными ошибками ИИ. Использование ИИ может провоцировать нечестное поведение и формировать зависимость, при которой учащиеся (особенно дети) начинают всецело полагаться на ИИ, что ограничивает их автономию и агентность [Li et al., 2024].

Стоит отметить, что обратная связь от преподавателей является отражением отношений, складывающихся между ними и учащимися, и поэтому обладает как бóльшим потенциалом для поддержки, так и бóльшими рисками — именно потому, что ее предоставляет человек.

4. Анализ данных тестирования

Обработка, анализ и интерпретация данных тестирования являются ключевыми этапами в оценивании. Эти процессы позволяют не только оценивать качество инструментов, но и выявлять уровень сформированности оцениваемых навыков тестируемых, а также обобщенные паттерны и закономерности. Однако в условиях роста объема информации, включая данные из цифровой среды, такие задачи становятся все более трудоемкими и времязатратными. Для их решения может применяться искусственный интеллект, который способен проводить как количественный, так и качественный анализ данных. Применение ИИ при анализе данных предоставляет ряд преимуществ, среди которых:

- **автоматизация калибровки заданий**, что позволяет обеспечивать высокую точность и надежность результатов тестирования, особенно в ситуациях массовой оценки;
- **выявление мошенничества и аномалий в данных**, что значительно повышает достоверность результатов тестирования и качество инструмента измерения;
- **анализ процессных данных**, предоставляющий детальную информацию о поведении тестируемого, стратегиях решения для повышения надежности измерения и обогащения обратной связи;
- **прогнозирование результатов учащихся** на основе данных;
- **анализ данных разных типов**, включая количественные и качественные, в большом объеме.

4.1. Некоторые аспекты применения ИИ в анализе данных тестирования

В данном разделе будут представлены некоторые аспекты применения ИИ в анализе данных тестирования.

Калибровка заданий заключается в оценке параметров заданий (трудности, дискриминативности и др.) с использованием математических моделей (например, моделей современной теории тестирования) [Wainer, Mislevy, 2000]. Этот процесс необходим на этапе пилотных испытаний для выявления заданий, которые не соответствуют требованиям (например, слишком легкие или трудные) или имеют низкую дискриминативность, что делает их менее эффективными для

задачи различения тестируемых с высоким и низким уровнем способности. Информация о параметрах заданий, полученных по итогам калибровки, используется для создания банков заданий или адаптивных систем тестирования.

Однако калибровка заданий требует их предварительного тестирования на выборке участников. Это связано с риском раскрытия содержания заданий, что особенно проблематично для тестов с высокими ставками, где важна конфиденциальность материалов. Кроме того, банки заданий нуждаются в регулярном обновлении, чтобы сохранить актуальность заданий и соответствовать требованиям образовательных стандартов. Это, в свою очередь, требует периодического пилотного тестирования новых заданий, что может быть ресурсоемким и долгим процессом.

Технологии ИИ имеют перспективы для упрощения процесса калибровки без предварительного тестирования заданий или для сокращения объема пилотных испытаний. Например, в одном исследовании использовалась модель BERT-LLTM для оценки параметров заданий, которые не проходили пилотные испытания [McCarthy et al., 2021]. Результаты модели оказались сопоставимы с оценками, полученными через двухпараметрическую модель современной теории тестирования (2PL).

В работе [Pereira et al., 2024] исследовался потенциал генеративных моделей ИИ для автоматизации калибровки заданий на примере экзамена в Португалии. Модель была протестирована 100 раз на 150 вопросах с выбором ответа. Для каждого задания рассчитывались показатели трудности (процент решаемости) с использованием классической теории тестирования, которые затем сравнивались с фактическими результатами калибровки. Корреляция между предсказанными и реальными значениями трудности заданий оказалась умеренной ($r = 0.372$), что недостаточно, чтобы считать попытку удачной. Тем не менее, исследование подчеркивает перспективы использования генеративных моделей для упрощения и ускорения процесса калибровки тестовых заданий и сообщения результатов широкой аудитории, которая не имеет углубленных знаний в области измерений.

Технологии ИИ могут выступать помощником для решения иных специфических задач в области измерения и работы с данными тестирования. В работе [Qin, Guo, 2024] классификационные алгоритмы машинного обучения использовались для валидизации Q-матрицы в

моделях когнитивной диагностики IRT (CDM). Q-матрица определяет соответствие заданий и измеряемых ими атрибутов (характеристик). Например, в этой статье рассматривались задания в тесте по чтению международного исследования PISA со следующими атрибутами: извлечение информации, обобщение основной идеи, интерпретация прочитанного, формулирование выводов, их оценка. Качество результатов модели и надежности измерения напрямую зависит от правильности соотнесения задания и атрибутов. Как правило, Q-матрицу разрабатывают эксперты, однако они не всегда достигают высокой согласованности мнений, что отражается на качестве результатов оценки. Алгоритмы машинного обучения позволяют автоматически подобрать Q-матрицу, которая повысит точность диагностики.

Другой важный аспект анализа данных тестирования — *выявление аномальных ответов* (aberrant responses), связанных с мошенничеством в контексте тестирования с высокими ставками или небрежным выполнением теста при низких ставках. Модели современной теории тестирования без использования ИИ оснащены инструментами для выявления подозрительных или невнимательных ответов. Например, используются статистики для выявления несогласованности ответов респондентов (person fit) [Meijer, Sijtsma, 2001] или более сложные модели смеси распределений для выявления невнимательных ответов [Ulitzsch et al., 2022]. Однако ИИ может быть помощником в этом процессе, анализируя паттерны ответов участников для выявления аномальных схем поведения, таких как подозрительно похожие ответы у разных участников или быстрое выполнение теста, что может свидетельствовать о мошенничестве [Alsabhan, 2023]. Также в литературе описаны случаи выявления мошенничества на экзамене на основе предыдущих результатов [Meng, Ma, 2023].

С ростом популярности компьютерного тестирования появилась возможность анализировать не только итоговые ответы респондентов, но и то, как они решают задания. Информация, собираемая в процессе решения задачи, получила название *процессные данные* (process data). К процессным данным относят время выполнения задания, количество нажатий на элементы экрана, последовательность действий, движение курсора мыши и другое. Базовые модели современной теории тестирования разрабатывались для анализа простых типов данных тестирования: бинарные (дихотомические задания 0/1) или порядковые (задания с частично верным ответом: 0/1/2). Несмо-

тря на то, что современные модели IRT включают анализ процессных данных [Liu et al., 2018], очевиден потенциал ИИ для анализа большого объема неструктурированных данных, снимаемых в цифровой среде. Например, майнинг данных (data mining) использовался для анализа поведения тестируемого в онлайн-курсах [Arpasat et al., 2021] или в интерактивных цифровых заданиях [Ulitzsch et al., 2022].

К процессным данным также относят данные движения глаз во время компьютерного тестирования. В работе [Ndiaye et al., 2023] проведен обзор применения технологий айтрекинга (отслеживания движения глаз, eye-tracking) и искусственного интеллекта для оценки компетенций в инженерном образовании. Авторы отмечают, что технологии ИИ имеют потенциал для улучшения процесса визуального отслеживания, а также используются для анализа данных айтрекинга. В контексте образования информация о движении глаз может помочь выявить сложности в восприятии материала тестирования, отслеживать когнитивную нагрузку, выявлять случаи нечестного выполнения тестирования.

Развивая направление использования данных тестирования, отметим, что на их основе можно делать *образовательные предсказания*, которые стали более доступными с появлением методов ИИ. Наиболее частые направления предсказания — академическая успеваемость, окончание учебы в образовательной организации, попадание в группу риска, академическая вовлеченность и т.д. [Pelima et al., 2024]. Для построения предсказательных моделей используют разные предикторы: от характеристик самого учащегося (поведение, академические показатели, семейный бэкграунд) [там же] до коллатеральной информации, получаемой из цифровой среды. Например, результаты формирующего оценивания учащихся и время, затраченное на выполнение заданий, были определены как сильные предикторы итоговых оценок этих учащихся [Jiao et al., 2022]. Сегодня модели ИИ используются для предсказания принадлежности студента к той или иной группе: какую оценку получит студент [например, Qu et al., 2018], будет ли студент исключен из образовательного учреждения [Figueroa-Cañas, Sancho-Vinuesa, 2020], является ли студент прокрастинатором [Akram et al., 2019], и т.д. Существующие исследования показывают высокое качество предсказательной способности моделей [Qu et al., 2018], однако исследователи выделяют ряд ограничений таких исследований, касающихся используемых данных, ограничений самих алгоритмов, негенерализуемости и неадаптивности моделей, а также этической составляющей [Pelima et al., 2024].

ИИ активно используется для *количественного и качественного анализа данных.* Он способствует эффективному извлечению и интерпретации ключевых показателей, таких как описательная статистика, анализ распределений, регрессионные модели и корреляционный анализ [Combrinck, 2024], а также упрощает процесс подготовки данных для дальнейшего анализа [Mumuni, Mumuni, 2024].

В работе Combrinck (2024) предлагается методология использования генеративного ИИ в смешанном анализе данных, объединяющем количественный и качественный подходы. В рамках количественного анализа ИИ можно применять на всех этапах, начиная с создания рандомизированной базы данных и заканчивая построением таблиц, графиков и интерпретацией результатов статистического анализа. Например, Combrinck (2024) предлагает следующие промпты, чтобы:

- получить рекомендации по видам анализа собранных данных: «У меня есть данные исследования. Пожалуйста, подскажи, какой тип анализа можно применить к этим данным. Переменные включают [демографические вопросы, такие как пол, возраст и раса]. Измеряемые конструкты в опроснике включают [мотивацию посещения университета и долгосрочные карьерные цели]. [Справочная информация о переменных вставляется в промпт]»;
- провести анализ и сформулировать выводы на основе данных: «На основе ранее предоставленных данных проведите статистический анализ, определите значимые элементы результатов, добавьте p-значения и размеры эффекта. Представьте результаты в виде таблицы с кратким изложением интерпретации».

ИИ также применяется в качественном анализе данных. Исследователи отмечают, что он может предложить теоретическую и аналитическую рамку для исследования, выполнить кодирование текстовых данных и провести нарративный анализ [Combrinck, 2024].

Тем не менее, хотя ИИ успешно справляется с задачами описательной статистики и базового статистического анализа, его способность к более сложным методам, таким как моделирование структурных уравнений, требует дальнейших исследований и верификации [там же].

4.2. Выводы по главе

Анализ данных, полученных в результате тестирования, играет ключевую роль в оценивании. Современные технологии искусствен-

ного интеллекта предоставляют исследователям и аналитикам образования инструменты для автоматизированного и быстрого анализа данных, охватывая широкий спектр задач — от психометрических исследований до прогностического моделирования. Несмотря на значительные достижения, результаты анализа, выполненного ИИ, требуют критического осмысления, так как многие исследования не опираются на устоявшиеся теории измерений. Перспективным направлением исследования является синтез технологий ИИ с теоретически обоснованными подходами к оцениванию.

5. Этические вопросы использования ИИ в оценивании

Очевидно, что общественные последствия использования ИИ требуют скорейшего и тщательного рассмотрения. Для решения этих задач в последние годы были созданы целые новые организации и сообщества, такие как AI Now, Partnership on AI и IEEE Global Initiative on the Ethics of Autonomous and Intelligent Systems (Глобальная инициатива по этике автономных и интеллектуальных систем). Появились даже специализированные академические конференции, такие как AAAI/ACM AI, Ethics, and Society (ИИ, этика и общество) или Fairness, Accountability, and Transparency (ACM FAccT, Справедливость, подотчетность и прозрачность) [Schiff, 2022].

В 2021 г. ЮНЕСКО приняла первый глобальный стандарт по этическим аспектам искусственного интеллекта. Рекомендации представляют собой нормативную основу, общую для 194 членов организации [UNESCO, 2022], направленную на обеспечение того, чтобы технологии ИИ приносили пользу человечеству, минимизируя потенциальные риски. Выделены четыре ключевые ценности, определяющие использование ИИ: уважение к правам человека, многообразие и инклюзивность, устойчивое развитие и защита окружающей среды, обеспечение благополучия человечества.

5.1. Классификация этических проблем использоваания ИИ в оценивании

Постараемся немного углубиться в этот вопрос и понять, как исследователи классифицируют этические проблемы, возникающие с развитием и использованием ИИ именно в оценивании и образовании. Например, в недавней статье в журнале «Nature» авторы делят этические вызовы, стоящие перед ИИ, на следующие составляющие: вопросы алгоритмической предвзятости и дискриминации, конфиденциальности, прозрачности и подотчетности в образовательных системах, управляемых искусственным интеллектом [Al-Zahrani, Alasmari, 2024].

Алгоритмическая предвзятость в моделях ИИ представляет собой сложную этическую проблему, связанную с различными источниками искажений [Radford et al., 2019]. Одним из ключевых факторов является предвзятость данных, на которых модель обучается. Обучение LLM проходит два основных этапа: предварительное обучение (pre-training) на масштабных корпусах текстов и последующую тонкую настройку (fine-tuning) на более специализированных данных [Liu et al., 2023]. На каждом из этих этапов могут возникать свои типы искажений. Во-первых, предвзятость может возникнуть из исходных наборов данных, если в них преобладают определенные точки зрения, культурные представления или стереотипы [Bender et al., 2021]. Во-вторых, предвзятость может усиливаться или модифицироваться на этапе тонкой настройки, если выбор данных или принципы дообучения отражают определенные предпочтения разработчиков [Gallegos et al., 2024]. В-третьих, влияние оказывает и сам механизм генерации: модели могут по-разному интерпретировать запросы пользователей, адаптируясь к уже существующим языковым паттернам, что может приводить к непреднамеренному воспроизведению предвзятых формулировок [Li et al., 2024].

При создании заданий для целей оценивания алгоритмическая предвзятость может проявляться сразу в нескольких аспектах. LLM может генерировать формулировки, содержащие незнакомые или неподходящие термины для определенных групп учащихся, что затрудняет интерпретацию заданий и снижает их валидность. Кроме того, в заданиях могут неявно воспроизводиться социальные предубеждения, связанные с полом, расой, возрастом, религией, национальностью и другими характеристиками, что ставит некоторых учащихся в неравные условия. Помимо этого, модели могут демонстрировать системную предвзятость в интерпретации сложных понятий, опираясь на доминирующие нарративы в исходных данных, что ограничивает возможность учета альтернативных точек зрения и культурных контекстов.

Один из описанных в научной литературе примеров, хоть и не из области оценивания, но весьма наглядный — онлайн-эксперимент, показавший, что ИИ-алгоритмы в Facebook демонстрируют дискриминационный таргетинг [Cecere et al., 2024]. В частности, авторы описали исследование алгоритмов предъявления объявлений о найме, где выяснилось, что 91% пользователей, которым показывали вакансии автомехаников, — мужчины, а 79 % аудитории, видевшей объ-

явления о вакансиях для воспитателей детских садов, — женщины. Авторы подчеркивают наличие алгоритмической гендерной предвзятости, проявляющейся в различных секторах, особенно в профессиях в области науки, технологий, инженерии и математики (STEM). Тем не менее, авторы указывают на возможность разработки методов аудита алгоритмической предвзятости. Также в их исследовании показаны возможные настройки алгоритмических решений в условиях, когда прозрачность и подотчетность алгоритмов затруднены, как чаще всего и происходит на практике.

Общие рекомендации, которые различные авторы предлагают для решения этой этической проблемы, включают прежде всего следующее:

• Важно, чтобы модели ИИ обучались на более широком спектре контекстов, что снижает вероятность создания предвзятого контента. В частности, должны использоваться гетерогенные и сбалансированные наборы данных, отражающие многообразие социально-культурных характеристик.

• Важно выстраивать обучение моделей на принципах справедливости (Fairness-aware Learning) — применять специальные методы вмешательств, которые корректируют дисбаланс как в данных для обучения, так и в выходных результатах моделей [Zhang, 2024].

• Применительно к оцениванию в образовании процесс генерации заданий с использованием ИИ должен включать обязательную экспертизу заданий экспертами (принцип «human in the loop»). Сгенерированные задания могут быть скорректированы в процессе экспертизы, что позволит смягчить предвзятость, например, путем обнаружения и замены вредных или неподходящих слов [Gallegos et al., 2024].

• Кроме того, очевидна необходимость во внедрении тщательного аудита самих ИИ-систем, используемых в образовании, чтобы как можно раньше выявлять возможные проблемы, связанные с предвзятостью алгоритмов [Cecere et al., 2024].

• Наконец, собственно создание и соблюдение этических кодексов в области использования и разработки ИИ вообще, и в оценивании в образовании в частности, поможет установить рамки для справедливого и ответственного функционирования ИИ-технологий [UNESCO, 2022].

Прозрачность алгоритмов — этическая проблема, тесно связанная с проблемой предвзятости. Каковы этические последствия не-

возможности легко проанализировать, каким образом ИИ принимает решения («черный ящик») [Miao et al., 2021]? Важно обеспечить понимание принципов работы ИИ для участников процесса оценивания, а это, в свою очередь, будет способствовать и решению обозначенной проблемы предвзятости и ошибок в оценках.

ИИ-технологии непрерывно совершенствуются, происходит постоянная оптимизация параметров моделей, с которыми взаимодействуют пользователи, и этот процесс становится все более непрозрачным. Неспособность человека понять, как именно большая языковая модель приходит к тому или иному выводу, отвечая на его вопрос, является сегодня характерной чертой взаимодействия людей и ИИ [Illia et al., 2023].

Однако непрозрачность ИИ-алгоритмов обусловлена не только технической сложностью машинного обучения, но и распределенной моральной ответственностью между различными участниками процесса разработки ИИ [Floridi, Chiriatti, 2020]. Такая распределенная ответственность затрудняет формулирование конкретных и адресных общественных ожиданий в отношении этических принципов, определяющих поведение ИИ. Распространенные рекомендации по повышению прозрачности можно сформулировать следующим образом:

• Чтобы уменьшить непрозрачность, некоторые исследователи предлагают возложить ответственность за нее на бизнес, предлагающий ИИ-услуги и продукты, и создателей ИИ-алгоритмов [Illia et al., 2023]. Представляется вполне ожидаемым, что любая компания, создающая ИИ-продукт, несет ответственность за использование продукта, который должен работать корректно и в соответствии с замыслом.

• Снижать непрозрачность можно и повышением образованности конечных пользователей. Если пользователи ИИ, например, преподаватели, учащиеся и родители, будут иметь базовое понимание того, как работают ИИ-системы в образовании, они смогут принимать более информированные решения и критически оценивать результаты работы ИИ, в том числе сообщать разработчикам о фактах ошибок, дискриминации или предвзятости. Также необходимо разрабатывать способы, позволяющие объяснять решения ИИ-систем понятным для рядовых пользователей образом, например, включать визуализацию данных, предоставлять технически неподкованному читателю более адаптированные подробные отчеты о факторах, влияющих на решения ИИ [Сберобразование].

Академическая честность. Какие педагогические подходы в работе с ИИ являются этически оправданными? Как быть с плагиатом [Dehouche, 2021]? Эти вопросы касаются использования ИИ как преподавателями, так и учащимися, и в оценке, и в общей практике.

Так, часто упоминаемым примером критики в адрес ИИ в образовании является ИИ-ассистент «Джилл Уотсон», разработанный в Технологическом институте Джорджии (США), который сортировал сообщения на форумах и отвечал на вопросы, где это возможно (например, «Когда мне нужно сдать задание?»), направляя более сложные вопросы ассистентам преподавателей. Этот ИИ-ассистент был основан на платформе IBM Watson [Goel, Polepeddi, 2018]. Критика касалась того, что студентов как будто обманывали, заставляя думать, что ИИ-ассистент был реальным человеком. К примеру, ответы давались с задержкой, некоторые реплики содержали юмор.

Другая сторона вопроса связана с тем, что учащиеся тоже могут обманывать с помощью ИИ, — как в ситуациях обычного внутриклассного оценивания, так и в оценивании с более высокими ставками. Например, учащиеся могут использовать генеративный ИИ, чтобы готовить эссе, рефераты, да и любые другие письменные или творческие задания и выдавать их за собственную работу. Это подрывает саму суть образования и может привести к девальвации дипломов. Отличить собственную работу учащегося от ответов, сгенерированных ИИ, может быть непросто. Однако исследователи и практики пытаются формулировать стратегии, которые могут в этом помочь. Например, в работе с говорящим названием «Chatting and cheating…» (чатинг и читинг) авторы дают следующие рекомендации [Cotton et al., 2024]:

• Знакомить учеников с тем, что такое плагиат. Одним из наиболее эффективных способов предотвращения плагиата является разъяснение ученикам, что такое плагиат и почему он недопустим в обучении.

• Заранее требовать предоставлять черновики проверочных и других оценочных работ. Обязательное предоставление черновых версий работ до финальной сдачи позволяет преподавателям выявить возможные признаки использования ИИ и дать студентам обратную связь для улучшения работы.

• Использовать специальные инструменты для обнаружения плагиата/выдачи текста, написанного генеративной моделью, за свой. Существует множество институциональных инструментов, которые помогают выявлять случаи классического плагиата в студенческих ра-

ботах. Особенно технически продвинутые преподаватели также могут рассмотреть возможность использования продвинутых технологий и методов, например, применять алгоритмы обработки естественного языка для анализа стиля и языка работы и выявления аномалий, указывающих на использование ChatGPT или других ИИ-моделей. К примеру, GPT-2 Output Detector Demo разработан для анализа текста и выявления аномалий, свидетельствующих о том, что текст мог быть сгенерирован ИИ.

• Анализировать языковые паттерны и аномалии. В целом, можно посмотреть текст «невооруженным глазом». Чат-боты часто имеют ограничения в своих языковых возможностях и могут генерировать текст с повторяющимися фразами, словами или некорректным и непоследовательным использованием языка. Анализ этих особенностей может помочь определить, был ли текст сгенерирован ИИ.

• Проверять на фактические ошибки. Хотя языковые модели ИИ способны создавать связные тексты, содержащаяся в них информация не всегда является достоверной.

• Включать в задание конкретный контекст. Наконец, тексты, созданные людьми, как правило, более контекстуальны, чем созданные ИИ, которые могут быть более общими и менее адаптированными к конкретной ситуации или требованию, изложенному в задании.

Конфиденциальность данных. Этические вопросы конфиденциальности при использовании ИИ в оценке в образовании, отраженные в рекомендациях ЮНЕСКО, очерчивают довольно широкий круг проблем [Miao et al., 2021]. Как временный характер интересов и эмоций учеников, а также сложность процесса обучения влияют на интерпретацию данных и этические аспекты применения ИИ в образовательных контекстах? Какие критерии следует учитывать при определении и обновлении этических границ в сборе и использовании данных об учащихся? Как школы, ученики и учителя могут отказаться от участия или оспорить свое «присутствие» в больших наборах данных?

Очевидно, что широкое внедрение технологий ИИ влечет за собой многочисленные риски и вызовы, связанные с правом собственности на данные (например, эксплуатация данных в коммерческих целях) или согласием на предоставление данных (например, способны ли учащиеся дать действительно осознанное согласие на участие в процедурах, напрямую или косвенно связанных с обучением или другим использованием их данных ИИ) [Bulut, Wongvorachan, 2022; Miao et al., 2021].

Вопросы предоставления данных ИИ тесно связаны с моральными дилеммами, поскольку искажения в работе алгоритмов ИИ могут подрывать основные права человека. В частности, применение ИИ в оценивании в образовании подвергается критике за свою инвазивность. Инвазивность заключается в том, что некоторые приложения требуют постоянного мониторинга и распознавания действий, жестов и эмоций учащихся, что может снижать и степень самостоятельности учащихся, и их удовлетворенность обучением, и чего они могут просто не хотеть [Miao et al., 2021].

Так, в некоторых школах камеры, управляемые ИИ, используются для мониторинга поведения учащихся [Loizos, 2017], и это, как представляется, явно нарушает этические границы. В цитируемом исследовании сообщается, что в систему наблюдения встроена технология распознавания лиц, чтобы проверить, насколько внимательно ученики ведут себя на уроке. Каждое движение учеников фиксируется множеством камер, установленных над доской. Система работает, определяя выражения лиц и передавая эту информацию компьютеру для оценки того, сосредоточены ли ученики или отвлеклись. В одном из примеров компьютер анализирует семь различных эмоций: нейтральную, счастливую, грустную, разочарованную, сердитую, испуганную и удивленную. Если система решает, что ученик отвлекся, она отправляет уведомление учителю для принятия мер. Однако такие камеры повышают уровень тревожности и изменяют естественное поведение учеников. Ученики сообщили, что чувствуют, будто за ними постоянно следят невидимые глаза.

В другом недавнем исследовании [Hossen, Uddin, 2023] авторы также описывают успешно внедренную ими систему наблюдения за учащимися на онлайн-занятиях. Система включает различные возможности аутентификации пользователей, в частности, обнаружение лиц, отслеживание рук, распознавание мобильных телефонов и модули оценки позы. Используя эти компоненты, исследователи извлекают чрезвычайно обширные данные, которые используют для обучения моделей машинного обучения. Их цель — опять же оценка уровня внимания учащихся во время онлайн-занятий. Система генерирует отчет о внимательности учащихся, доступный через специализированную веб-страницу, которая включает сводку поведения учащихся на протяжении онлайн-уроков. И хотя авторы подчеркивают, что отчет анонимный, этические вопросы к приемлемости такой ситуации оста-

ются. Возможно, что именно подобные ситуации вызывают у пользователей закономерный вопрос, а не является ли происходящее на наших глазах тем, что отражено в антиутопических произведениях — от романа «Мы» Замятина до, например, известного японского анимационного сериала «Психопаспорт».

Часть обозначенных аспектов проблемы конфиденциальности только предстоит решить. Однако многое могут сделать обычные пользователи уже сейчас.

• Например, очень важно соблюдать базовые правила защиты информации, если педагог анализирует данные учеников, или исследователь — данные респондентов. Важно помнить об этом, так как ответы тестируемых могут содержать личную информацию, и поэтому всем категориям пользователей ИИ в образовании — родителям, учителям и детям, исследователям — нужно обучиться основам защиты персональных данных и безопасного использования ИИ.

• Рекомендуется воздерживаться от загрузки, например, в чатботы, любых конфиденциальных данных. Прежде чем использовать генеративный ИИ для анализа данных, важно убедиться, что вся идентифицирующая информация из наборов данных удалена.

Кроме того, важно подтверждать наши — пользователей — требования к организациям, которые разрабатывают ИИ. По мнению исследователей, они должны включать в себя, например, следующее [Vasa, Thakkar, 2023]:

• разделять данные — обучать модели на синтетических данных или обезличенных наборах, по возможности избегая работы с оригинальными персональными данными;

• дифференцировать приватность — добавлять случайный шум в данные перед обучением, чтобы невозможно было восстановить исходные записи;

• фильтровать данные — проверять наборы данных на наличие потенциально чувствительных сведений перед их использованием в обучении моделей.

С точки зрения хранения данных, в том числе данных о несовершеннолетних, важно требовать от компаний, производящих и в том или ином виде работающих с ИИ, выполнение следующего [Zhang et al., 2024]:

• избегать, где это возможно, длительного хранения данных и внедрять автоматические механизмы удаления данных после истечения срока хранения;

• обеспечивать пользователям — детям и их родителям — право на удаление (Right to be Forgotten), то есть возможность удаления своих данных по запросу;

• разрабатывать процедуры для независимого контроля за использованием данных несовершеннолетних.

5.2. Выводы по главе

Круг этических вызовов, с которыми сталкиваются пользователи ИИ в образовании, не исчерпывается описанными проблемами. Но мы обсудили те из них, которые упомянуты в большинстве рассмотренных нами академических источников. Обозначенные этические проблемы важно анализировать и пытаться решать, причем решать на уровне повседневных действий каждого пользователя ИИ.

Несмотря на то что рассмотренные нами рекомендации по преодолению или смягчению этических вызовов в области использования ИИ носят довольно общий характер, они находят свое отражение и углубление в нормативных актах организаций в различных сферах. Так, в июне 2024 г. ученый совет НИУ ВШЭ утвердил Декларацию этических принципов создания и использования систем искусственного интеллекта[10], устанавливающую общие принципы, которыми следует руководствоваться преподавательскому и научному сообществу, администрации, студентам при создании и использовании ИИ. Университет рассматривает ИИ как средство обогащения процессов учебной, научной, экспертно-аналитической и административной деятельности и формулирует ряд принципов, которые должны соблюдать преподаватели, научные сотрудники, администрация и студенты при использовании ИИ для решения конкретных задач, критически оценивая получаемые результаты и возможные риски.

В Декларации подчеркиваются следующие основные принципы[11]:

• *принцип академической честности*: ИИ должен использоваться как дополнение, а не как замена естественного интеллекта;

• *принцип прозрачности*: от сотрудников и студентов НИУ ВШЭ требуется выделять результаты своей деятельности, в которой был использован ИИ, указывая характер и объем работ, выполненных с

[10] https://www.hse.ru//mirror/pubs/share/937054455.pdf.

[11] С полным перечнем принципов, утверждаемых в Декларации, можно ознакомиться по ссылке из предыдущей сноски.

его помощью, а также публично сообщать, когда используется ПО, имитирующее человеческое общение;

• *принцип конфиденциальности и соблюдения интеллектуальных прав*: разработка и использование ИИ должны обеспечивать защиту персональных данных;

• *принцип разумного ограничения*: студентам и сотрудникам в ходе учебного процесса и выполнения исследовательских работ можно использовать ИИ как полезный инструмент, однако системы оценивания и элементы контроля планируемых результатов обучения могут быть пересмотрены по мере развития ИИ, и университет оставляет за собой право вводить ограничения на разработку и использование ИИ.

Не только НИУ ВШЭ, но и другие университеты и организации в мире стараются определить свое видение перспектив развития и использования — в том числе с учетом этических аспектов — искусственного интеллекта, и это свидетельствует об ответственности и растущем понимании влияния ИИ на общество и академическую среду: так научное сообщество пытается адаптироваться к вызовам новой цифровой эпохи. Обсуждение вопросов этики использования ИИ в образовании, и в частности в оценивании, должно помочь сформировать у пользователей установки и отношения, позволяющие критически воспринимать новые, не лишенные потенциальных противоречий технологии, и осознанно подходить к их применению.

Заключение

Использование искусственного интеллекта (ИИ) в оценивании демонстрирует значительный потенциал для улучшения связанных с ним процесов. Разработка заданий, автоматическая оценка различных типов ответов, персонализированная обратная связь и анализ данных с применением ИИ открывают новые возможности для более объективного, точного и оперативного оценивания. Эти технологии помогают не только ускорить процессы, но и повысить их адаптивность и гибкость, делая оценивание более справедливым и инклюзивным.

Однако с внедрением ИИ возникают и новые вызовы, связанные с этическими аспектами. Прозрачность алгоритмов, конфиденциальность данных, предотвращение предвзятости, а также обеспечение правильного понимания роли ИИ в оценивании — все это требует серьезного внимания и разработки этических норм и стандартов. Важно, чтобы технологии не заменяли человеческое суждение, а дополняли его, сохраняя центральную роль преподавателей и экспертов в учебном процессе.

Таким образом, интеграция ИИ в образовательное оценивание открывает перспективы, которые могут значительно трансформировать процессы обучения и оценки. Однако успех этих изменений зависит от правильного и ответственного использования технологий, где этические и практические аспекты играют ключевую роль.

Литература

Азбель А.А., Илюшин Л.С., Казакова Е.И., Морозова П.А. Отношение учеников и учителей к обратной связи: противоречия и тенденции развития // Образование и наука. 2022. Т. 24. № 7. С. 76–109.

Азбель А.А., Илюшин Л.С., Морозова П.А. Обратная связь в обучении глазами российских подростков // Вопросы образования. 2021. № 1. С. 195–212.

Лапошина А.Н., Лебедева М.Ю. Текстометр: онлайн-инструмент определения уровня сложности текста по русскому языку как иностранному // Русистика. 2021. Т. 19. № 3. С. 331–345.

Сберобразование. Текущее состояние и перспективы ИИ в образовании. 025courses.sberuniversity.ru/ai-education (дата обращения: 28.02.2025).

Угланова И.Л., Гельвер Е.С., Тарасов С.В., Грачева Д.А., Вырва Е.Е. Оценивание креативности на основе анализа изображений с помощью нейронных сетей // Искусственный интеллект и принятие решений. 2021. № 1. С. 86–97.

Экопси. Echo: Автоматическая оценка кандидатов на массовые позиции. 2024. https://digital.ecopsy.ru/products/echo?utm_source=telegram-digital&utm_medium=post&utm_campaign=30.09.2024 (дата обращения: 30.09.2024).

Agarwal R. et al. Many-shot in-context learning // arXiv preprint arXiv:2404.11018. 2024.

Agbavor F., Liang H. Artificial Intelligence-enabled end-to-end detection and assessment of alzheimer's disease using voice // Brain sciences. 2022. Vol. 13. No. 1. P. 28.

Agostini D. et al. Large language models for sustainable assessment and feedback in higher education: Towards a pedagogical and technological framework // CEUR WORKSHOP PROCEEDINGS. CEUR Workshop Proceedings. 2024.

Akram A. et al. Predicting students' academic procrastination in blended learning course using homework submission data // Ieee Access. 2019. Vol. 7. P. 102487–102498.

Al-Ansi A.M. et al. Analyzing augmented reality (AR) and virtual reality (VR) recent development in education // Social Sciences & Humanities Open. 2023. Vol. 8. No. 1. P. 100532.

Al Balushi J.S.G. et al. Incorporating Artificial Intelligence Powered Immersive Realities to Improve Learning using Virtual Reality (VR) and Augmented Reality (AR) Technology / 2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC). IEEE, 2024. P. 760–765. https://doi.org/10.1109/ICAAIC60222.2024.10575046.

Al-Zahrani A.M., Alasmari T.M. Exploring the impact of artificial intelligence on higher education: The dynamics of ethical, social, and educational implications // Humanities and Social Sciences Communications. 2024. Vol. 11. No. 1. P. 1–12.

Alsabhan W. Student cheating detection in higher education by implementing machine learning and LSTM techniques // Sensors. 2023. Vol. 23. No. 8. P. 4149. https://doi.org/10.3390/s23084149.

Ariely M., Nazaretsky T., Alexandron G. Machine learning and Hebrew NLP for automated assessment of open-ended questions in biology //International journal of artificial intelligence in education. 2023. Vol. 33. No. 1. P. 1–34. https://doi.org/10.1007/s40593-021-00283-x.

Arpasat P. et al. Applying process mining to analyze the behavior of learners in online courses // International Journal of Information and Education Technology. 2021. Vol. 11. No. 10. P. 436–443.

Attali Y. et al. The interactive reading task: Transformer-based automatic item generation // Frontiers in Artificial Intelligence. 2022. Vol. 5.

Banihashem S.K. et al. Feedback sources in essay writing: peer-generated or AI-generated feedback? // International Journal of Educational Technology in Higher Education. 2024. Vol. 21. No. 1. P. 23.

Banville F. et al. Validation of a sorting task implemented in the virtual multitasking task-2 and effect of aging / Human Interface and the Management of Information. Information in Applications and Services: 20th International Conference, HIMI 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15-20, 2018. Proceedings, Part II 20. Springer International Publishing, 2018. P. 41–54. https://doi.org/10.1007/978-3-319-92046-7_4.

Becerra Á. et al. A generative AI-based personalized guidance tool for enhancing the feedback to MOOC learners / 2024 IEEE Global Engineering Education Conference (EDUCON). IEEE, 2024. P. 1–8.

Becker J. et al. Text generation: A systematic literature review of tasks, evaluation, and challenges // arXiv preprint arXiv:2405.15604. 2024.

Beliaeva A.Y., Yusupova E.M., Talov D.P. The agreement of neural-based and feature-based models for autoscoring students' written answers // Assessing Writing. 2024.

Bender E. M. et al. On the dangers of stochastic parrots: Can language models be too big? / Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. 2021. P. 610–623.

Bezirhan U., von Davier M. Automated reading passage generation with OpenAI's large language model // Computers and Education: Artificial Intelligence. 2023. Vol. 5. P. 100161.

Bhandari S., Liu Y., Pardos Z. A. Evaluating ChatGPT-generated Textbook Questions using IRT / Proceedings of the Generative AI for Education Workshop (GAIED) at the Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS). New Orleans, LA. 2023.

Botelho A. et al. Leveraging natural language processing to support automated assessment and feedback for student open responses in mathematics // Journal of computer assisted learning. 2023. Vol. 39. No. 3. P. 823–840.

Bozkurt A., Sharma R.C. Generative AI and prompt engineering: The art of whispering to let the genie out of the algorithmic world // Asian Journal of Distance Education. 2023. Vol. 18. No. 2. P. i-vii. https://www.asianjde.com/ojs/index.php/AsianJDE/article/view/749.

Brohi S.N. et al. Accuracy comparison of machine learning algorithms for predictive analytics in higher education / Emerging Technologies in Computing: Second International Conference, iCETiC 2019, London, UK. August 19–20, 2019. Proceedings 2. Springer International Publishing, 2019. P. 254–261.

Brown T.B. et al. Language models are few-shot learners // arXiv preprint arXiv:2005.14165. 2020. Vol. 1.

Bulut O. et al. The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges // arXiv preprint arXiv:2406.18900. 2024.

Bulut O., Wongvorachan T. Feedback generation through artificial intelligence // The Open / Technology in Education, Society, and Scholarship Association Conference. 2022. Vol. 2. No. 1. P. 1–9.

Bulut O., Yildirim-Erbasli S.N., Gorgun G. Assessment Analytics for Digital Assessments Identifying, Modeling, and Interpreting Behavioral Engagement / Assessment Analytics in Education: Designs, Methods and Solutions. Cham: Springer International Publishing, 2024. P. 35–60. doi: 10.1007/978-3-031-56365-2 3.

Burgues M., Goujet R., Zaraik J. Learning Soft Skills with an AI-Based Simulation Role-Play: A Literature Review / EDULEARN24 Proceedings. Palma, 2024. P. 6285–6293. doi: 10.21125/edulearn.2024.1484.

Burstein J. et al. A theoretical assessment ecosystem for a digital-first assessment — The Duolingo English Test // DRR-21-04. 2021.

Carless D. Feedback as dialogue / Encyclopedia of educational philosophy and theory. 2016. P. 1–6.

Cathoven. Free text difficulty analyzer. https://www.cathoven.com/free-text-difficulty-analyzer/.

Cavalcanti A P. et al. Automatic feedback in online learning environments: A systematic literature review // Computers and Education: Artificial Intelligence. 2021. Vol. 2. P. 100027.

Cecere G. et al. Artificial intelligence and algorithmic bias? Field tests on social network with teens // Technological Forecasting and Social Change. 2024. Vol. 201. P. 123204.

Çınar A. et al. Machine learning algorithm for grading open-ended physics questions in Turkish // Education and information technologies. 2020. Vol. 25. No. 5. P. 3821–3844.

Combrinck C.A tutorial for integrating generative AI in mixed methods data analysis // Discover Education. 2024. Vol. 3. No. 1. P. 116. https://doi.org/10.1007/s44217-024-00214-7.

Cotton D.R.E., Cotton P.A., Shipway J.R. Chatting and cheating: Ensuring academic integrity in the era of ChatGPT // Innovations in education and teaching international. 2024. Vol. 61. No. 2. P. 228–239.

Cropley D.H., Marrone R.L. Automated scoring of figural creativity using a convolutional neural network / Psychology of Aesthetics, Creativity, and the Arts. 2022. https://doi.org/10.1037/aca0000510.

Darwish S.M., Mohamed S.K. Automated essay evaluation based on fusion of fuzzy ontology and latent semantic analysis / The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019) 4. Springer International Publishing, 2020. P. 566–575. https://doi.org/10.1007/978-3-030-14118-9_57.

Davis B. Methods and system for reducing implicit bias with virtual environments : [заяв. пат.] 16168460 США. 2019. https://patents.google.com/patent/US20190369837A1/en.

Davison S.M.C., Deeprose C., Terbeck S. A comparison of immersive virtual reality with traditional neuropsychological measures in the assessment of executive functions // Acta Neuropsychiatrica. 2018. Vol. 30. No. 2. P. 79–89.

Dehouche N. Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3) // Ethics in Science and Environmental Politics. 2021. Vol. 21. P. 17–23. https://doi.org/10.3354/esep00195.

Devlin J. BERT: Pre-training of deep bidirectional transformers for language understanding // arXiv preprint arXiv:1810.04805. 2018. doi: 10.48550/arXiv.1810.04805.

Dong F., Zhang Y., Yang J. Attention-based recurrent convolutional neural network for automatic essay scoring // Proceedings of the 21st conference on computational natural language learning (CoNLL 2017). 2017. P. 153–162.

Dood A.J. et al. Automated Text Analysis of Organic Chemistry Students' Written Hypotheses // Journal of Chemical Education. 2024. Vol. 101. No. 3. P. 807–818.

Doughty J. et al. A comparative study of AI-generated (GPT-4) and human-crafted MCQs in programming education / Proceedings of the 26th Australasian Computing Education Conference. 2024. P. 114–123.

Erickson J.A. et al. The automated grading of student open responses in mathematics / Proceedings of the tenth international conference on learning analytics & knowledge. 2020. P. 615–624.

Estejab H., Bayramzadeh S. The Application of Augmented Reality in Simulation-Based Design Evaluations of Trauma Rooms // HERD. 2025. Vol. 18. No. 1. P. 70–85. doi: https://doi.org/10.1177/1937586724130241.

Figueroa-Cañas J., Sancho-Vinuesa T. Early prediction of dropout and final exam performance in an online statistics course // IEEE Revista Iberoamericana de Tecnologias del Aprendizaje. 2020. Vol. 15. No. 2. P. 86–94.

Floridi L., Chiriatti M. GPT-3: Its nature, scope, limits, and consequences // Minds and Machines. 2020. Vol. 30. P. 681–694.

Fokides E., Peristeraki E. Comparing ChatGPT's correction and feedback comments with that of educators in the context of primary students' short essays written in English and Greek / Education and Information Technologies. 2024. P. 1–45.

Gabbay H., Cohen A. Combining LLM-generated and test-based feedback in a MOOC for programming / Proceedings of the Eleventh ACM Conference on Learning@ Scale. 2024. P. 177–187.

Gallegos I.O. et al. Bias and fairness in large language models: A survey / Computational Linguistics. 2024. P. 1–79.

Garg S., Sharma S. Impact of artificial intelligence in special need education to promote inclusive pedagogy // International Journal of Information and Education Technology. 2020. Vol. 10. No. 7. P. 523–527.

Gao T., Fisch A., Chen D. Making pre-trained language models better few-shot learners // arXiv preprint arXiv:2012.15723. –2020.

Gierl M.J., Lai H. Automatic item generation / Handbook of test development. Routledge, 2015. P. 410–429.

Goel A.K., Polepeddi L. Jill Watson: A virtual teaching assistant for online education / Learning engineering for online education. Routledge, 2018. P. 120–143. https://smartech.gatech.edu/handle/1853/59104.

Grover M.S. et al. Multi-modal automated speech scoring using attention fusion // arXiv preprint arXiv:2005.08182. 2020.

Gupta S. et al. Towards Building a Language-Independent Speech Scoring Assessment / Proceedings of the AAAI Conference on Artificial Intelligence. 2024. Vol. 38. No. 21. P. 23200–23206.

Gwern. GPT-3 creative fiction & nonfiction writing. 2023. https://www.gwern.net/ GPT-3.

Haley D.T. et al. Measuring improvement in latent semantic analysis-based marking systems: using a computer to mark questions about HTML / In: S. Mann & Simon (Eds.), Proceedings of the 9th australasian conference on computing education, volume 66 of ACE. Ballarat: Australian Computer Society, 2007. P. 35–42.

Haughney K., Wakeman S., Hart L. Quality of feedback in higher education: A review of literature // Education Sciences. 2020. Vol. 10. No. 3. P. 60.

Hickman L., Tay L., Woo S.E. Are automated video interviews smart enough? Behavioral modes, reliability, validity, and bias of machine learning cognitive ability assessments // Journal of Applied Psychology. 2024. Vol. *110*(3). P. 314–335. https://doi.org/10.1037/apl0001236.

Higgins D. et al. Evaluating multiple aspects of coherence in student essays / Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004. 2004. P. 185–192.

Hossen M.K., Uddin M.S. Attention monitoring of students during online classes using XGBoost classifier // Computers and Education: Artificial Intelligence. 2023. Vol. 5. P. 100191.

Hsiao S.W. et al. Toward automating oral presentation scoring during principal certification program using audio-video low-level behavior profiles // IEEE Transactions on Affective Computing. 2017. Vol. 10. No. 4. P. 552–567.

Hua Zhen Lv. Innovative music education: Using an AI-based flipped classroom // Education and Information Technologies. 2023. Vol. 28. No. 11. P. 1–16. DOI:10.1007/s10639-023-11835-0.

Idrisoglu A. et al. Applied machine learning techniques to diagnose voice-affecting conditions and disorders: Systematic literature review // Journal of Medical Internet Research. 2023. Vol. 25. P. e46105.

Illia L., Colleoni E., Zyglidopoulos S. Ethical implications of text generation in the age of artificial intelligence // Business Ethics, the Environment & Responsibility. 2023. Vol. 32. No. 1. P. 20–210.

Irvine S. H., Kyllonen P. C. (ed.). Item generation for test development. Routledge, 2013.

Jeon J. Chatbot-assisted dynamic assessment (CA-DA) for L2 vocabulary learning and diagnosis // Computer Assisted Language Learning. 2023. Vol. 36. No. 7. P. 1338–1364.

Jiang L., Bosch N. Short answer scoring with GPT-4 / Proceedings of the Eleventh ACM Conference on Learning@ Scale. 2024. P. 438–442.

Jiao P. et al. Artificial intelligence-enabled prediction model of student academic performance in online engineering education // Artificial Intelligence Review. 2022. Vol. 55. No. 8. P. 6321–6344. https://doi.org/10.1007/s10462-022-10155-y.

Jurafsky D., Martin J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Online manuscript released August 20, 2024. https://web.stanford.edu/~jurafsky/slp3.

Kakarla S. et al. Using large language models to assess tutors' performance in reacting to students making math errors // arXiv preprint arXiv:2401.03238. 2024.

Kirkham R. et al. Immersive Virtual Reality–Based Methods for Assessing Executive Functioning: Systematic Review // JMIR Serious Games. 2024. Vol. 12. P. e50282. https://doi.org/10.2196/50282.

Kostic M. et al. LLMs in Automated Essay Evaluation: A Case Study // Proceedings of the AAAI Symposium Series. 2024. Vol. 3. No. 1. P. 143–147. https://doi.org/10.1609/aaaiss.v3i1.31193.

Koutsoumpis A. et al. Beyond traditional interviews: Psychometric analysis of asynchronous video interviews for personality and interview performance evaluation using machine learning // Computers in Human Behavior. 2024. Vol. 154. P. 108128.

Küchemann S. et al. Physics task development of prospective physics teachers using ChatGPT // arXiv preprint arXiv:2304.10014. 2023.

Lancaster T. Artificial intelligence, text generation tools and ChatGPT–does digital watermarking offer a solution? // International Journal for Educational Integrity. 2023. Vol. 19. No. 1. P. 10.

Landauer T.K., Foltz P.W., Laham D. An introduction to latent semantic analysis // Discourse processes. 1998. Vol. 25. No. 2–3. P. 259-284. doi:10.1080/01638539809545028.

Laverghetta Jr. A., Licato J. Generating better items for cognitive assessments using large language models / Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023). 2023. P. 414–428.

Li J. et al. Pre-trained language models for text generation: A survey // ACM Computing Surveys. 2024. Vol. 56. No. 9. P. 1-39.

Li Y., Zhou X., Chiu T.K.F. Systematics review on artificial intelligence chatbots and ChatGPT for language learning and research from self-determination theory (SDT): what are the roles of teachers? // Interactive Learning Environments. 2024. P. 1–15.

Liang Z. et al. Towards the automated generation of readily applicable personalised feedback in education / International Conference on Artificial Intelligence in Education. Cham: Springer Nature Switzerland, 2024. P. 75–88. https://doi.org/10.1007/978-3-031-64299-9_6.

Lin J. et al. How Can I Improve? Using GPT to Highlight the Desired and Undesired Parts of Open-ended Responses // arXiv preprint arXiv:2405.00291. 2024.

Liu H., Liu Y., Li M. Analysis of process data of PISA 2012 computer-based problem solving: Application of the modified multilevel mixture IRT model // Frontiers in psychology. 2018. Vol. 9. P. 1372.

Liu P. et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing // ACM Computing Surveys. 2023. Vol. 55. No. 9. P. 1–35. https://doi.org/10.1145/3560815.

Lo T. H. et al. An Effective Automated Speaking Assessment Approach to Mitigating Data Scarcity and Imbalanced Distribution // preprint arXiv:2404.07575. 2024.

Loizos C. AltSchool wants to change how kids learn, but fears have surfaced that it's failing students // TechCrunch. Recuperado en. 2017. Vol. 29.

Maier U., Klotz C. Personalized feedback in digital learning environments: Classification framework and literature review // Computers and Education: Artificial Intelligence. 2022. Vol. 3. P. 100080.

Marvin G. et al. Prompt engineering in large language models / International conference on data intelligence and cognitive informatics. Singapore: Springer Nature Singapore, 2023. P. 387–402.

Masikisiki B., Marivate V., Hlophe Y. Investigating the Efficacy of Large Language Models in Reflective Assessment Methods through Chain of Thought Prompting / Proceedings of the 4th African Human Computer Interaction Conference. 2023. P. 44–49.

Mathias S., Bhattacharyya P. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores / Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018). 2018.

McCarthy A.D. et al. Jump-starting item parameters for adaptive language tests / Proceedings of the 2021 conference on empirical methods in natural language processing. 2021. P. 883–899.

McNichols H. et al. Automated distractor and feedback generation for math multiple-choice questions via in-context learning. 2023.

McNichols H. et al. Can Large Language Models Replicate ITS Feedback on Open-Ended Math Questions? // arXiv preprint arXiv:2405.06414. 2024.

Mehta A. et al. Can ChatGPT play the role of a teaching assistant in an introductory programming course? // arXiv preprint arXiv:2312.07343. 2023.

Meijer R. R., Sijtsma K. Methodology review: Evaluating person fit // Applied psychological measurement. 2001. Vol. 25. No. 2. P. 107–135.

Meng H., Ma Y. Machine Learning–Based Profiling in Test Cheating Detection // Educational Measurement: Issues and Practice. 2023. Vol. 42. No. 1. P. 59–75. doi: 10.1111/emip.12541.

Miao F. et al. AI and education: A guidance for policymakers. Unesco Publishing, 2021.

Miao J. et al. Chain of thought utilization in large language models and application in nephrology // Medicina. 2024. Vol. 60. No. 1. P. 148.

Mikolov T. et al. Efficient estimation of word representations in vector space // arXiv preprint arXiv:1301.3781. 2013.

Minaee S. et al. Large language models: A survey // arXiv preprint arXiv:2402.06196. 2024.

Mishra S. et al. Reframing instructional prompts to GPTk's language // arXiv preprint arXiv:2109.07830. 2021.

Morris W. et al. Automated scoring of constructed response items in math assessment using large language models // International Journal of Artificial Intelligence in Education. 2025. Vol. 35. P. 559–586. https://doi.org/10.1007/s40593-024-00418-w.

Mumuni A., Mumuni F. Automated data processing and feature engineering for deep learning and big data applications: A survey // Journal of Information and Intelligence. 2024. doi: 10.1016/j.jiixd.2024.01.002.

Ndiaye Y., Lim K. H., Blessing L. Eye tracking and artificial intelligence for competency assessment in engineering education: a review // Frontiers in Education. Frontiers Media SA, 2023. Vol. 8. P. 1170348. DOI: 10.3389/feduc.2023.1170348.

Nieminen J.H., Carless D. Feedback literacy: A critical review of an emerging concept // Higher Education. 2023. Vol. 85. No. 6. P. 1381–1400.

Nocera A., Condino S., Ferrari V. Proof of Concept of Using HoloLens 2 for AR Immersive Training in Complex Medical Scenarios / 2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE). IEEE, 2023. P. 1–5. doi: 10.1109/MetroXRAINE58569.2023.10405738.

Norrthon A., Schörling E. Generating Feedback for Multiple Choice Questions with the Help of AI. 2023.

Omopekunola M.O., Kardanova E.Y. Automatic generation of physics items with Large Language Models (LLMs) // REID (Research and Evaluation in Education). 2024. Vol. 10. No. 2. P. 168–185.

Ossa C., Willatt C. Uso de Inteligencia Artificial Generativa para retroalimentar escritura académica en procesos de Formación Inicial Docente // European Journal of Education and Psychology. 2023. Vol. 16. No. 2. P. 1–16.

Otaki B. Feedback in the era of generative AI. 2023. https://hdl.handle.net/2077/77610.

Page E.B. Grading essays by computer: Progress report / Proceedings of the invitational Conference on Testing Problems. 1967.

Panfilova A.S., Valueva E.A., Ilyin I.Y. The application of explainable artificial intelligence methods to models for automatic creativity assessment // Frontiers in Artificial Intelligence. 2024. Vol. 7. P. 1310518. doi: 10.3389/frai.2024.1310518.

Panhoon S., Wongwanich S. An analysis of teacher feedback for improving teaching quality in primary schools // Procedia-Social and Behavioral Sciences. 2014. Vol. 116. P. 4124–4130.

Pankiewicz M., Baker R.S. Large Language Models (GPT) for automating feedback on programming assignments // arXiv preprint arXiv:2307.00150. 2023.

Pelima L. R., Sukmana Y., Rosmansyah Y. Predicting university student graduation using academic performance and machine learning: a systematic literature review / IEEE Access. 2024.

Pennington J., Socher R., Manning C.D. Glove: Global vectors for word representation / Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014. P. 1532–1543.

Pereira D., Flores M.A., Niklasson L. Assessment revisited: a review of research in Assessment and Evaluation in Higher Education // Assessment & Evaluation in Higher Education. 2016. Vol. 41. No. 7. P. 1008–1032.

Pereira D.S.M. et al. ChatGPT as an item calibration tool: Psychometric insights in a high-stakes examination // Medical Teacher. 2025. Vol. 47. No. 4. P. 677–683. doi: 10.1080/0142159X.2024.2376205.

Polat F., Tiddi I., Groth P. Testing Prompt Engineering Methods for Knowledge Extraction from Text / Semantic Web. Under Review. 2024.

Qin H., Guo L. Using machine learning to improve Q-matrix validation // Behavior Research Methods. 2024. Vol. 56. No. 3. P. 1916–1935.

Qu S. et al. Predicting achievement of students in smart campus // IEEE access. 2018. Vol. 6. P. 60264–60273.

Radford A. et al. Improving Language Understanding by Generative Pre-Training. 2018. https://cdn.openai.com/research-covers/language-

unsupervised/language_understanding_paper.pdf (date of access: 06.03.2025).

Radford A. et al. Language models are unsupervised multitask learners // OpenAI blog. 2019. Vol. 1. No. 8. P. 9.

Raksasat R. et al. Attentive pairwise interaction network for AI-assisted clock drawing test assessment of early visuospatial deficits // Scientific Reports. 2023. Vol. 13. No. 1. P. 18113. 10.2139/ssrn.4327538.

Ramesh D., Sanampudi S.K. An automated essay scoring systems: a systematic literature review // Artificial Intelligence Review. 2022. Vol. 55. No. 3. P. 2495–2527. https://doi.org/10.1007/s10462-021-10068-2.

Rangapur A., Rangapur A. The Battle of LLMs: A Comparative Study in Conversational QA Tasks // arXiv preprint arXiv:2405.18344. 2024.

Rao K. Universal design for learning and multimedia technology: Supporting culturally and linguistically diverse students // Journal of Educational Multimedia and Hypermedia. 2015. Vol. 24. No. 2. P. 121–137.

Raz T. et al. Automated Scoring of Open-Ended Question Complexity: A Large Language Model Approach. 2024.

Rock Paper Reality. How AI is making immersive experiences more powerful. 2024. https://rockpaperreality.com/insights/ar-development/how-ai-is-making-immersive-experiences-more-powerful/.

Rudovic O. et al. Personalized machine learning for robot perception of affect and engagement in autism therapy // Science Robotics. 2018. Vol. 3. No. 19. https://doi.org/10.1126/scirobotics.aao6760.

Ruengchaijatuporn N. et al. An explainable self-attention deep neural network for detecting mild cognitive impairment using multi-input digital drawing tasks // Alzheimer's Research & Therapy. 2022. Vol. 14. No. 1. P. 111. https://doi.org/10.1186/s13195-022-01043-2.

Salas-Pilco S.Z., Xiao K., Oshima J. Artificial Intelligence and New Technologies in Inclusive Education for Minority Students: A Systematic Review // Sustainability. 2022. Vol. 14. No. 20. P. 13572. doi: https://doi.org/10.3390/su142013572.

Schiff D. Education for AI, not AI for education: The role of education and ethics in national AI policy strategies //I nternational Journal of Artificial Intelligence in Education. 2022. Vol. 32. No. 3. P. 527–563.

Schneider J. et al. Towards LLM-based Autograding for Short Textual Answers // arXiv e-prints. 2023. P. arXiv: 2309.11508.

Sharma K., Giannakos M. Multimodal data capabilities for learning: What can multimodal data tell us about learning? // British Journal of Educational Technology. 2020. Vol. 51. No. 5. P. 1450–1484. doi: https://doi.org/10.1111/bjet.12993.

Simkute A. et al. Ironies of Generative AI: Understanding and Mitigating Productivity Loss in Human-AI Interaction // International Journal of Human–Computer Interaction. 2024. P. 1–22.

Singla Y.K. et al. Speaker-conditioned hierarchical modeling for automated speech scoring / Proceedings of the 30th ACM international conference on information & knowledge management. 2021. P. 1681–1691.

Strivr. (n.d.). The role of generative AI in immersive learning and VR training experiences. Strivr Blog. https://www.strivr.com/blog/role-genai-immersive-learning-vr-training-experiences.

Song C. et al. Taxonprompt: Taxonomy-aware curriculum prompt learning for few-shot event class ification // Knowledge-Based Systems. 2023. Vol. 264. P. 110290.

Su J. et al. EssayJudge: A Multi-Granular Benchmark for Assessing Automated Essay Scoring Capabilities of Multimodal Large Language Models // arXiv preprint arXiv:2502.11916. 2025.

Tan B. et al. A Review of Automatic Item Generation Techniques Leveraging Large Language Models. 2024. https://doi.org/10.35542/osf.io/6d8tj.

Tyack L., Khorramdel L., von Davier M. Using Convolutional Neural Networks to Automatically Score Eight TIMSS 2019 Graphical Response Items // Computers and Education: Artificial Intelligence. 2024. P. 100249. doi: 10.1016/j.caeai.2024.100249.

Ulitzsch E., He Q., Pohl S. Using sequence mining techniques for understanding incorrect behavioral patterns on interactive tasks // Journal of Educational and Behavioral Statistics. 2022. Vol. 47. No. 1. P. 3–35. https://doi.org/10.3102/10769986211010467.

UNESCO. Recommendation on the Ethics of Artificial Intelligence / UNESCO. SHS/BIO/PI/2021/1. 2022. https://unesdoc.unesco.org/ark:/48223/pf0000381137 (date of access: 07.03.2025).

Vasa J., Thakkar A. Deep learning: Differential privacy preservation in the era of big data // Journal of computer information systems. 2023. Vol. 63. No. 3. P. 608–631.

Van der Linden W. J. (ed.). Handbook of item response theory: Volume 3: Applications. CRC press, 2017.

Van Dijk L.J. AI as the assistant of the teacher: an adaptive math application for primary schools : M.Sc. Thesis / University of Twente, Faculty of Electrical Engineering, Mathematics & Computer Science. Enschede, 2021.

Von Davier M. Automated item generation with recurrent neural networks // Psychometrika. 2018. Vol. 83. No. 4. P. 847–857.

Von Davier M., Tyack L., Khorramdel L. Scoring graphical responses in TIMSS 2019 using artificial neural networks // Educational and Psychological Measurement. 2023. Vol. 83. No. 3. P. 556–585. https://doi.org/10.1177/00131644221098021

Wainer H., Mislevy R.J. Item response theory, item calibration, and proficiency estimation / Computerized adaptive testing. Routledge, 2000. P. 61–100.

Wang L. et al. ChatGPT's capabilities in providing feedback on undergraduate students' argumentation: A case study // Thinking Skills and Creativity. 2024. Vol. 51. P. 101440.

Wisniewski B., Zierer K., Hattie J. The power of feedback revisited: A meta-analysis of educational feedback research // Frontiers in psychology. 2020. Vol. 10. P. 487662.

Yao L., Jiao H. Comparing performance of feature extraction methods and machine learning models in essay scoring // Chinese/English Journal of Educational Measurement and Evaluation| 教育测量与评估双语期刊. 2023. Vol. 4. No. 3. P. 1. DOI: https://doi.org/10.59863/DQIZ8440.

Youn Y.C. et al. Use of the Clock Drawing Test and the Rey–Osterrieth Complex Figure Test-copy with convolutional neural networks to predict cognitive impairment // Alzheimer's research & therapy. 2021. Vol. 13. P. 1–7. https://doi.org/10.1186/s13195-021-00821-8.

Yusupova E.M., Antipkina I.V., Bakay E.A. What Features Improve the Automated Scoring of Short Open Answers in a Reading Test // International Journal of Artificial Intelligence in Education. 2024.

Zhai X. et al. Applying machine learning in science assessment: a systematic review // Studies in Science Education. 2020. Vol. 56. No. 1. P. 111–151. doi:10.1080/03057267.2020.1735757.

Zhang D. et al. Right to be forgotten in the era of large language models: Implications, challenges, and solutions // AI and Ethics.2024. Vol. 5. No. 3. P. 2445-2454. DOI:10.1007/s43681-024-00573-9.

Zhang L. et al. An automatic short-answer grading model for semi-open-ended questions // Interactive learning environments. 2022. Vol. 30. No. 1. P. 177–190. https://doi.org/10.1080/10494820.2019.1648300.

Zhang M., Li J. A commentary of GPT-3 in MIT Technology Review 2021 // Fundamental Research. 2021. Vol. 1. No. 6. P. 831–833.

Zhang W. AI fairness in practice: Paradigm, challenges, and prospects // Ai Magazine. 2024. Vol. 45. No. 3. P. 386–395.

Zhong Q. et al. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert // arXiv preprint arXiv:2302.10198. 2023.

НОВЫЕ ПОДХОДЫ К ОЦЕНИВАНИЮ: ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ КАК ДРАЙВЕР ИЗМЕНЕНИЙ В ОБРАЗОВАНИИ

Карданова Елена Юрьевна,
кандидат физико-математических наук, научный руководитель Центра психометрики и измерений в образовании Института образования НИУ ВШЭ и магистерской программы «Обучение и оценивание как наука».
E-mail: ekardanova@hse.ru

Тарасов Сергей Владимирович,
заместитель директора Центра психометрики и измерений в образовании Института образования НИУ ВШЭ.
E-mail: svtarasov@hse.ru

Иванова Алина Евгеньевна,
кандидат наук об образовании, заместитель директора Центра психометрики и измерений в образовании Института образования НИУ ВШЭ.
E-mail: aeivanova@hse.ru

Юсупова Элен Магомедовна,
кандидат наук об образовании, научный сотрудник Центра психометрики и измерений в образовании Института образования НИУ ВШЭ.
E-mail: eabdurakhmanova@hse.ru

Грачева Дарья Александровна,
кандидат наук об образовании, научный сотрудник Центра психометрики и измерений в образовании Института образования НИУ ВШЭ, академический руководитель магистерской программы «Обучение и оценивание как наука».
E-mail: dgracheva@hse.ru

Тарасова Ксения Вадимовна,
кандидат педагогических наук, директор Центра психометрики и измерений в образовании Института образования НИУ ВШЭ.
E-mail: ktarasova@hse.ru

Денисов Илья Сергеевич,
младший научный сотрудник, аналитик Центра психометрики и измерений в образовании Института образования НИУ ВШЭ.
E-mail: idenisov@hse.ru

Талов Даниил Павлович,
младший научный сотрудник Центра психометрики и измерений в образовании Института образования НИУ ВШЭ.
E-mail: dtalov@hse.ru

Струкова Александра Сергеевна,
младший научный сотрудник Центра психометрики и измерений в образовании Института образования НИУ ВШЭ.
E-mail: alstrukova@hse.ru

Аннотация. Искусственный интеллект (ИИ) — удобный инструмент, открывающий новые возможности в сфере образовательного оценивания. Авторы рассматривают применение ИИ на всех ключевых этапах оценивания: автоматическая разработка заданий, проверка работ, предоставление персонализированной обратной связи и анализ результатов. Особое внимание уделено этическим вопросам, связанным с прозрачностью, предвзятостью, конфиденциальностью и ограничениями использования ИИ в оценивании.

В выпуске приводятся практические примеры и современные исследования, что делает его полезным ресурсом для специалистов в области образования, психометрики и цифровых технологий.

Ключевые слова: искусственный интеллект, оценивание в образовании, автоматическая генерация заданий, автоматическая оценка, обратная связь, этика ИИ, образовательные технологии.

NEW APPROACHES TO ASSESSMENT: ARTIFICIAL INTELLIGENCE AS A DRIVER OF CHANGE IN EDUCATION

Kardanova Elena,
Ph.D. in Physics and Mathematics, Scientific Supervisor of the Center for Psychometrics and Measurement in Education at the Institute of Education, HSE University, and Scientific Supervisor of the Master's Program «Science of Learning and Assessment».
E-mail: ekardanova@hse.ru

Tarasov Sergei,
Deputy Director of the Center for Psychometrics and Measurement in Education at the Institute of Education, HSE University.
E-mail: svtarasov@hse.ru

Ivanova Alina,
Ph.D. in Education, Deputy Director of the Center for Psychometrics and Measurement in Education at the Institute of Education, HSE University.
E-mail: aeivanova@hse.ru

Yusupova Elen,
Ph.D. in Education, Research Fellow at the Center for Psychometrics and Measurement in Education at Institute of Education, HSE University.
E-mail: eabdurakhmanova@hse.ru

Gracheva Daria,
Ph.D. in Education, Research Fellow at the Center for Psychometrics and Measurement in Education at the Institute of Education, HSE University, and Academic Supervisor of the Master's Program Master's Program «Science of Learning and Assessment».
E-mail: dgracheva@hse.ru

Tarasova Ksenia,
Ph.D. in Education, Director of the Center for Psychometrics and Measurement in Education at the Institute of Education, HSE University.
E-mail: ktarasova@hse.ru

Denisov Ilya,
Junior Research Fellow and Analyst at the Center for Psychometrics and Measurement in Education at the Institute of Education, HSE University.
E-mail: idenisov@hse.ru

Talov Daniil,
Junior Research Fellow at the Center for Psychometrics and Measurement in Education at the Institute of Education, HSE University.
E-mail: dtalov@hse.ru

Strukova Alexandra,
Junior Research Fellow at the Center for Psychometrics and Measurement in Education at the Institute of Education, HSE University.
E-mail: alstrukova@hse.ru

Abstract. Artificial intelligence is a convenient tool that offers new opportunities in the field of educational assessment. The authors examine the application of AI across all key stages of assessment: automatic item generation, scoring of responses, personalized feedback provision, and results analysis. Particular attention is given to ethical issues related to transparency, bias, privacy, and the limitations of AI use in assessment.

The issue presents practical examples and recent research, making it a valuable resource for professionals in education, psychometrics, and digital technologies.

Keywords: artificial intelligence, educational assessment, automatic item generation, automatic scoring, feedback, AI ethics, educational technologies.

Один из сильнейших университетов страны приглашает на бюджетные места

Институт образования НИУ ВШЭ предоставляет уникальную возможность для профессионального развития и карьерного роста. Образовательные программы построены с учетом научных разработок и изменений в законодательстве. Среди преподавателей — ведущие российские и зарубежные ученые, признанные эксперты-практики российского образования.

МАГИСТЕРСКИЕ ПРОГРАММЫ

Для будущих ученых

■ **Трек «Магистратура — аспирантура»**
Период обучения: 5 лет
Форма обучения: очно-заочная

Для старта карьеры в образовании

Период обучения: 2 года.
Форма обучения: очная

■ **«Доказательное развитие образования»**
Академический руководитель — В.А. Мальцева

■ **«Обучение и оценивание как наука»**
Академический руководитель — Д.А. Грачева
Научный руководитель — Е.Ю. Карданова

■ **«Педагогическое образование»**
Академический руководитель — Ю.Н. Корешникова

Для руководителей вузов и школ

Период обучения: 2,5 года
Форма обучения: очно-заочная

■ **«Управление в высшем образовании»**
Академический руководитель — Н.К. Габдрахманов

■ **«Управление образованием»**
Академические руководители — Н.В. Исаева, А.А. Кобцева

■ **«Цифровая трансформация образования»**
Академический руководитель — А.А. Кобцева

Обучение осуществляется как бесплатно на бюджетной основе, так и с оплатой на договорной основе. Работникам бюджетных учреждений предоставляется 50%-я скидка на обучение при поступлении на коммерцию.

Департамент образовательных программ Института образования НИУ ВШЭ:
https://ioe.hse.ru/masters
Тел.: +7 495 772-95-90 (доб. 23094, 23452)

АСПИРАНТСКАЯ ШКОЛА ПО ОБРАЗОВАНИЮ

Институт образования НИУ ВШЭ приглашает к поступлению в уникальную для России Аспирантскую школу по образованию. Аспирантская школа открывает возможность проводить исследования на стыке наук, применяя междисциплинарный подход. После защиты соискатели получают степень кандидата наук НИУ ВШЭ об образовании / PhD HSE in Education

Преимущества программы:

- ✓ Практика исследований и возможность трудоустройства с первых дней
- ✓ Система финансовой поддержки аспирантов
- ✓ Онлайн-стажировки в ведущих мировых университетах по теме исследования
- ✓ Доступ ко всем образовательным и академическим ресурсам ВШЭ
- ✓ Трек по «Измерениям и оцениванию в образовании»
- ✓ Регулярные презентации новых исследований

Школа предлагает две формы обучения и подготовки диссертации:

Классическая аспирантура — для тех, кто хочет полностью сфокусироваться на развитии научной карьеры. Это очная аспирантура, дающая все плюсы обучения в аспирантской школе: статус аспиранта, комплексную поддержку на протяжении всего периода обучения и подготовки диссертации, возможность трудоустройства в центры и проекты Института образования и т.д.

Профессиональная аспирантура — для тех, кто уже нашел себя в бизнес- и управленческих структурах сферы образования. Эта очная программа дает возможность совмещать обучение с занятостью вне стен Института.

Как поступить?
Подробная информация на сайте: https://aspirantura.hse.ru/ed/howtoapply

Обучение очное и бесплатное — три года.

Аспирантская школа по образованию:
https://aspirantura.hse.ru/ed
Тел.: +7 495 772-95-90 (доб. 22714)

Научное издание

Серия
Современная аналитика образования

№ 5 (88)

НОВЫЕ ПОДХОДЫ К ОЦЕНИВАНИЮ: ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ КАК ДРАЙВЕР ИЗМЕНЕНИЙ В ОБРАЗОВАНИИ

Редактор: *И. Гумерова*
Компьютерная верстка: *Н. Пузанова*

Подписано в печать 08.08.2025. Формат 60×84 1/16
Усл.-печ. л. 5,12. Уч.-изд. л. 4,9. Тираж 100 экз.

Национальный исследовательский университет
«Высшая школа экономики»
101000, Москва, ул. Мясницкая, д. 20
Тел.: +7 495 624-40-27

Институт образования
101000, Москва, Потаповский пер., д. 16, стр. 10
Тел.: +7 495 623-52-49
ioe@hse.ru

ISSN 2500-0608