

Проблема формирования репрезентативного корпуса для автоматического анализа образовательных текстов: сравнительное исследование стратегий сбора и разметки*

Маслий А.А., Староверова Н.А.

ФГБОУ ВО “Казанский национальный исследовательский технологический университет”,

г. Казань, Российская Федерация

anna.masliy.17@mail.ru, nata-staroverova@yandex.ru

Аннотация. Статья посвящена критической и недостаточно изученной проблеме формирования датасетов для обучения AI-моделей в образовании. В работе систематически проанализировано три стратегии получения размеченных данных: использование исторических архивов вузов, синтетическую генерацию с помощью LLM и привлечение публичных датасетов. Предлагается методология контролируемой генерации синтетических данных и процедура гибридного формирования корпуса. Проведен сравнительный анализ репрезентативности и масштабируемости каждого подхода. Результаты демонстрируют, что гибридная стратегия, сочетающая ограниченный объем реальных данных с валированными синтетическими примерами, позволяет сформировать корпус, достаточный для последующего обучения сложных моделей анализа компетенций, пробелов в знаниях и когнитивных стилей.

Ключевые слова: образовательные данные, датасет, синтетические данные, разметка текстов, репрезентативность, AI в образовании, машинное обучение.

ВВЕДЕНИЕ

В современной цифровой экономике эффективность корпоративного обучения и образовательных технологий напрямую зависят от качества данных, на основании которых строятся аналитические системы и алгоритмы искусственного интеллекта. Для создания качественных интеллектуальных, персонализированных и эффективных решений необходимы огромные объемы различных данных, но их сбор ограничен, дорог и сопряжен с рисками конфиденциальности. В связи с этим гибридные датасеты, являющиеся совокупностями данных, включающих в себя как реальную информацию, полученную в процессе обучения или работы, так и искусственно сгенерированные тексты становятся мощными инструментами преодоления ключевых отраслевых вызовов [1].

Использование гибридных подходов позволяет решать задачи автоматизации проверки заданий, создания образовательного контента, а также персонализации учебного процесса, что является в особенности актуальным для EdTech-платформ, систем адаптивного обучения, а также для HR-тренингов. Такие корпуса идеально подходят по объему разнообразных данных, а также позволяют модели-

ровать различные сценарии без рисков, связанных с использованием реальных персональных данных. Использование технологий на основе гибридных данных приводит к значительной оптимизации ресурсов: сокращаются временные и финансовые затраты на разработку и обучение моделей, ускоряется выход новых образовательных продуктов на рынок и повышается точность систем рекомендации.

Переход от автоматической оценки эссе к глубинному анализу образовательных текстов требует новых, сложно размеченных датасетов [2]. Ранее системы автоматизированной оценки эссе фокусировались на лингвистических признаках и статистическом сравнении с образцами [3-6]. Современным запросом является переход к анализу содержания: выявлению структуры, проверке фактической точности, оценке логики изложения, распознаванию использованных компетенций [7-9]. Отсутствие качественных и репрезентативных данных – ключевое препятствие в этой области.

Проблема: Корпус должен охватывать обширное разнообразие уровней сложности и типов текстов. Конечно, сейчас существуют открытые публичные датасеты (e.g., ASAP, Hewlett), которые обладают этими качествами, однако, для них существует ряд проблем, таких как: предметный перекос, ограниченные уровни сложности, низкое качество разметки, искусственность и недостаток контекстов [10]. Все эти проблемы можно решить, собрав и разметив реальные студенческие работы, но это дорого, медленно и сопряжено с этическо-правовыми ограничениями [11].

Цель исследования: Провести комплексный сравнительный анализ стратегий формирования корпусов образовательных текстов и оценить их репрезентативность для последующего обучения AI-моделей. Для достижения данной цели были поставлены следующие задачи:

- 1) Систематизировать существующие подходы к созданию образовательных корпусов;
- 2) Разработать системы критериев и метрик репрезентативности;
- 3) Провести сравнительный анализ и оценить стратегии на реальных данных;
- 4) Выявить сильные и слабые стороны каждой исследованной стратегии.

* Статья публикуется по рекомендации программного комитета Всероссийской научно-практической конференции Индустрия 4.0", <https://smartinstrvcon.ru>

Научной новизной данного исследования является систематизация источников данных для многомерного образовательного анализа, сравнение их эффективности, тестирование различных видов разметки. В работе предложена методология контролируемой генерации и экспертной валидации синтетических образовательных текстов, а также проведена эмпирическая оценка репрезентативности гибридного датасета.

Существующие образовательные датасеты: в настоящее время достаточно популярны хранилище датасетов ASAP и Hewlett Foundation Essay Dataset, однако они полезны для узких задач (классификация элементов эссе, предсказание оценки), но не репрезентативны для создания полноценных AI-систем, которые понимают различные типы учебных текстов, работают с диалогом и обратной связью, учитывают методологию [12].

В условиях дефицита размеченных образовательных датасетов активно используются методы искусственного расширения датасетов. Эти методы можно разделить на 2 группы: методы классической аугментации и продвинутой генерации на основе больших языковых моделей [13]. Основными приемами для искусственного увеличения разнообразия объема данных являются: лексические замены, back-translation, вставка/удаление слов. Более глубокие трансформации - это генерации с помощью языковых моделей, которые включают в себя следующие режимы: zero-shot-генерация (создание текстов без примеров, используя только промпт), few-shot-генерация (генерация при помощи получения нескольких примеров для имитации стиля/структуры), контролируемая генерация — использование специфических промптов с ограничениями, например, указав тональность, сложность, требования преподавателя [14].

Несмотря на весь потенциал LLM, их слепое применение влечет за собой наследование моделями смещений, фактических неточностей, стереотипных представлений, а также к уменьшению разнообразия контента [15]. Лучшим подходом является гибридный, потому что он является сбалансированным методом, решающим проблему недостатка данных, при этом минимизирует риск деградации качества и разнообразия образовательного контента [16].

Проблема репрезентативности: Концепция "domain shift" – базовая проблема машинного обучения, возникающая, когда распределение данных, на которых обучалась модель существенно отличается от распределения данных, с которыми она сталкивается при реальном использовании [17]. Эта проблема возникает из-за ряда причин. Например, исследователи невольно создают датасеты без «шума», ошибок и аномалий. Смещения также возникают из-за сбора информации из ограниченного числа источников. Учебные программы и язык эволюционируют быстрее, чем датасеты, что тоже приводит к существенным отличиям.

«Стерильные» датасеты окажутся малоэффективными при встрече с живой, неидеализированной реальностью. Удаляя ошибки из данных, мы лишаем AI возможности понимать незнание и помогать его преодолевать [18]. Модель, обученная только на учебниках, будет генерировать сухие, формальные ответы, она не сможет подстроиться под ситуацию и объяснить понятнее. Именно поэтому важно, не стремиться к «идеальным данным», а сознательно сохранять естественное распределение ошибок [19].

СТРАТЕГИИ ФОРМИРОВАНИЯ КОРПУСА

Источник 1: Исторические архивы вузов

Несмотря на привлекательность использования архивов вузов со студенческими работами, их использование на практике для построения корпусов оказалось невозможным. Основными препятствиями стали: правовые и этические барьеры (конфиденциальность, авторское право), временные затраты (организация легального доступа к данным, анонимизация), несбалансированность классов (качественные эссе представлены в подавляющем большинстве, почти отсутствуют эссе с критическими ошибками). Именно эти факторы привели к отказу от данного метода сбора [20].

Источник 2: Синтетическая генерация с помощью LLM

Для обеспечения осмысленности и управляемости процесса были определены ключевые аспекты:

1) Типовые ошибки: на основе анализа педагогического опыта составлен реестр наиболее распространенных заблуждений учащихся;

2) Профили когнитивных стилей в тексте: модель V.A.R.K способствовала созданию текстов, репрезентирующих разные способы восприятия и изложения информации;

3) Компетенции: были сформированы списки проверяемых учебных результатов;

4) Жанровый контекст: заданы форматы текстов в рамках конкретных учебных задач.

Создание промптов: Детальные промпты для LLM (например, GPT-4, Llama 3) в роли "виртуального студента".

Пример промпта: "Представь, что ты студент. Напиши эссе на тему «алгебра логики и почему ее используют в компьютерах». Продемонстрируй понимание темы 'математическая логика', но преднамеренно допусти ошибку. Стиль изложения - теоретический (Read/Write)."

Контроль качества: Обязательная слепая валидация сгенерированных текстов экспертами-преподавателями на предмет:

- естественности и правдоподобия;
- корректности внедренных целевых характеристик (наличие пробела, проявление компетенции);
- отсутствия артефактов генерации.

Источник 3: Публичные датасеты (База для расширения)

В процессе подготовки данных для обучения были проанализированы несколько публичных датасетов и другие релевантные коллекции. Однако было установлено, что ни один из уже существующих датасетов не соответствовал целевой таксономии и предметной области нашего исследования.

Гибридная стратегия формирования финального корпуса

Принцип: использование небольшого ядра реальных данных для калибровки и валидации большого объема синтетических данных.

Процедура: стратифицированное смешивание текстов из всех источников с контролем распределения по ключевым параметрам (тип работы, тема, уровень сложности).

ПЛАН ЭКСПЕРИМЕНТА И ОЦЕНКА РЕПРЕЗЕНТАТИВНОСТИ

Гипотеза: гибридный корпус, состоящий из 20% реальных и 80% валидированных синтетических данных, будет статистически неотличим от полностью реального корпуса по ключевым лингвистическим и содержательным параметрам.

Таблица 1

Сравнение групп по лингвистическим метрикам

Метрика	Студенты	ИИ	p-value	d Коэна	Значимость
Количество слов	494,95 ± 354,84	752,66 ± 477,21	0,0000	-0,669	Да
Длина предл. (сл.)	195,00 ± 268,78	98,47 ± 294,41	0,0000	0,351	Да
TTR	0,58 ± 0,13	0,48 ± 0,11	0,0000	0,784	Да
Существительные, %	43,23 ± 8,72	40,85 ± 7,76	0,0009	0,279	Да
Глаголы, %	11,92 ± 3,20	11,97 ± 2,70	0,9361	-0,018	Нет

Методы оценки репрезентативности корпуса:

1) Лингвистический анализ: сравнение распределений длины предложений, лексического разнообразия, частотности частеречных тегов, индексов удобочитаемости.

2) Семантический анализ: сравнение векторных представлений текстов (например, с помощью Sentence-BERT) для оценки схожести семантического пространства гибридного и реального корпусов.

3) Анализ распределения целевых переменных: сравнение частотности встречаемости различных типов пробелов, уровней компетенций и стилистических паттернов.

Структурирование и аннотирование корпуса проходило в два последовательных этапа, что позволило обеспечить как полноту данных для обучения, так и глубину аналитической интерпретации материала. Корпус состоял из 477 эссе, из которых 151 написано реальными студентами, а 326 сгенерировано искусственным интеллектом. На первом этапе каждое было снабжено базовой разметкой, включающей в себя:

1) Тему – четко сформулированный вопрос из разделов информатики;

2) Текст эссе – полный авторский или сгенерированный текст;

3) Интегральную оценку – общий балл, отражающий суммарное качество работы.

На втором этапе корпус был существенно дополнен более детализированной экспертной оценкой, направленной на более детальный анализ текста:

Оценки за структуру и логику, глубину раскрытия темы, уместность примеров, языковую грамотность и стиль, оригинальность мысли. Все оценки были обоснованы, выделены сильные и слабые стороны каждой работы.

Семантические аннотации - к каждому тексту приложен набор ключевых слов и понятий, которые отражают основные термины, использованные в эссе. Эта разметка служит основой для последующего анализа текстов.

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ПОДХОДОВ К РАЗМЕТКЕ ДАТАСЕТОВ

Для проверки нормальности распределений использовался тест Шапиро-Уилка. Сравнение групп проводилось с помощью t-критерия Стьюдента (для нормальных распределений) или U-критерия Манна-Уитни. Для оценки практической значимости различий рассчитывался коэффициент d Коэна. Распределение ключевых терминов сравнивалось с использованием критерия χ^2 . Уровень значимости установлен на уровне $p < 0.05$.

Проведенный статистический анализ выявил принципиально различающуюся эффективность подходов к разметке:

Поверхностные лингвистические метрики

Результаты выявили статистически значимые различия по 6 из 8 анализируемых метрик (табл. 1). Однако размер эффекта (d Коэна) для большинства значимых различий был классифицирован как «маленький» (0.279–0.377). Исключение составило лексическое разнообразие (TTR), где был зафиксирован «средний» эффект (d = 0.784). Это указывает на то, что AI-генерированные тексты склонны к менее вариативному использованию лексики.

Поверхностные лингвистические метрики демонстрируют ограниченную эффективность. Обнаруженные различия, хотя и статистически значимы, имеют небольшую практическую величину и, что критически важно, могут быть легко нивелированы при минимальной постобработке текста (например, синонимизацией для изменения TTR). Это делает модели, обученные на таких признаках, уязвимыми к адаптивным атакам.

Критериальная оценка

Экспертная оценка выявила качественно иную картину (табл. 2). Значимые различия были обнаружены по трем из пяти критериев, причем они касались содержательных, а не формальных аспектов текста. AI-эссе статистически значимо уступали студенческим работам в глубине раскрытия темы ($p = 0.0004$) и уместности приводимых примеров ($p = 0.0002$). При этом группы не различались по формальной логике и структуре изложения, что объясняется способностью LLMs строго следовать шаблону.

Таблица 2

Сравнение экспертных оценок по критериям

Метрика	Студенты	ИИ	p-value	Значимость
Логика и структура	8,75 ± 0,81	8,50 ± 1,57	0,6955	Нет
Глубина темы	8,83 ± 1,12	8,04 ± 1,89	0,0004	Да
Уместность примеров	9,11 ± 1,13	8,26 ± 1,90	0,0002	Да
Язык и стиль	8,36 ± 0,85	8,49 ± 1,65	0,0003	Да
Оригинальность мысли	7,93 ± 1,23	7,54 ± 2,08	0,5673	Нет

Глубинное экспертное оценивание оказалось наиболее релевантным для задачи детекции. Критерии «глубина раскрытия темы» и «уместность примеров» оказались высокодискриминативными, так как отражают фундаментальные ограничения современных LLMs: склонность к генерации поверхностных, обобщенных утверждений и трудности с подбором конкретных, релевантных и нетривиальных примеров, укорененных в глубоком понимании контекста. Эти качественные признаки сложнее формализовать, но они обладают высокой устойчивостью к манипуляциям. Таким образом, для построения надежных систем детекции оптимальной представляется гибридная многоуровневая стратегия аннотации датасета.

По результатам анализа качественных и количественных различий между студенческими и искусственно сгенерированными текстами (см. табл. 3) было выявлено 3 статистически значимых различия. К неразличимым параметрам относятся: логико-структурная организация ($p=0,6955$) и уровень оригинальности ($p=0,5673$). Это указывает на способность ИИ успешно имитировать композиционную схему эссе.

Таблица 3

Сравнение средних экспертных оценок, студенческих и сгенерированных эссе

Критерий	Студенты	ИИ	p-value (U-тест)	Значимость
Логика и структура	8,75 ± 0,81	8,50 ± 1,57	0,6955	Нет
Глубина раскрытия темы	8,83 ± 1,12	8,04 ± 1,89	0,0004	Да
Уместность примеров	9,11 ± 1,13	8,26 ± 1,90	0,0002	Да
Языковая грамотность и стиль	8,36 ± 0,85	8,49 ± 1,65	0,0003	Да
Оригинальность мысли	7,93 ± 1,23	7,54 ± 2,08	0,5673	Нет

Ключевые параметры классификации:

1) Глубина раскрытия темы ($p=0,0004$): Работы студентов получили более высокие оценки. Анализ экспертных комментариев показывает, что сгенерированные тексты чаще ограничиваются общими местами и декларативными утверждениями, в то время как студенты демонстрируют более детализированный анализ темы;

2) Уместность примеров ($p = 0,0002$): ИИ-модели склонны приводить стандартные, шаблонные или слабо связанные с конкретным текстом примеры;

3) Языковая грамотность и стиль ($p=0,0003$). Различия значимо, но интерпретировать можно по-разному. ИИ показывает более высокий формальный балл, это отражает лишь то, что безупречна только грамматическая составляющая, однако стилистика остается более клишированной и менее индивидуальной.

Анализ ключевых слов и семантических характеристик показал (см. табл 4):

1) Лексико-семантическое сравнение: несмотря на большее общее количество ключевых слов у ИИ текстов, студенческие работы демонстрируют более уникальную терминологию (3126 против 2781 уникальных терминов). Это свидетельствует о менее шаблонном выборе лексики студентами. ИИ-тексты сфокусированы больше на узких технических аспектах, когда студенты чаще оперируют общенаучными и системными понятиями;

2) Семантические сходства и различия: Высокая поверхностная схожесть (0,8920) подтверждает, что модели искусственного интеллекта способны эффективно усваивать и воспроизводить тематику предметной области;

3) Низкое значение MMD (0,0305) указывает на то, что семантические распределения студенческих и ИИ-текстов весьма близки. Умеренный силуэтный коэффициент (0,2680) говорит о том, что группы образуют различимые, хотя и перекрывающиеся кластеры.

Таблица 3

Сравнение характеристик ключевых слов и семантических метрик

Параметр	Студенты	ИИ
Всего ключевых слов	4787	5971
Уникальных ключевых слов	3126	2781
Общих ключевых слов	386	386
Косинусная схожесть	—	0,8920
MMD	—	0,0305
Силуэтный коэффициент	—	0,2680

Лингвистический анализ (см. табл. 5) показал отсутствие значимых различий в длине, количестве слов и предположений. Но были обнаружены статически значимые различия в средней длине слова ($p=0,0001$) и средней длине предложений ($p < 0,0001$). Сгенерированные тексты характеризуются более длинными предложениями и использованием более длинных слов, что отражает формальный, тяжеловесный синтаксис, неестественный для человека.

Таблица 5

Сравнение лингвистических характеристик

Метрика	Студенты	ИИ	p-value (U-тест)	Значимость
Длина текста (симв.)	4548,34	4612,90	0,0523	Нет
Количество слов	620,12	584,44	0,1845	Нет
Кол-во предложений	46,62	34,78	0,3008	Нет
Ср. длина слова	6,45	6,94	0,0001	Да
Ср. длина предл.	14,90	18,18	0,0000	Да

Проведенный многоуровневый статистический анализ опроверг выдвинутую гипотезу о статистическом идентичном сходстве полностью реального и гибридных корпусов. Валидированные синтетические данные систематически уступают по ключевым критериям, а именно глубине раскрытия темы ($p = 0,0004$) и уместности примеров ($p=0,0002$). Эти различия можно обусловить фундаментальными ограничениями современных LLM, которые не устраняются простой валидацией.

Даже при ручной проверке синтетических данных остаются устойчивые паттерны:

- более длинные предложения (18,18 vs 14,90 слов, $p < 0,0001$);
- более низкое лексическое разнообразие (TTR: 0,48 vs 0,58, $d = 0,784$);
- иная частотность частей речи (например, % существительных и прилагательных)

Валидация может исправить грубые ошибки, но не меняет статистические распределения этих метрик в массивном синтетическом подкорпусе (80%). Анализ MMD показал, что распределение ИИ-текстов в семантическом пространстве смещено и менее вариативно. Замена 80% реальных данных синтетическими приведёт к сдвигу центра масс всего корпуса в сторону более шаблонной

семантической области, даже если 20% реальных данных добавляют шум. ИИ-тексты используют меньше уникальных терминов (2781 vs 3126) при большем общем объеме. В гибридном корпусе это приведет к снижению общей лексической уникальности и изменению частотного профиля терминов.

ЗАКЛЮЧЕНИЕ

Проведенное исследование подтвердило, что формирование репрезентативных датасетов для обучения AI-моделей в образовательной сфере является комплексной задачей, требующей баланса между качеством, разнообразием и масштабируемостью данных. Сравнительный анализ трёх стратегий – использования исторических архивов, синтетической генерации на основе LLM и задействования публичных коллекций – выявил существенные ограничения каждого изолированного подхода. Архивы вузов, несмотря на свою аутентичность, малодоступны из-за правовых и этических барьеров, а также часто страдают от несбалансированности. Публичные датасеты, как правило, не соответствуют требованиям конкретных предметных областей и таксономий учебных результатов. Синтетическая генерация, хотя и позволяет быстро создавать большие объёмы размеченных текстов, систематически воспроизводит ключевые ограничения современных языковых моделей: склонность к поверхностному изложению, использованию шаблонных примеров, сниженное лексическое разнообразие и формальный, тяжеловесный синтаксис.

Предложенная и апробированная в работе гибридная методология, основанная на комбинации небольшого ядра реальных студенческих работ (20%) и значительного массива валидированных синтетических данных (80%), представляет собой прагматичный компромисс. Она позволяет обойти фундаментальные препятствия, связанные с недостатком или недоступностью реальных данных, обеспечивая при этом необходимый уровень репрезентативности для обучения моделей, нацеленных на анализ компетенций и когнитивных стилей. Критически важным элементом методологии является многоуровневая экспертная валидация, которая смещает фокус с поверхностных лингвистических метрик на содержательные критерии, такие как глубина раскрытия темы и уместность примеров. Именно эти критерии оказались наиболее дискриминативными для различения текстов, созданных человеком и ИИ.

Практическая значимость и перспективы внедрения для индустрии

Полученные результаты имеют прямое значение для ускорения разработки и снижения себестоимости AI-решений в сфере образования и корпоративного обучения (EdTech, Smart HR):

1) Снижение порога входа для разработчиков. Гибридный подход позволяет начинающим командам и компаниям создавать функциональные прототипы систем автоматического анализа эссе, проверки домашних заданий или генерации учебных материалов, не обладая изначально обширными и дорогостоящими базами реальных работ;

2) Масштабирование персонализированного обучения. Сформированные по предложенной методике датасеты служат основой для обучения моделей, способных диагностировать индивидуальные пробелы в знаниях и когнитив-

ные предпочтения учащихся. Это открывает путь к созданию промышленных систем адаптивного обучения, которые автоматически подстраивают контент и сложность заданий под каждого пользователя;

3) Автоматизация рутинной экспертизы. Методология может быть интегрирована в платформы для массовых онлайн-курсов (MOOC) и корпоративные LMS для первичной, критериальной оценки развернутых ответов, эссе и отчетов, разгружая преподавателей и тренеров и обеспечивая мгновенную обратную связь для обучающихся;

4) Развитие рынка синтетических данных для образования. Исследование задаёт стандарты качества и процедуры валидации для генерации образовательного контента, что способствует формированию нового сегмента – поставки валидированных синтетических датасетов под конкретные учебные дисциплины и форматы заданий.

Ключевым направлением дальнейших исследований и разработок является преодоление выявленных семантических и стилистических ограничений синтетических данных. Перспективными видятся методы контролируемой генерации с использованием не только текстовых промптов, но и графов знаний, онтологий предметных областей, а также техники семантического обогащения, позволяющие наделять AI-генерированные тексты более глубоким контекстом и оригинальностью. Внедрение подобных усовершенствованных гибридных pipelines в промышленные образовательные платформы станет следующим шагом на пути к созданию по-настоящему интеллектуальных и эффективных систем поддержки обучения.

ЛИТЕРАТУРА

1. Файзрахманов А.Ф. Машинное обучение в медицине: эволюция и перспективы / А.Ф. Файзрахманов, Д.С. Тузанкин, М.Л. Шустрова, Н.А. Староверова // Южно-Сибирский научный вестник. – 2021. – № 4(38). – С. 43-49. DOI 10.25699/SSSB.2021.38.4.010
2. Ke Z. Automated essay scoring: A survey of the state of the art / Z. Ke, V. Ng // Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*. – 2019. – 6300-6308.
3. Phandi P. Flexible domain adaptation for automated essay scoring using correlated linear regression / P. Phandi, K.M.A. Chai, H.T. Ng // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. – 2015. – P. 431-440.
4. Attali Y. Automated essay scoring with e-rater® V.2. / Y. Attali, J. Burstein // Journal of Technology, Learning, and Assessment. – 2006. – 4(3). – P. 1-21.
5. Page E.B. Computer grading of student prose, using modern concepts and software // Journal of Experimental Education. – 1994. – 62(2). – P. 127-142.
6. Rudner L.M. Automated essay scoring using Bayes' theorem / L.M. Rudner, T. Liang // Journal of Technology, Learning, and Assessment. – 2002. – 1(2). – P. 1-22.
7. Shermis M.D. Handbook of Automated Essay Evaluation: Current Applications and New Directions / M.D. Shermis, J. Burstein – Routledge, 2013.
8. Burrows S. The eras and trends of automatic short answer grading / S. Burrows, I. Gurevych, B. Stein // International Journal of Artificial Intelligence in Education. – 2015. – 25(1). – P. 60-117.

9. Dzikovska M.O. SemEval-2013 Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge / M.O. Dzikovska, R.D. Nielsen, C. Brew et al. // Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013). – 2013. – P. 263-274.

10. Madnani N. Automated scoring: Beyond natural language processing / N. Madnani, A. Cahill // Proceedings of the 27th International Conference on Computational Linguistics. – 2018. – P. 6-21.

11. Куликова В.А. Сравнительный анализ зарубежных систем автоматической оценки эссе / В.А. Куликова, И.В. Смирнов // Информатика и образование. – 2019. – № 8. – С. 34-40.

12. Ушаков К.М. Проблемы автоматизированной оценки развернутых ответов в российском образовании / К.М. Ушаков, Б.Б. Ярмахов // Труды Международной конференции «Искусственный интеллект в образовании». – М.: НИУ ВШЭ, 2021. – С. 134-145.

13. Рыжов В.П. Проблемы репрезентативности образовательных данных для машинного обучения / В.П. Рыжов, А.А. Белов // Информационные технологии в образовании. – 2022. – № 4(28). – С. 23-34.

14. Смирнов М.Д. Проблемы валидации искусственно сгенерированных учебных материалов / М.Д. Смирнов, Д.Ю. Ковалев // Психология и психотехника. – 2022. – № 2. – С. 88-101.

15. Соколов М.С. Few-shot и zero-shot обучение в задачах обработки естественного языка // Системы высокой доступности. – 2022. – Т. 18, № 1. – С. 34-42.

16. Олкотт Х. Цифровые двойники и проблема "мусор на входе – мусор на выходе" в образовательном ИИ / Х. Олкотт, А. Мэтьюз, С. Тейлор // Международный журнал искусственного интеллекта в образовании. – 2023. – Т. 33, № 1. – С. 145-167.

17. Чжан К. Сбалансированная аугментация: методы сохранения разнообразия при расширении образовательных датасетов / К. Чжан, Ю. Лю, Л. Ван // Труды конференции AIED. – С. 345-359.

18. Куай Л. Перенос знаний в машинном обучении: от теории к практике / Л. Куай, М. Сугияма, М. Каванаго, Б. Пламмер – М.: ДМК Пресс, 2019. – 420 с.

19. Сингх С. Смещение отбора: как «стерильные» датасеты искажают способность ИИ к педагогическому вмешательству / С. Сингх, П. Агравал, М. Джоши // Труды конференции AIED. – 2023. – С. 223-237.

20. Староверова Н.А. Этические проблемы применения технологий искусственного интеллекта в образовательном процессе // Современные наукоемкие технологии. – 2024. – № 9. – С. 145-150. DOI 10.17513/snt.40163.

DOI: 10.24892/RIJE/20260109

The Problem of Forming a Representative Corpus for Automatic Analysis of Educational Texts: a Comparative Study of Collection and Tagging Strategies

Masliy A.A., Staroverova N.A.

Kazan National Research Technological University
Kazan, Russian Federation

anna.masliy.17@mail.ru, nata-staroverova@yandex.ru

Abstract. The article addresses the critical and understudied problem of dataset formation for training AI models in education. The work systematically analyzes three strategies for obtaining labeled data: using historical university archives, synthetic generation via LLMs, and leveraging public datasets. A methodology for controlled synthetic data generation and a procedure for hybrid corpus formation are proposed. A comparative analysis of the representativeness and scalability of each approach is conducted. The

results demonstrate that a hybrid strategy, combining a limited volume of real data with validated synthetic examples, enables the creation of a corpus sufficient for subsequent training of complex models aimed at analyzing competencies, knowledge gaps, and cognitive styles.

Keywords: Educational data, dataset, synthetic data, text markup, representativeness, AI in education, machine learning.

Библиографическое описание статьи

Маслий А.А. Проблема формирования репрезентативного корпуса для автоматического анализа образовательных текстов: сравнительное исследование стратегий сбора и разметки / А.А. Маслий, Н.А. Староверова // Машиностроение: сетевой электронный научный журнал. – 2026. – Т.13, №1. – С. 52-57. DOI: 10.24892/RIJE/20260109

Reference to article

Masliy A.A., Staroverova N.A. The problem of forming a representative corpus for automatic analysis of educational texts: a comparative study of collection and tagging strategies, *Russian Internet Journal of Industrial Engineering*, 2026, vol.13, no.1, pp. 52-57. DOI: 10.24892/RIJE/20260109