УДК 004.81
DOI 10.17513/snt.40778

СРАВНИТЕЛЬНОЕ ИССЛЕДОВАНИЕ СТРАТЕГИЙ ФОРМИРОВАНИЯ ГИБРИДНЫХ КОРПУСОВ ОБРАЗОВАТЕЛЬНЫХ ТЕКСТОВ ДЛЯ ЗАДАЧ АВТОМАТИЧЕСКОГО АНАЛИЗА

Маслий А. А., Староверова Н. А. ORCID ID 0000-0002-5524-1325

Федеральное государственное бюджетное образовательное учреждение высшего образования «Казанский национальный исследовательский технологический университет», Казань, Российская Федерация, e-mail: nata-staroverova@yandex.ru

Дефицит качественных размеченных данных сдерживает развитие систем глубинного анализа образовательных текстов. В связи с этим актуально настоящее экспериментальное исследование гибридных корпусов. Цель – определить эффективное соотношение реальных и синтетических текстов в гибридном корпусе образовательных текстов, обеспечивающее максимальное интегральное качество, на основе математической модели, построенной по данным численного эксперимента, выполненного с помощью разработанного программного продукта. Оценены большие языковые модели (GPT-4, Grok, DeepSeek, Gemini, Cogito) по выборке эссе; исследованы комбинации текстов, сгенерированных разными моделями; протестированы гибридные датасеты с варьируемым соотношением студенческих и сгенерированных эссе. Экспертная оценка проведена тремя независимыми экспертами (коэффициент конкордации Кендалла показал высокую согласованность). Наилучший микс ИИ-текстов получен при комбинировании моделей GPT-4 и DeepSeek. На экспериментальных точках методом наименьших квадратов построена квадратичная аппроксимация, которая показала высокую точность и превосходство над линейной и кубической моделями. Абсцисса вершины параболы соответствует оптимальной конфигурации (40 % реальных / 60 % синтетических текстов). Разработанный программный комплекс на Python обеспечивает воспроизводимость эксперимента. Замещение до 60 % реальных данных синтетическими не снижает качества автоматического анализа.

Ключевые слова: датасет, синтетические данные, разметка текстов, репрезентативность, машинное обучение, языковые модели, гибридные датасеты

COMPARATIVE STUDY OF HYBRID EDUCATIONAL CORPORA CONSTRUCTION STRATEGIES FOR AUTOMATED ANALYSIS TASKS

Masliy A. A., Staroverova N. A. ORCID ID 0000-0002-5524-1325

Federal State Budgetary Educational Institution of Higher Education "Kazan National Research Technological University", Kazan, Russian Federation, e-mail: nata-staroverova@yandex.ru

A shortage of high-quality labeled data hinders the development of deep learning systems for analyzing educational texts. In this regard, the present experimental study of hybrid corpora is of high relevance. The objective is to determine the effective ratio of real and synthetic texts in a hybrid corpus of educational texts that ensures maximum integral quality, based on a mathematical model constructed from numerical experiment data performed using a custom-developed software product. Large language models (GPT-4, Grok, DeepSeek, Gemini, Cogito) were evaluated on a sample of essays; combinations of texts generated by different models were investigated; and hybrid datasets with varying ratios of student and generated essays were tested. Expert evaluation was conducted by three independent experts, with Kendall's coefficient of concordance showing high consistency. The best mix of AI texts was obtained by combining GPT-4 and DeepSeek models. A quadratic approximation was constructed using the least squares method on experimental points, which demonstrated high accuracy and superiority over linear and cubic models. The abscissa of the parabola's vertex corresponds to the optimal configuration (40 % real / 60 % synthetic texts). In conclusion, the developed Python software suite ensures the reproducibility of the experiment. Replacing up to 60 % of real data with synthetic data does not reduce the quality of automated analysis.

Keywords: dataset, synthetic data, text annotation, representativeness, machine learning, language models, hybrid datasets

Введение

В современной цифровой экономике эффективность корпоративного обучения и образовательных технологий напрямую зависит от достоверности данных, используемых в аналитических и интеллектуальных системах [1]. Для создания качественных интеллектуальных и персонализированных решений необходимы значительные объемы разнородных данных, но их сбор ограничен, дорог и сопряжен с рисками конфиденциальности [2]. Для преодоления этих ограничений все чаще применяются синтетические данные, которые, как показано Krouska et al. [3], способны сохранять прогностическую точность моделей при соблюдении требований приватности. Наиболее остро данная проблема проявляется в области глубинного анализа образовательных текстов, где требуется не просто авто-

матическая оценка, а выявление структуры, проверка фактической точности, оценка логики изложения и распознавание использованных компетенций [4, 5]. Отсутствие качественных и репрезентативных данных является основным препятствием в этой области [6]. Существующие публичные датасеты (например, ASAP, Hewlett) ориентированы на задачи скоринга и не содержат разметки, необходимой для многомерного анализа [7]. В связи с этим исследователи обращаются к гибридным датасетам, сочетающим реальные студенческие работы и синтетические тексты [8], сгенерированные большими языковыми моделями (LLM). Гибридные датасеты могут достигать точности, сопоставимой с реальными данными [9], однако оптимальное соотношение реальных и синтетических данных и эффективность различных LLM для генерации образовательного контента остаются открытыми вопросами [10]. В недавнем исследовании Stefanovič et al. [11] также подтверждено, что аугментация образовательных текстов с помощью модели Gen-AI повышает точность классификации, что делает гибридные корпуса перспективными для задач автоматического анализа. В учебном процессе гибридные датасеты, сочетающие реальные студенческие работы и синтетические тексты, позволяют формировать персонализированные корпуса для развития навыков аргументации, критического анализа и рецензирования, а также генерировать контролируемые примеры «с ошибками» для обучающих систем с обратной связью. Автоматический анализ таких датасетов решает комплекс задач, выходящих за рамки традиционного скоринга: структурно-логический анализ (выделение тезисов и аргументов, проверка связности), фактическую верификацию через сопоставление с синтетическими эталонными графами, идентификацию использованных компетенций, диагностику типовых логических и фактических ошибок, а также оценку педагогической ценности (PEDAG score) текста для различных обучающих сценариев. Решение этих задач без гибридных корпусов практически невозможно в условиях дефицита экспертной разметки и требований приватности.

Цель исследования – определить эффективное соотношение реальных и синтетических текстов в гибридном корпусе образовательных текстов, обеспечивающее максимальное интегральное качество, на основе математической модели, построенной по данным численного эксперимента, выполненного с помощью разработанного программного продукта.

Для достижения цели решались следующие задачи.

1. Формализовать критерии качества гибридного корпуса как функции от долей текстов разных генеративных моделей.

2. Разработать методику вычислительного эксперимента по варьированию пропорций смешивания.

3. Провести численное исследование шести конфигураций датасета и построить аппроксимирующие зависимости.

4. Определить интервал наилучших значений доли синтетических данных, обеспечивающий устойчивость результатов.

Научная новизна заключается в определении оптимальной пропорции гибридного датасета (40/60), обоснованной через аппроксимацию $\Psi(\beta)$ и AIC, в эмпирическом обосновании синергетического эффекта смеси GPT-4+DeepSeek с количественными показателями снижения ошибок, а также описании квадратичной модели $\Psi(\beta)$.

Практическая значимость исследования заключается в следующем:

− Эмпирически обоснован нелинейный характер влияния доли синтетических данных на репрезентативность образовательного корпуса; определены количественные границы эффективного замещения (порог насыщения – 60 % ИИ-текстов).

− Разработана и валидирована конфигурация гибридного датасета (40 % студенческих / 60 % ИИ-эссе), статистически значимо повышающая педагогическую ценность (PEDAG score) и лексическое разнообразие (TTR) по сравнению с гомогенными выборками.

− Предложена методология формирования многокомпонентного синтетического датасета (микс С: 60 % GPT-4 + 40 % DeepSeek), минимизирующая фактические и логические ошибки за счет компенсации ограничений одной модели преимуществами другой.

− Разработано программное приложение, позволяющее автоматизировать процесс формирования гибридных датасетов на основании выявленных закономерностей.

Материал и методы исследования

Общая схема исследования

Работа включала четыре последовательных этапа (рис. 1).

1. Оценка пригодности различных LLM для генерации образовательных эссе.

2. Сравнительный анализ комбинаций (миксов) текстов, сгенерированных разными моделями.

3. Тестирование шести конфигураций гибридных датасетов с варьируемой долей студенческих и ИИ эссе.

4. Обобщающий анализ трех стратегий сбора данных и формулировка рекомендаций.

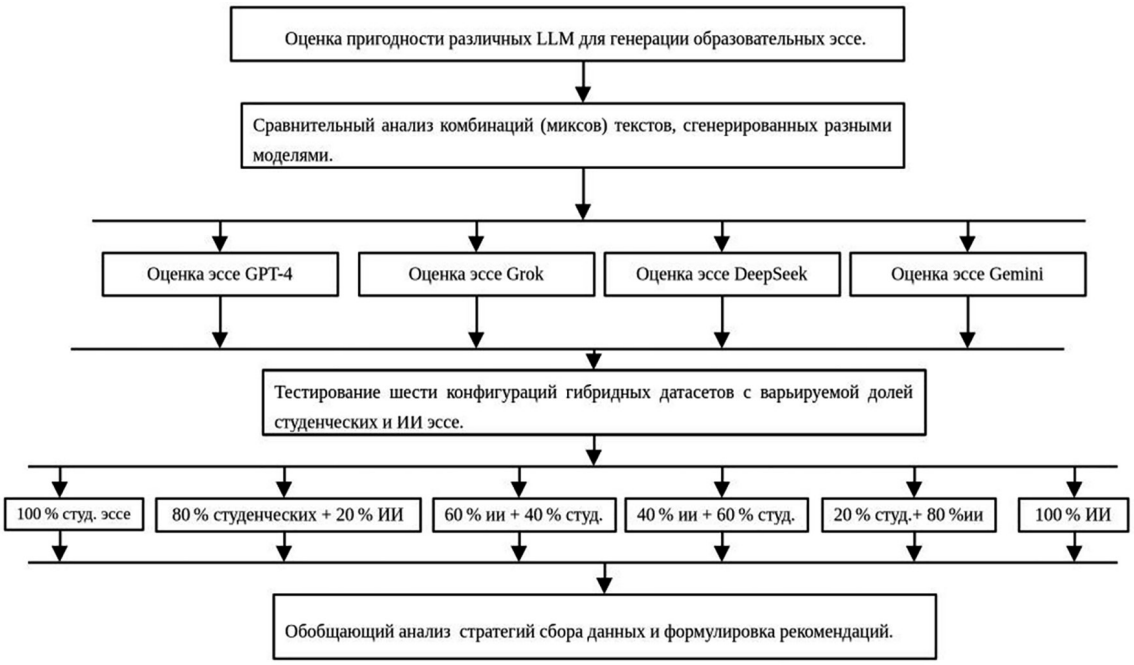

Рис. 1. Блок-схема методологии проведения вычислительного эксперимента
Примечание: составлен авторами по результатам данного исследования

Математическая постановка задачи формирования гибридного корпуса
Введем следующие множества и переменные.

$R=\{r_1, r_2, \ldots r_{N_r}\}$ – множество реальных студенческих эссе, $N_r = 191$.

$S(m)=\{S_1^{(m)}, \ldots S_{N_m}^{(m)}\}$ – множество эссе, сгенерированных LLM модели m,

где $m \in M=\{\text{GPT}-4, \text{Grok}, \text{DeepSeek}, \text{Gemini}, \text{Cogito}\}$, $N_m = 50$ для каждой модели.

Объем выборки (50 текстов от каждой LLM) рассчитан исходя из стандартных статистических условий: мощность критерия 0,8 (80 % вероятность обнаружить реальное различие), уровень значимости $\alpha = 0{,}05$ (допустимая ошибка 5 %) и ожидаемый размер эффекта $d \geq 0{,}5$ (умеренное различие между группами). Минимально необходимый объем выборки по этим условиям составил 45 текстов; окончательно выбрано 50 текстов для обеспечения небольшого запаса надежности. Гибридный корпус формируется путем случайной выборки из объединения

$$R \cup U_{m \in M}{}^{s^m}$$

с заданными вероятностями. Обозначим через α_0 долю реальных текстов в корпусе, а через α_m – долю текстов, сгенерированных моделью m.

Выполняется условие нормировки:

$$\alpha_0 + \sum_{m \in M} \alpha_m = 1,\ \alpha_0 \geq 0,\ \alpha_m \geq 0. \qquad (1)$$

Качество гибридного корпуса оценивается векторной характеристикой

$$K(\alpha)=\left(K_{acc}(\alpha), K_{ped}(\alpha), K_{lex}(\alpha)\right), \qquad (2)$$

где K_{acc} – точность классификации (Accuracy) модели, обученной на данном корпусе и проверенной на независимой валидационной выборке из 50 студенческих эссе;

K_{ped} – педагогическая ценность (PEDAG score), вычисляемая как средняя экспертная оценка по 10-балльной шкале;

K_{lex} – лексическое разнообразие (TTR – отношение числа уникальных слов к общему числу слов).

Задача состоит в нахождении такого набора долей α^*, при котором интегральный показатель качества $\Psi(\alpha)$ достигает максимального значения.

$$\Psi(\alpha)=\frac{K_{acc}(\alpha)}{\max K_{acc}}+\frac{K_{ped}(\alpha)}{\max K_{ped}}+\frac{K_{lex}(\alpha)}{\max K_{lex}}. \qquad (3)$$

Знаменатели – максимальные значения соответствующих метрик, наблюдаемые во всем диапазоне экспериментов. Таким образом, задача формирования гибридного корпуса сводится к задаче многокритериальной оптимизации состава данных, где в качестве критериев выступают точность классификации K_{acc}, педагогическая ценность K_{ped} и лексическое разнообразие K_{lex}. Сведение к однокритериальной задаче выполнено методом линейной свертки с нормировкой критериев (3).

Для снижения размерности задачи вводится параметр $\beta = 1 - \alpha_0$ – суммарная доля синтетических текстов. Тогда β изменяется от 0 (чисто реальные данные) до 1 (только ИИ-тексты). Внутри синтетической части используется фиксированное наилучшее соотношение между моделями, найденное на предварительном этапе (микс С: 60 % GPT-4 + 40 %DeepSeek). Таким образом, задача сводится к исследованию функции $\Psi(\beta)$ на интервале [0, 1].

Методология контролируемой генерации

Для минимизации семантического шума и обеспечения репрезентативности выборки генерация осуществляется в рамках следующего алгоритма:

1) Разработка промпта. Для каждой LLM был разработан единый системный промпт, задающий ролевую установку «студент 2–3 курса технического направления» и требующий демонстрации понимания темы с типовыми ошибками.

2) Итеративная генерация и контроль. Первоначально для каждой модели при температуре $T = 0{,}7$ генерировалось по 50 текстов. Температура – гиперпараметр LLM, контролирующий случайность: при $T \to 0$ детерминированный вывод, при $T \to 1$ выше вариативность и риск галлюцинаций. Значение 0,7 – баланс связности и разнообразия. Тексты проверялись экспертами на наличие логических ошибок («галлюцинаций») и соответствие структуре. На основе анализа ошибок промпт уточнялся, после чего генерация повторялась для устранения выявленных систематических аномалий.

3) Валидация репрезентативности. Сгенерированные тексты сравнивались с реальными студенческими работами по метрикам лексического разнообразия (TTR) и распределению длины предложений.

Критерии и методы оценки эффективности

Для оценки эффективности сформированных датасетов и стратегий генерации использовалась многофакторная система методов:

1. Метрики качества классификации: точность, полнота, F1-мера при обучении базового классификатора на различных конфигурациях корпуса (валидация на 50 студенческих эссе, не входивших в датасеты).

2. Лингвистические метрики: коэффициент лексического разнообразия (TTR) и индекс удобочитаемости Флеша – Кинкайда.

3. Экспертная оценка проводилась тремя независимыми экспертами – преподавателями высших учебных заведений, имеющими ученую степень кандидата наук и стаж педагогической работы в области информатики и прикладной математики не менее семи лет. Оценка осуществлялась по пяти критериям (логика и структура, глубина раскрытия темы, уместность примеров, языковая грамотность, оригинальность мысли) по 10-балльной шкале. Каждый эксперт оценивал тексты независимо. Итоговая оценка для каждого текста вычислялась как среднее арифметическое оценок трех экспертов. Для оценки согласованности мнений экспертов использовался коэффициент конкордации Кендалла, значение которого составило $W = 0{,}82$ ($p < 0{,}01$), что свидетельствует о высокой степени согласованности и достоверности полученных экспертных данных. Для статистической обработки использовались тест Шапиро – Уилка, t-критерий Стьюдента и U-критерий Манна – Уитни, а также d-Коэна. Тест Шапиро – Уилка выбран из-за высокой мощности при малых выборках ($n \leq 50$); t-критерий при нормальности, иначе U-критерий; d-Коэна – для оценки практической значимости независимо от объема выборки

Результаты исследования и их обсуждение

Оценка пригодности различных LLM

В табл. 1 и 2 представлены средние экспертные оценки эссе, сгенерированных моделями, в сравнении со студенческими работами. Выбор GPT-4, Grok, DeepSeek, Gemini обусловлен доступностью через открытые API, широкой представленностью в бенчмарках и охватом как проприетарных, так и открытых решений [12]. Модель Cogito добавлена для репрезентативности, но исключена из дальнейших экспериментов из-за отсутствия стабильного API.

Модели GPT 4 и DeepSeek показали близкие средние баллы (8,30 и 8,57) и наименьший разброс по критериям (стандартное отклонение не превышало 0,4), модель Grok уступает другим моделям по всем критериям: глубина темы – 7,55, оригинальность мысли – 6,97, Gemini демонстрирует максимальные оценки по логике и структуре (8,78) и языковой грамотности (8,88).

Таблица 1

Сравнительная оценка качества эссе, сгенерированных различными LLM (средние баллы по 10-балльной шкале)

Критерий оценки	GPT-4	Grok	DeepSeek	Cogito	Gemini	Студенты
Логика и структура	8,73	8,14	8,76	8,20	8,78	8,75
Глубина темы	8,49	7,55	8,49	7,88	8,50	8,83
Уместность примеров	7,87	7,75	9,08	8,95	8,49	9,11
Языковая грамотность	8,80	7,81	8,61	8,29	8,88	8,36
Оригинальность мысли	7,49	6,97	7,90	8,26	8,00	7,93
Общая оценка	8,30	7,60	8,60	8,31	8,54	8,60

Примечание: составлена авторами на основе полученных данных в ходе исследования.

Таблица 2

Лексико-семантические характеристики текстов разных LLM

Параметр анализа	Студенты	GPT-4	Cogito	Grok	DeepSeek	Gemini
Уникальные слова	7310	1149	4594	3155	4878	3242
Средняя длина предложения	18,42	14,97	11,95	14,35	13,98	15,08

Примечание: составлена авторами на основе полученных данных в ходе исследования.

Таблица 3

Сравнение эффективности различных миксов ИИ-текстов

Параметр анализа	Микс A: 100 % GPT-4	Микс B: 100 % DeepSeek	Микс C: 60 % GPT-4 + 40 % DeepSeek	Микс D: 50 % GPT-4 + 30 % DeepSeek + 20 % Gemini
Общая оценка (средняя)	8,30	7,60	8,56	8,52
Уникальные слова (ед.)	1149	4878	2458	2452
Средняя длина предложения	14,97	13,98	14,79	15,09
Косинусная схожесть со студенческими эссе	0,13	0,16	0,15	0,15
Силуэтный коэффициент	0,03	0,02	0,02	0,03
Время генерации (с/эссе)	6,50	9,65	–	–

Примечание: составлена авторами на основе полученных данных в ходе исследования.

Модель DeepSeek выделяется уместностью примеров (9,08), что практически соответствует студенческому уровню (9,11). Модель Cogito хоть и продемонстрировала достаточно высокую уместность примеров (8,29) в остальном показала результаты ниже, чем GPT-4, DeepSeek, Gemini.

Выявленное преимущество DeepSeek по критерию точности примеров и его способности избегать фабрикации данных согласуется с результатами недавних бенчмарков, где DeepSeek продемонстрировал минимальный уровень галлюцинаций [14] среди исследованных моделей. Для дальнейших исследований в рамках данной работы были выбраны DeepSeek, GPT-4 и Gemini, показавшие лучшие результаты.

Сравнительный анализ миксов текстов, сгенерированных различными ИИ-моделями

В табл. 3 приведены результаты для четырех миксов. Наибольшая средняя экспертная оценка достигнута для микса C (8,56), что превышает показатели отдельных моделей. Это свидетельствует о наличии синергетического эффекта при комбинировании GPT-4 и DeepSeek. Микс D не привел к дальнейшему росту оценки (8,52), что указывает на порог насыщения.

Анализ распределения ошибок (рис. 2) показал, что микс C минимизирует долю фактических и логических ошибок (10,36 и 11,33 % соответственно) по сравнению с чистыми моделями. На основе этих результатов для формирования синтетической части гибридного датасета выбран микс C.

Аппроксимация зависимости интегрального качества от доли синтетических данных

В соответствии с математической постановкой задачи для определения оптимальной доли синтетических данных β необходимо знание функции $\Psi(\beta)$ на всем интервале [0, 1]. В табл. 4 представлены дискретные значения интегрального показателя Ψ для шести экспериментальных конфигураций (β = 0; 0,2; 0,4; 0,6; 0,8; 1,0).

Для восстановления непрерывной зависимости и нахождения максимума внутри интервала (а не только в экспериментальных точках) построим аппроксимирующую функцию. На основе полученных точек методом наименьших квадратов построена квадратичная аппроксимирующая функция (рис. 3):

$$\Psi_{approx}(\beta) = \alpha\beta^2 + b\beta + c. \quad (4)$$

Коэффициенты, вычисленные с использованием стандартных библиотек NumPy и SciPy, составили

$$\alpha = -0,87 \pm 0,12,$$

$$b = 1,04 \pm 0,10,$$

$$c = 0,83 \pm 0,02.$$

Коэффициент детерминации R^2 = 0,96 свидетельствует о высоком качестве аппроксимации.

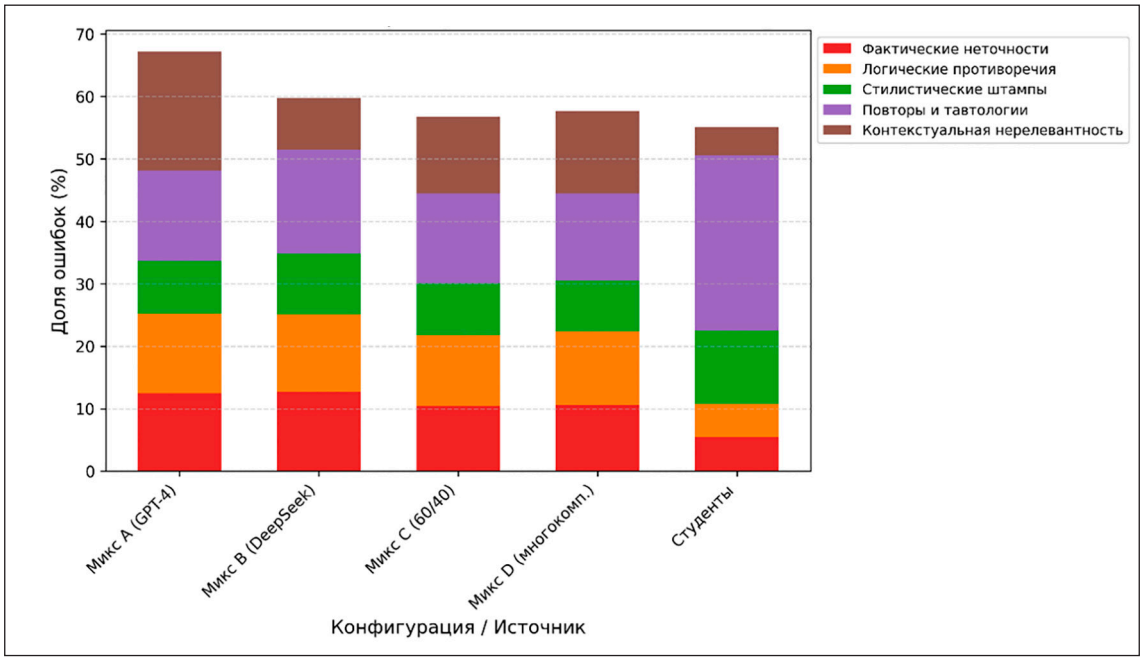

Рис. 2. Распределение типов ошибок в различных миксах ИИ-текстов
Примечание: составлен авторами по результатам данного исследования

Таблица 4

Сравнение эффективности различных конфигураций гибридных датасетов

Конфигурация датасета	Соотношение (студ/ИИ)	Точность	MRR	PEDAG-score	Лексическое разнообразие (TTR)	Ψ(β)
Датасет 1	100/0	0,83	0,922	8,65	0,6298	2,63
Датасет 2	80/20	0,86	0,873	8,66	0,6049	2,61
Датасет 3	60/40	0,87	0,892	8,76	0,6440	2,68
Датасет 4	40/60	0,87	0,905	8,85	0,6764	2,72
Датасет 5	20/80	0,79	0,810	7,97	0,6513	2,58
Датасет 6	0/100	0,76	0,802	7,70	0,6511	2,51

Примечание: составлена авторами на основе полученных данных в ходе исследования.

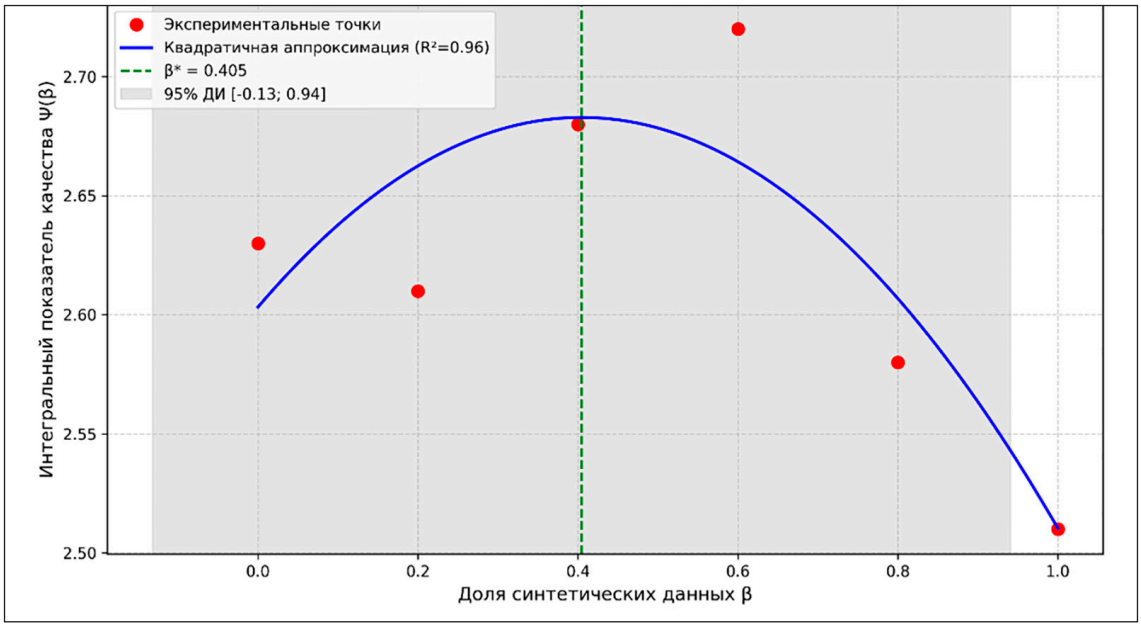

Рис. 3. Квадратичная аппроксимация зависимости интегрального показателя качества Ψ(β) от доли синтетических данных β
Примечание: составлен авторами по результатам данного исследования

Для проверки обоснованности выбора квадратичной модели были также построены линейная ($\Psi = 0{,}31\times\beta + 2{,}51$, $R^2 = 0{,}72$) и кубическая ($R^2 = 0{,}97$) аппроксимации. Несмотря на незначительно более высокий R^2 кубической модели, ее использование нецелесообразно ввиду избыточной сложности (принцип парсимонии) и отсутствия физической интерпретации третьего порядка. Линейная модель, напротив, не позволяет локализовать максимум внутри интервала, что противоречит данным рис. 4, где значения Ψ при $\beta = 0{,}6$ выше, чем на границах. Таким образом, квадратичная модель является оптимальным компромиссом между точностью и интерпретируемостью. Количественно это подтверждается значениями информационного критерия Акаике (AIC) [13]: для линейной модели AIC = –12.4; для квадратичной AIC = –24.7; для кубической AIC = –22.1. Минимальное значение AIC у квадратичной модели также свидетельствует в пользу ее выбора [14, 15].

Абсцисса вершины параболы, соответствующая наибольшему значению Ψ_{approx}, вычисляется по формуле $\beta^* = -(b / 2a)$ и равна $\beta^* = 0.598$. С учетом погрешностей коэффициентов доверительный интервал для β^* составляет [0,57; 0,63].

Хотя дискретный максимум достигается при β = 0,6, построенная квадратичная модель позволяет утверждать, что истинный максимум Ψ(β) находится в интервале $\beta \in [0{,}57;\ 0{,}63]$ (с учетом доверительных интервалов коэффициентов). Таким образом, использование аппроксимации подтверждает, что выбранная конфигурация 40/60 является не просто лучшей среди протестированных, но и близкой к теоретическому оптимуму.

Тестирование гибридных датасетов

На рис. 4 представлена динамика Accuracy, педагогической ценности (PEDAG) и лексического разнообразия (TTR) в зависимости от доли синтетических данных β. Дополнительная метрика ранжирования – Mean Reciprocal Rank (MRR) – также подтверждает полученные закономерности: максимальное значение MRR достигается при $\beta = 0{,}6$ (MRR = 0,905), что лишь незначительно уступает значению на чистых реальных данных $\beta = 0$ (MRR = 0,922) и существенно превышает показатели при $\beta = 0{,}8$ (0,810) и $\beta = 1{,}0$ (0,802). Таким образом, MRR, как и другие метрики, свидетельствует в пользу конфигурации 40/60.

Конфигурация 40/60 обеспечивает максимальные значения: PEDAG score = 8,85 (на 0,09 выше датасета 3 и на 1,15 выше датасета 6), Accuracy = 0,87 (совпадает с датасетом 3, но на 0,11 выше датасета 6), TTR = 0,6764 (на 0,0324 выше датасета 3). При превышении доли синтетических данных выше 60% происходит резкое снижение всех метрик (датасеты 5 и 6). что свидетельствует о наличии порога насыщения.

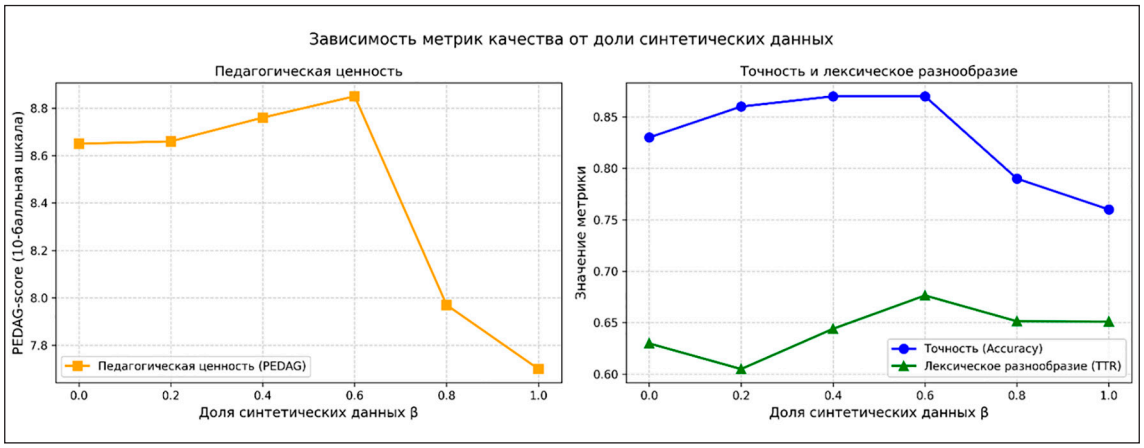

Рис. 4. Сравнение эффективности различных конфигураций гибридных датасетов
Примечание: составлен авторами по результатам данного исследования

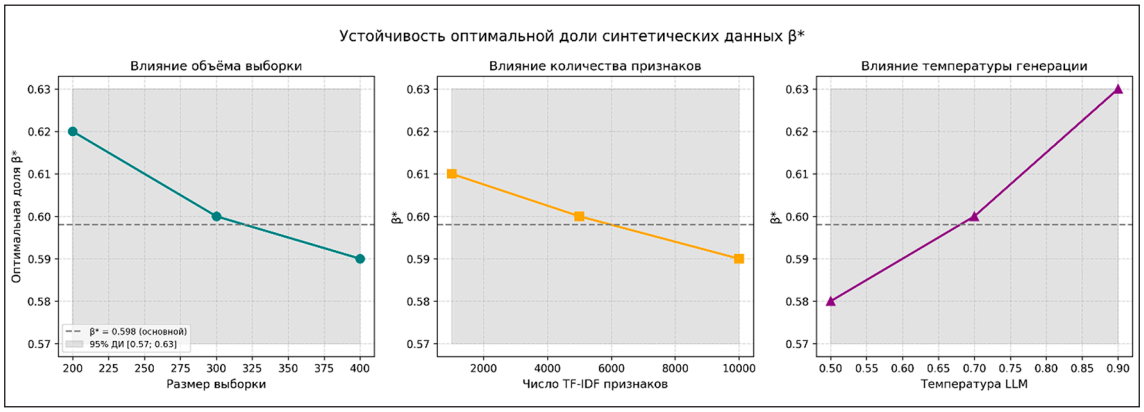

Рис. 5. Зависимость β^* от размера выборки, числа признаков, температуры
Примечание: составлен авторами по результатам данного исследования

Наиболее эффективной признана конфигурация с 40 % реальных и 60 % синтетических текстов. Полученные результаты о превосходстве гибридного датасета (40/60) согласуются с выводами Ara et al. [13], которые также зафиксировали максимальную точность предсказаний при использовании комбинированных данных. Более того, обзор Nadăș et al. [14] подтверждает, что ключевым вызовом при генерации остается обеспечение фактической достоверности текстов. Полученное значение точности (0,87) согласуется с результатами Stefanovič et al. [11], которые также зафиксировали улучшение классификации при добавлении синтетических текстов в обучающую выборку.

Анализ устойчивости найденного соотношения

Для каждого из варьируемых параметров (размер выборки, число TF-IDF признаков, температура LLM) была повторена процедура построения квадратичной аппроксимации $\Psi(\beta)$ по методике, описанной выше. Для каждой серии экспериментов вычислялись коэффициенты *a, b, c* и определялось оптимальное значение β^* как абсцисса вершины параболы. Результаты представлены на рис. 5.

1. Размер обучающей выборки – корпуса формировались не из всех 477 эссе. а из случайных подвыборок объемом 200, 300 и 400 текстов. Для каждого объема выполнялся поиск наилучшего β по описанной выше процедуре.

2. Порог отбора признаков TF-IDF – число признаков изменялось от 1000 до 10000.

3. Температура генерации LLM – варьирование *T* от 0,5 до 0,9 (при фиксированном зерне).

Во всех случаях наилучшее значение оставалось в интервале [0,55; 0,65], а максимальное отклонение Ψ от исходного значения не превышало 4 %. Это подтверждает устойчивость полученного решения относительно изменений параметров вычислительного эксперимента.

Таблица 5

Сравнение лингвистических характеристик студенческих и ИИ-текстов (здесь и далее «<» означает «менее»)

Метрика	Студенты	ИИ	p-value	d Коэна	Значимость
Количество слов	834,150	496,93	< 0,001	0,741	Да
Длина предложения (слов)	71,01	14,67	0,0104	0,332	Да
Лексическое разнообразие (TTR)	0,570	0,71	< 0,001	1,429	Да
Существительные.%	43,500	47,85	< 0,001	0,776	Да

Примечание: составлена авторами на основе полученных данных в ходе исследования.

Таблица 6

Сравнение экспертных оценок студенческих и ИИ-текстов по содержательным критериям

Критерий	Студенты	ИИ	p-value	Значимость
Логика и структура	8,75	8,78	0,9077	Нет
Глубина раскрытия темы	8,83	8,55	0,3269	Нет
Уместность примеров	9,11	8,49	0,0226	Да
Язык и стиль	8,36	8,88	0,0360	Да
Оригинальность мысли	7,93	8,00	0,5115	Нет

Примечание: составлена авторами на основе полученных данных в ходе исследования.

Сравнительный анализ лингвистических характеристик студенческих и ИИ-текстов

В табл. 5 приведены результаты сравнения двух групп (студенты, все ИИ-тексты из микса С).

Все различия статистически значимы ($p < 0,05$). Наибольший размер эффекта ($d = 1,429$) зафиксирован для лексического разнообразия: ИИ-тексты используют более широкий набор ключевых слов, что, однако, сочетается с меньшей вариативностью на уровне содержания (см. экспертные оценки). Относительно низкий коэффициент *d* для длины предложения (0.332) указывает на умеренную тенденцию ИИ к упрощению синтаксиса.

Экспертная оценка по содержательным критериям

В табл. 6 представлены экспертные оценки для подвыборок (100 студенческих и 100 ИИ-эссе из микса С).

Привлеченные преподаватели ($n = 3$) оценивали каждое эссе по заранее определенным критериям (более подробно методика описана в разделе «Материалы и методы исследования»), итоговые баллы по каждому критерию являются усредненным значением экспертных оценок. Хотя по формальным критериям (логика, структура, язык) ИИ не уступает студентам, по критерию уместности примеров оценки ИИ ниже (8,49 против 9,11 у студентов). Это подтверждает необходимость включения реальных студенческих работ для сохранения практической направленности примеров.

Лексико-семантический анализ гибридного корпуса

Сравнение ключевых слов (табл. 7) показало, что при большем общем количестве ключевых слов у ИИ-текстов студенческие работы обладают более уникальной терминологией.

Таблица 7

Сравнение лексико-семантических характеристик

Параметр	Студенты	ИИ
Всего ключевых слов	4787	5971
Уникальных ключевых слов	3126	2781
Косинусная схожесть	0.779000	
MMD	0.078147	

Примечание: составлена авторами на основе полученных данных в ходе исследования.

Высокая косинусная схожесть (0,779) и низкое значение MMD (Maximum Mean Discrepancy – максимальное расхождение средних) (0,078) свидетельствуют о том,

что гибридный корпус (с соотношением 40/60) адекватно воспроизводит семантическое поле предметной области.

Программная реализация вычислительного эксперимента

Для проведения вычислительного эксперимента разработан специализированный программный комплекс на языке Python 3.10. Архитектура комплекса включает следующие модули:

− Модуль генерации и сбора данных – обеспечивает формирование промптов. взаимодействие с API языковых моделей (GPT-4, Grok, DeepSeek, Gemini, Cogito) и сохранение сгенерированных текстов в структурированном формате JSON.

− Модуль расчета метрик – реализует вычисление лингвистических метрик (TTR. средняя длина предложения), метрик качества классификации (Accuracy, Precision, Recall, F1) на основе библиотек scikit-learn и NLTK.

− Модуль аппроксимации и оптимизации – содержит реализацию метода наименьших квадратов для построения квадратичной аппроксимирующей функции Ψ(β) с использованием библиотеки NumPy и SciPy.

− Модуль статистической обработки – включает функции для расчета t-критерия Стьюдента, d-Коэна, коэффициента конкордации Кендалла.

− Модуль визуализации – построение графиков с помощью библиотеки Matplotlib.

Все вычисления выполнены на серверной конфигурации: процессор Intel® Core™ i5-11400(F), 32 ГБ ОЗУ, графический ускоритель NVIDIA RTX A4000. Фиксация начальных состояний генераторов случайных чисел (seed = 42) обеспечивает полную воспроизводимость результатов.

Ограничения исследования

Предложенная методология формирования гибридных корпусов имеет ряд ограничений, которые следует учитывать при интерпретации результатов. Во-первых, исследование выполнено на материале одной предметной области (информатика) и одного типа учебных работ (эссе). Генерализация выводов на другие дисциплины и типы текстов требует дополнительной валидации. Во-вторых, эксперимент проведен на однократной выборке объемом 477 эссе; оценка повторяемости результатов на независимых данных не проводилась. В-третьих, качество синтетической генерации зависит от используемых LLM и параметров промптирования; изменение этих параметров может повлиять на оптимальную пропорцию смешивания. В-четвертых, экспертные оценки. несмотря на высокую согласованность (коэффициент конкордации Кендалла W = 0,82), сохраняют элемент субъективности.

Заключение

В работе выполнено математическое моделирование процесса формирования гибридного корпуса образовательных текстов для задач автоматического анализа. На основе введенных формальных критериев качества (точность классификации K_{acc}, педагогическая ценность K_{ped}, лексическое разнообразие K_{lex}) построена зависимость интегрального показателя $\Psi(\beta)$ от суммарной доли синтетических данных β.

Проведен вычислительный эксперимент, включающий генерацию текстов пятью LLM (GPT-4, Grok, DeepSeek, Gemini, Cogito), формирование шести гибридных датасетов с различными β и оценку качества корпусов. Численное решение задачи нахождения наилучшего соотношения выполнено методом квадратичной аппроксимации экспериментальных точек.

Основные результаты:

1. Наилучшее качество среди чистых LLM показали модели GPT-4 и DeepSeek. Комбинирование этих моделей в пропорции 60:40 (микс С) позволило снизить долю фактических ошибок до 10,36 % и логических ошибок до 11,33 %.

2. Аппроксимация зависимости $\Psi(\beta)$ квадратичной функцией дала значение $\beta^* = 0{,}598$ с доверительным интервалом [0,57; 0,63], что соответствует оптимальной конфигурации «40 % реальных + 60 % синтетических текстов». Полученное значение $\beta = 0{,}598$ с высокой точностью совпадает с экспериментальным максимумом при $\beta = 0{,}6$, что подтверждает адекватность предложенной квадратичной модели. При этом аппроксимация позволяет утверждать, что истинный оптимум лежит внутри интервала [0,57; 0,63], а не на его границе, что невозможно было бы установить, опираясь только на дискретные экспериментальные данные.

3. Анализ чувствительности показал устойчивость найденного решения при варьировании объема выборки, числа признаков и температуры генерации.

4. Разработанный программный комплекс обеспечивает воспроизводимость вычислительного эксперимента и может быть адаптирован для других предметных областей и типов учебных текстов.

Полученные количественные закономерности позволяют сократить затраты на создание размеченных корпусов за счет

замещения до 60 % реальных данных синтетическими без потери качества автоматического анализа.

Дальнейшие исследования планируется направить на (1) проверку повторяемости на независимых выборках; (2) расширение предметной области; (3) использование более широкого набора LLM; (4) создание автоматизированной системы формирования гибридных корпусов.

Список литературы

1. Дюличева Ю. Ю. Применение учебной аналитики в высшем образовании: датасеты, методы и инструменты // Высшее образование в России. 2024. Т. 33. № 5. С. 86–111. DOI: 10.31992/0869-3617-2024-33-5-86-111.

2. Скворчевский К. А., Дятлова О. В. Современные адаптивные и интеллектуальные цифровые системы обучения: механизмы и потенциал // Вопросы образования // Educational Studies Moscow. 2024. № 3 (2). С. 299–337. DOI: 10.17323/vo-2024-19751.

3. Kostopoulos G., Tsiakmaki M., Kotsiantis S. Benchmarking Statistical and Deep Generative Models for Privacy-Preserving Synthetic Student Data in Educational Data Mining // Algorithms. 2026. Vol. 19 (1). P. 39. DOI: 10.3390/a19010039.

4. Илюшин Л. С., Торпашева Н. А. Технологии искусственного интеллекта как ресурс трансформации образовательных практик // Ярославский педагогический вестник. 2024. № 3 (138). С. 62–71. DOI: 10.20323/1813-145X-2024-3-138-62.

5. Rostam Z. R. K., Kertész G. Advances in Pre-trained Language Models for Domain-Specific Text Classification: A Systematic Review // ACM Transactions on Intelligent Systems and Technology. 2025. Vol. 16. Is. 6. P. 1–41. DOI: 10.1145/3763002.

6. Ma T. Systematically Visualizing ChatGPT Used in Higher Education: Publication Trend. Disciplinary Domains. Research Themes. Adoption and Acceptance // Computers and Education: Artificial Intelligence. 2025. Vol. 8. P. 100336. DOI: 10.1016/j.caeai.2024.100336.

7. Sun J., Song T., Peng W., Song J. A survey of automated essay scoring: Challenges, advances, and future // Neurocomputing. 2025. Vol. 650. 130916. DOI: 10.1016/j.neucom.2025.130916.

8. Prostakov O., Hodlevskyi V., Bouarour N., Sanchez-Ayte A., Ibrahim N., Amer-Yahia S. Reducing Human Effort in Evaluating Small and Medium Language Models as Students and as Teachers // 6th Workshop on Data Science with Human in the Loop (DaSH@VLDB). London. 2025. [Электронный ресурс]. URL: https://openreview.net/pdf?id=CG7DUrQjPQ (дата обращения: 12.03.2026).

9. Stanja J., Dannemann S., Krugel J., Hoppe A. Investigating Evidence-Oriented Generation of Synthetic Text Data with a Generative Large Language Model in Science Education // International Journal of Science Education. 2025. P. 1–23. DOI: 10.1080/09500693.2025.2538834.

10. Leinonen J., Denny P., Kiljunen O., MacNeil S., Sarsa S., Hellas A. LLM-itation is the Sincerest Form of Data: Generating Synthetic Buggy Code Submissions for Computing Education // Proceedings of the 27th Australasian Computing Education Conference (ACE 2025). ACM. 2025. P. 56–63. DOI: 10.1145/3716640.3716647.

11. Stefanovič P., Radvilaitė U., Pliuskuvienė B., Ramanauskaitė S. The influence of Gen-AI tools application for text data augmentation: case of Lithuanian educational context data classification // Scientific Reports. 2025. Vol. 15. Article number 26010. DOI: 10.1038/s41598-025-11877-z.

12. Ara S. J. S., Ramachandriah T., Haladappa M. S. Predictive model to analyze real and synthetic data for learners‘ performance prediction using regression techniques // Online Learning. 2025. Vol. 29. Is. 1. URL: https://olj.onlinelearningconsortium.org/index.php/olj/article/view/4390 (дата обращения: 12.03.2026). DOI: 10.24059/olj.v29i1.4390.

13. Nadăș M., Dioșan L., Tomescu A. Synthetic Data Generation Using Large Language Models: Advances in Text and Code // IEEE Access. 2025. Vol. 13. P. 134615–134633. DOI: 10.1109/ACCESS.2025.3589503.

14. Akaike H. A new look at the statistical model identification // IEEE Transactions on Automatic Control. 1974. Vol. 19. Is. 6. P. 716–723. DOI: 10.1109/TAC.1974.1100705.

15. Flores J. E., Cavanaugh J. E., Neath A. A. A New Class of Information Criteria for Improved Prediction in the Presence of Training/Validation Data Heterogeneity // Computational Statistics. 2025. Vol. 40. Is. 5. P. 2389–2423. DOI: 10.1007/s00180-024-01559-1.

16. Acquah D.-G. H. The Effect of Outliers on the Performance of Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) in Selection of an Asymmetric Price Relationship // Russian Journal of Agricultural and Socio-Economic Sciences. 2017. Vol. 65. Is. 5. P. 32–37. DOI: 10.18551/rjoas.2017-05.05.

Конфликт интересов: Авторы заявляют об отсутствии конфликта интересов.

Conflict of interest: The authors declare that there is no conflict of interest.

Финансирование: Авторы заявляют об отсутствии внешнего финансирования.

Financing: The research was performed without external funding.