MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 33, NO. 1, 2026
journal homepage: www.mais-journal.ru

ARTIFICIAL INTELLIGENCE

The Impact of the Size of Training Sets on Quality of Automatic Short Answers Grading

L. S. Rogulin[1], A. Y. Poletaev[1], K. V. Lagutina[1]

DOI: 10.18255/1818-1015-2026-1-62-77

[1]P.G. Demidov Yaroslavl State University, Yaroslavl, Russia

MSC2020: 68T50
Research article
Full text in Russian

Received January 16, 2026
Revised February 23, 2026
Accepted February 27, 2026

The paper investigates the impact of training set size on the quality of automatic short answers grading, formulated as a classification task. The impact was evaluated using a method based on measuring the similarity between the assessed answer and a given reference answer, calculated via embedding vectors, in combination with a logistic regression classifier. Experiments were conducted on corpora of answers to questions in computer science, history, and software development using Qt framework. The sizes of the corpora were 547, 522, and 931 answers, respectively. Two experiments were conducted during the study. In the first experiment, the change in classification quality was assessed as the training set size was reduced. It showed that when the binary classification is utilized (an answer can be either correct or incorrect), reducing the size of the training set leads to a smaller decline in quality compared to ternary classification (which includes a class of partially correct answers). In the second experiment, the possibility of improving classification quality by expanding small-sized training sets through data augmentation was investigated. It demonstrated that augmentation performed using the DeepSeek generative model can significantly improve results in several cases, which is important for practical applications under data scarcity conditions. Additionally, the experiments revealed that when different language models are used to generate embeddings, the magnitude of change in classification quality with varying training set sizes can differ significantly. Specifically, using certain models — such as rubert-tiny2 and MiniLM-L12-v2 — to produce embeddings yields more stable results than using other models.

Keywords: ASAG; data augmentation; text classification; neural network language models; assessing students' open responses; artificial intelligence in education

INFORMATION ABOUT THE AUTHORS

Rogulin, Lev S.	ORCID iD: 0009-0000-9551-6007. E-mail: rogulev0805@gmail.com Student
Poletaev, Anatoliy Y. (corresponding author)	ORCID iD: 0000-0003-0116-4739. E-mail: anatoliy-poletaev@mail.ru PhD, Senior Lecturer
Lagutina, Ksenia V.	ORCID iD: 0000-0002-1742-3240. E-mail: lagutinakv@mail.ru PhD, Associate Professor

Funding: Yaroslavl State University (project VIP-021).

For citation: L. S. Rogulin, A. Y. Poletaev, and K. V. Lagutina, "The impact of the size of training sets on quality of automatic short answers grading", *Modeling and Analysis of Information Systems*, vol. 33, no. 1, pp. 62–77, 2026.
DOI: 10.18255/1818-1015-2026-1-62-77.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 33, № 1, 2026
сайт журнала: www.mais-journal.ru

Влияние размера обучающей выборки на качество автоматической оценки коротких ответов

Л. С. Рогулин[1], А. Ю. Полетаев[1], К. В. Лагутина[1]

DOI: 10.18255/1818-1015-2026-1-62-77

[1]Ярославский государственный университет им. П.Г. Демидова, Ярославль, Россия

УДК 004.912
Научная статья
Полный текст на русском языке

Получена 16 января 2026 г.
После доработки 23 февраля 2026 г.
Принята к публикации 27 февраля 2026 г.

В работе исследуется влияние объёма обучающей выборки на качество автоматического оценивания правильности коротких ответов, представленного в виде задачи классификации. Влияние оценивалось на примере метода, основанного на оценке сходства между оцениваемым ответом и заданным эталонным ответом, рассчитываемого с помощью векторов эмбеддингов, и классификатора на основе логистической регрессии. Эксперименты проводились на корпусах ответов студентов на вопросы по компьютерным наукам, истории и разработке на Qt. Объём корпусов составил 547, 522 и 931 ответ соответственно. В ходе выполнения работы было поставлено два эксперимента. В ходе первого эксперимента оценивалось изменение качества классификации при уменьшении объёма обучающей выборки. Он показал, что при бинарной классификации (когда ответ может быть либо верным, либо неверным) уменьшение объёма обучающей выборки классификатора приводит к меньшему снижению качества, чем при тернарной классификации (когда выделяется класс частично верных ответов). В ходе второго эксперимента изучалась возможность повышения качества классификации за счёт расширения обучающей выборки малого объёма с помощью аугментации. Он показал, что аугментация, выполненная с помощью генеративной модели DeepSeek, позволяет в ряде случаев значительно улучшить результат, что представляется важным для практического применения в условиях дефицита данных. Также в ходе экспериментов было выявлено, что при использовании для генерации эмбеддингов различных языковых моделей величина изменения качества классификации при изменении объёма обучающей выборки может существенно различаться: при использовании некоторых для получения эмбеддингов моделей rubert-tiny2 и MiniLM-L12-v2 результаты оказываются более стабильными, чем при использовании других моделей.

Ключевые слова: автоматическая оценка коротких ответов; аугментация данных; классификация текстов; нейросетевые языковые модели; оценка открытых ответов учащихся; искусственный интеллект в образовании

ИНФОРМАЦИЯ ОБ АВТОРАХ

Рогулин, Лев Сергеевич	ORCID iD: 0009-0000-9551-6007. E-mail: rogulev0805@gmail.com Студент
Полетаев, Анатолий Юрьевич (автор для корреспонденции)	ORCID iD: 0000-0003-0116-4739. E-mail: anatoliy-poletaev@mail.ru Канд. тех. наук, старший преподаватель
Лагутина, Ксения Владимировна	ORCID iD: 0000-0002-1742-3240. E-mail: lagutinakv@mail.ru Канд. тех. наук, доцент

Финансирование: ЯрГУ (проект VIP-021).

Для цитирования: L. S. Rogulin, A. Y. Poletaev, and K. V. Lagutina, “The impact of the size of training sets on quality of automatic short answers grading”, *Modeling and Analysis of Information Systems*, vol. 33, no. 1, pp. 62–77, 2026.
DOI: 10.18255/1818-1015-2026-1-62-77.

Введение

В последние годы задачи автоматической оценки коротких ответов приобретают всё большую актуальность благодаря развитию систем автоматизированного обучения и интеллектуальных помощников [1]. Такая автоматизация позволяет освободить преподавателя от рутинной работы и дать студенту быструю обратную связь.

Одной из ключевых задач при создании методов автоматической оценки является отделение верных, то есть совпадающих по смыслу с эталонными, ответов от неверных. Для её решения часто используются языковые модели или методы, комбинирующие языковые модели с лингвистическими характеристиками, которые позволяют достичь F1-меры 0.7–0.8, а иногда — около 0.9 [2], но требуют обучения на достаточно крупных размеченных корпусах.

Многие исследователи работают или с имеющимися открытыми корпусами ответов, преимущественно англоязычными Mohler и SciEntsBank [3], или собирают собственные корпуса, обычно небольшие. Перед исследователями, как правило, возникает крайне сложная задача выбрать такой размера корпуса, чтобы, с одной стороны, его можно было качественно разметить при имеющихся ресурсах, а с другой — он позволил бы достаточно качественно обучить модель [4, 5].

Под аугментацией в данной работе понимается технология создания синтетических данных на основе имеющегося набора данных. Эти синтетические данные обычно содержат небольшие изменения, к которым предсказания модели должны быть инвариантны. Аугментация позволяет предотвратить переобучение нейросетевых алгоритмов и создать сбалансированную выборку текстов [6]. В современных исследованиях расширение корпусов за счёт аугментации редко применяется для задач обработки открытых ответов, особенно для русскоязычных текстов. Поэтому анализ влияния аугментации на качество оценки коротких ответов является актуальной задачей.

Цель данной работы — исследовать влияние объема обучающей выборки на качество автоматической оценки коротких ответов с использованием нейросетевых моделей. В ходе работы поставлены два эксперимента: первый направлен на изучение влияния размера обучающей выборки на качество автоматической оценки коротких ответов, а второй — на возможность повышения качества оценки коротких ответов за счёт расширения обучающей выборки с помощью аугментации. Необходимо отметить, что сравнение методов автоматической оценки правильности коротких ответов, так же как и создание нового метода решения этой задачи, выходит за пределы работы.

Работа организована следующим образом. В разделе 1 приведён обзор существующих методов, которые могут быть использованы для решения поставленной задачи. Раздел 2 описывает использованный в работе метод автоматической оценки правильности коротких ответов. В разделе 3 описаны корпуса коротких ответов, используемые для экспериментов. В разделе 4 приводятся результаты эксперимента по изучению влияния объёма обучающей выборки на качество оценки коротких ответов. В разделе 5 описан эксперимент, направленный на исследование влияния аугментации на качество оценки. В заключении обсуждаются полученные в ходе экспериментов результаты и подводятся итоги.

1. Обзор аналогичных научных исследований

Автоматическая оценка коротких ответов в последние годы наиболее часто решается с помощью больших языковых моделей [2]. Для англоязычных текстов лучшие результаты показывают предобученные языковые модели на основе BERT. Например, для RoBERTa F1-мера достигает около 0.80–0.90 на разных корпусах [7]. Модели генеративного искусственного интеллекта также активно применяются в исследованиях, но показывают результаты существенно ниже. На открытом корпусе SciEntsBank GPT-4 обеспечивает F1-меру около 0.74 , в то время как BERT и RoBERTa демонстрируют F1-меру в диапазоне 0.73–0.81. На открытом корпусе Beetle F1-мера у GPT-4 варьируется в диапазоне 0.52–0.61, а у BERT и RoBERTa — 0.73–0.91 [8].

Русскоязычные короткие ответы оцениваются автоматически при помощи BERT-эмбеддингов и ключевых слов с долей правильных ответов классификатора 0.77–0.90 [9]. В целом и для английского, и для других национальных языков качество классификации достигает более 0.9 для стандартных метрик качества только в отдельных случаях, но не на корпусах текстов в целом [2].

Повысить качество классификации может расширение корпуса. Нейросетевые классификаторы и методы машинного обучения обычно показывают более высокие результаты при увеличении размера корпуса для многих задач обработки текстов, например, для анализа тональности и определения информативности [4].

Так как сбор и разметка ответов обучающихся требует значительных трудозатрат от преподавателей, быстрее всего расширение корпуса может быть произведено путём автоматической аугментации данных. В некоторых случаях расширение корпуса может быть выполнено за счёт комбинирования реальных данных. В работе [3] аугментация выполнялась для корпусов Mohler и SemEval Task 7, содержащий около 3500 пар ответов по естественнонаучным дисциплинам. Исходными данными для классификации с помощью сиамских нейронных сетей служили пары ответ — эталон. Авторы считали ответы студентов с высшей оценкой как эталонные и формировали новые пары ответ — ответ с высшей оценкой. Таким образом данные оставались реальными, но количество данных для обучения увеличилось с 2500 и 5000 до нескольких десятков тысяч. RMSE уменьшилась с 0.87 до 0.83 для корпуса Mohler и с 0.845–1.096 до 0.758–0.996 для корпуса SemEval.

Другим сравнительно простым подходом к аугментации является обратный перевод. Авторы работы [10] расширяли таким образом данные SciEntsBank. Бинарная классификация при помощи сверточной графовой сети достигла F1-меры, равной 0.716, а при помощи BERT — 0.710.

Более высоких результатов на корпусе SciEntsBank достигли Lun и др. [11]. Они разработали MDA-ASAS — комплекс стратегий для повышения эффективности автоматической оценки коротких ответов. MDA-ASAS улучшен с помощью подходов к аугментации, которые включают обратный перевод, использование правильного ответа в качестве эталонного ответа и перестановку контента. При экспериментах с бинарной классификацией и языковой моделью BERT при использовании аугментации была достигнута F1-мера около 0.82, тогда как без аугментации F1-мера была 0.70–0.79.

В последние годы аугментация данных часто осуществляется с помощью генеративного искусственного интеллекта. В работе [12] корпус текстов Mohler расширялся за счет использования GPT-4 для перефразирования ответа, что служило дополнительной стратегией замены синонимов. Размер корпуса увеличился с 2500 до около 6000 ответов и стал более сбалансированным по классам оценок. Без аугментации качество классификации при помощи языковой модели SentenceTransformers достигало F1-меры, равной 0.7, а с аугментацией — около 0.92.

Большие языковые модели применяются для аугментации текстовых данных, чтобы генерировать разнообразные тексты. Zhong и др. [13] предложили метод улучшения корпусов текстов путем интеграции фрагментов с различной лексикой в исходный набор данных при помощи большой языковой модели Qwen2.5. Авторы исследования [14] расширяли данные методом дополненного поиска, применив Sentence-BERT для извлечения знаний из Википедии. В фреймворке ReGen [15] использовался BERT для поиска по документам и расширения обучающей выборки. Доля правильных ответов для классификации по темам и тональности возросла на 2–5 % в среднем для различных корпусов. ChatGPT тоже может выступать аугментатором, увеличивая точность обработки клинических данных на 1–11 % для разных корпусов [16].

Таким образом, аугментация данных позволяет расширять корпуса текстов в несколько раз или даже на порядок и при этом повышать качество классификации. Для автоматической оценки коротких ответов аугментация применяется очень редко, в основном для открытых англоязычных корпусов. Поэтому синтетическая генерация русскоязычных текстов для расширения обучающей выборки является актуальной задачей.

2. Метод автоматической оценки коротких ответов

В экспериментах используется метод автоматической оценки правильности коротких ответов, представленной в виде задачи классификаци, основанный на оценке сходства между оцениваемым ответом и заданным эталонным ответом с помощью векторов эмбеддингов и классификатора на основе логистической регрессии. Как показано в работе [17], этот метод позволяет достигать достаточно высоких показателей качества, достаточных для практического применения. Ниже приведено его более подробное описание.

Метод сопоставляет каждому оцениваемому ответу его метку правильности, выполняя следующие шаги.

1. Для каждой пары текстов — оцениваемый ответ и эталонный ответ — вычисляются эмбеддинги с помощью языковых моделей, которые представляют ответы в виде векторов.
2. Эмбеддинги двух текстов конкатенируются в один вектор.
3. Вектор передаётся в классификатор на основе логистической регрессии, обученный на размеченных данных для предсказания метки правильности ответа.

Метод может применяться как для бинарной классификации, когда возможны два варианта метки правильности — «верный» и «неверный», так и для тернарной, с добавлением метки «частично верный».

Особенностью метода является то, что при сопоставлении ответов непосредственно вопрос не подаётся в модель. Это решение обусловлено тем, что эталонный ответ уже сформулирован преподавателем в контексте данного вопроса и содержит все ключевые элементы, необходимые для оценки правильности студенческого ответа. Таким образом, пара «ответ студента — эталон» содержит информацию, с помощью которой можно достаточно качественно оценивать семантическое сходство, а следовательно, и определять правильность ответа. Добавление вопроса как отдельного входа могло бы избыточно увеличить размерность признаков и усложнить модель.

Для того, чтобы изучить влияние размера обучающей выборки на качество оценки коротких ответов при получении эмбеддингов с помощью различных языковых моделей, в экспериментах использовались следующие предобученные модели, опубликованные на платформе HuggingFace[1]:

- ai-forever/rugpt2large — крупномасштабная русскоязычная авторегрессионная модель семейства ruGPT, построенная на архитектуре GPT-2 и предназначенная для генерации и обработки текстов на русском языке.
- ai-forever/rugpt3medium_based_on_gpt2 — модель ruGPT-3 Medium, основанная на архитектуре GPT-2 и обученная на расширенном русскоязычном корпусе; поддерживает увеличенную длину контекста и демонстрирует более высокую выразительность при работе с русским текстом.
- bert-base-multilingual-uncased — многоязычная модель семейства BERT, предобученная более чем на 100 языках. Использует архитектуру Transformer, что обеспечивает устойчивость к регистру и высокую универсальность.
- cointegrated/rubert-tiny2 — компактная русскоязычная модель на основе BERT, оптимизированная для задач получения эмбеддингов предложений. Обладает малым размером и высокой скоростью работы при сохранении качества семантического кодирования.
- sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 — модель семейства Sentence Transformers, обученная для многозадачного парафразирования на множестве языков. Формирует компактные эмбеддинги и обеспечивает хорошее качество семантического сравнения текстов.

Выбор конкретных моделей продиктован результатами аналогичного исследования [17], в котором авторы использовали именно эти архитектуры. Было решено повторить данный эксперимент

[1] https://huggingface.co/

на русскоязычных данных, чтобы проверить, насколько хорошо данные модели справляются с задачей ASAG при различных объёмах обучающей выборки.

Практическая реализация описанного метода была выполнена на языке программирования Python. Полный исходный код, использованный в экспериментах, находится в открытом доступе по адресу: https://github.com/RainbollD/Reasearch-ASAG.

3. Корпуса коротких ответов

В работе использовались три корпуса коротких ответов из числа корпусов, опубликованных преподавателями Ярославского государственного университета имени П. Г. Демидова в репозитории по адресу: https://gitverse.ru/Shtepser/asag-datasets. Разметка данных корпусов выполнялась их составителями по четырём классам: верные ответы, неверные ответы, частично верные ответы и ответы не по теме. Однако, поскольку ответов не по теме в данных корпусах крайне мало (менее десятка), при проведении экспериментов в рамках данной работы было принято решение включить класс ответов не по теме в класс неверных ответов, то есть привести корпуса к тернарной разметке.

Первый из использованных корпусов — 8_questions_extended — включает 547 ответов на 8 вопросов по различным темам компьютерных наук; составителями корпуса также опубликован его вариант с бинарной разметкой (ответ может быть либо верным, либо неверным). Второй из корпусов — history_summer_2025 — содержит 522 ответа на 12 вопросов по истории России. Третий — qt_questions — включает 931 ответ на 12 вопросов, связанных с разработкой приложений на фреймворке Qt.

Распределение ответов по меткам для всех корпусов представлено в таблице 1. В ней представлены соотношения между названием корпуса, номером вопроса (файлом, содержащим данные по этому вопросу) и количеством ответов, распределенных в зависимости от степени их схожести. Таблица демонстрирует значительное доминирование верных ответов над всеми остальными. Особенно выраженный дисбаланс наблюдается в корпусе qt_questions, где для нескольких вопросов доля полностью корректных ответов превышает 70–80 %. Аналогичная тенденция отмечается и для корпуса history_summer_2025. В корпусе 8_questions_extended соотношение меток различается между вариантами разметки: тернарная версия содержит умеренную долю частично корректных ответов, в то время как бинарная версия характеризуется симметричным распределением классов.

Для решения практических задач полезно оценить влияние объёма обучающей выборки на качество определения правильности ответа как для случая, когда ответы считаются либо верными, либо неверными (бинарная классификация), так и для случая, когда выделяется класс частично верных ответов (тернарная классификация). Для корпуса 8_questions_extended оба варианта разметки — бинарная и тернарная — уже предоставлены его создателями; корпуса же history_summer_2025 и qt_questions исходно содержат только тернарную разметку. Чтобы поставить эксперименты с бинарной классификацией на всех трёх корпусах, тернарная разметка корпусов history_summer_2025 и qt_questions преобразуется в бинарную с помощью включения класса частично верных ответов в класс неверных ответов; класс верных ответов при этом сохраняется неизменным.

Для разделения корпуса данных на выборки была применена стратифицированная кросс-валидация с числом фолдов, равным 5. В процессе кросс-валидации были настроены гиперпараметры, включая случайное перемешивание данных, а также установлено начальное случайное значение 42.

Важно отметить, что обучение классификатора на основе логистической регрессии проводилось на данных, относящихся исключительно к конкретному вопросу, представленному в каждом файле, а не на данных всего корпуса, что позволяет более точно оценить их способность решать задачу классификации для каждого конкретного вопроса. Значения метрик для всех вопросов усреднялись для получения общей метрики качества.

Table 1. Combined table of score distributions across corpora

Таблица 1. Объединённая таблица распределений оценок по корпусам

Название корпуса	Номер вопроса	-1	0	1
8_questions_extended (ternary)	1	1	23	21
	2	5	20	121
	3	4	14	18
	4	0	10	26
	5	8	10	25
	6	11	22	10
	7	4	16	25
	8	12	17	124
8_questions_extended (binary)	1	0	44	44
	2	0	141	141
	3	0	33	33
	4	0	36	36
	5	0	35	35
	6	0	32	32
	7	0	41	41
	8	0	141	141
history_summer_2025	1	1	15	29
	2	15	12	17
	3	3	6	32
	4	2	4	36
	5	4	13	26
	6	2	4	35
	7	5	21	20
	8	2	8	35
	9	3	11	29
	10	0	6	40
	11	3	9	33
	12	3	11	27
qt_questions	1	0	8	22
	2	0	103	21
	3	6	81	28
	4	5	23	69
	5	0	1	84
	6	13	36	42
	7	2	4	6
	8	0	9	80
	9	0	0	5
	10	6	30	85
	11	2	8	30
	12	10	27	83

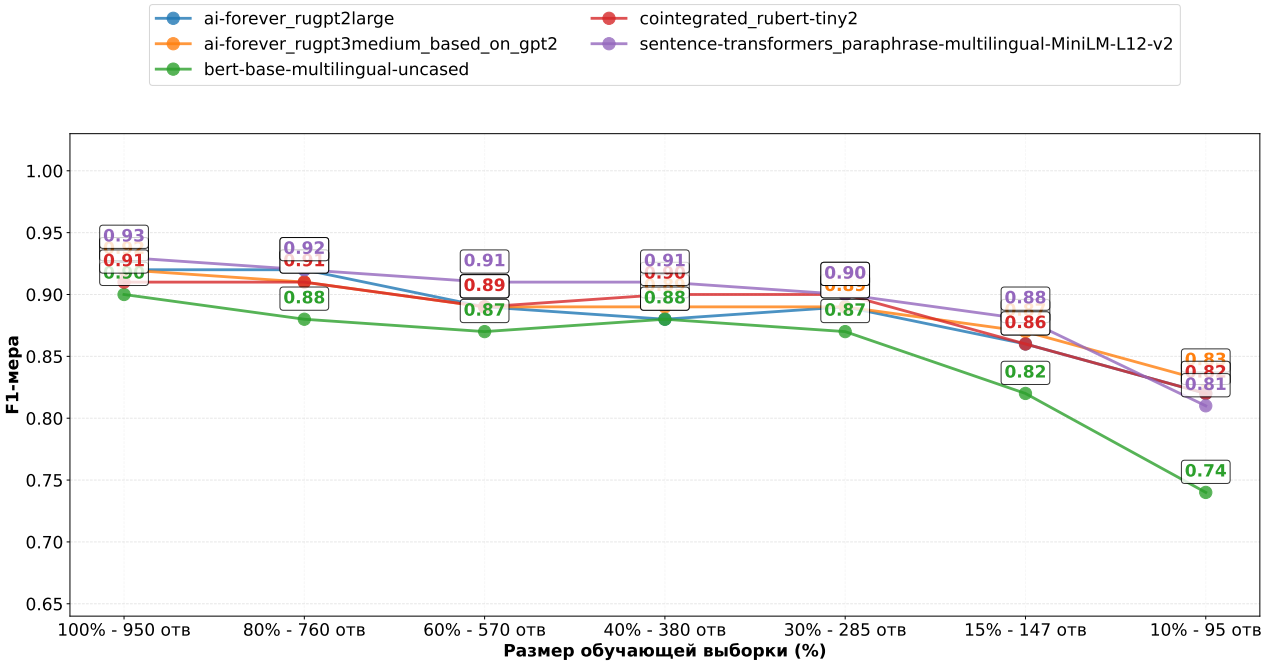

Fig. 1. F1-score on the 8_questions_extended corpus (binary classification)

Рис. 1. F1-мера на корпусе 8_questions_extended (бинарная классификация)

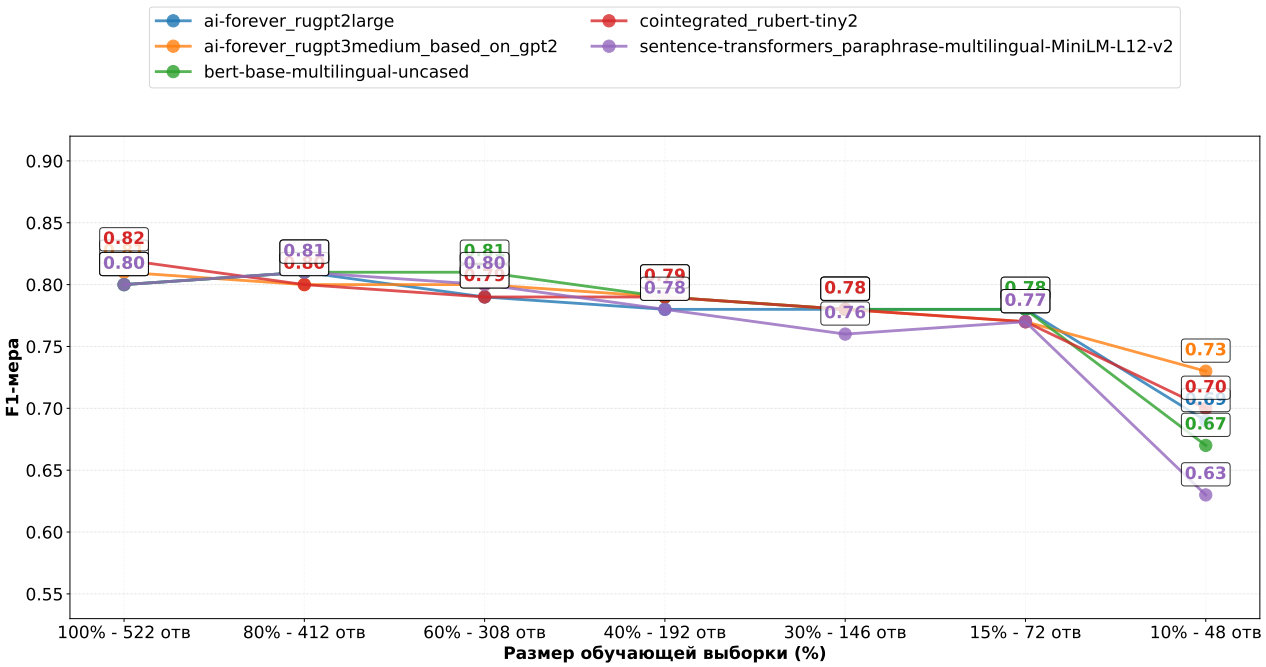

Fig. 2. F1-score on the history_summer_2025 corpus (binary classification)

Рис. 2. F1-мера на корпусе history_summer_2025 (бинарная классификация)

4. Влияние объёма обучающей выборки на качество определения оценки

Первый эксперимент, поставленный в ходе исследования, направлен на изучение влияния размера обучающей выборки на качество автоматической оценки коротких ответов. Он организован в несколько этапов, на каждом из которых классификатор обучался на уменьшенном объёме дан-

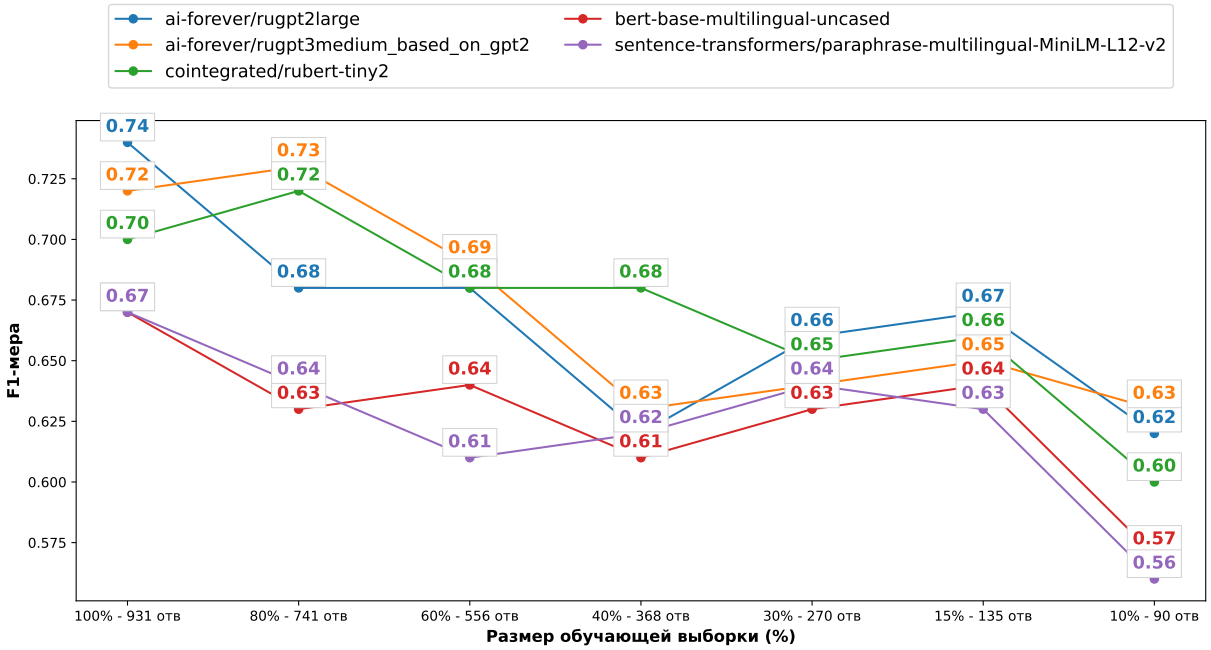

Fig. 3. F1-score on qt_questions corpus (binary classification)

Рис. 3. F1-мера на корпусе qt_questions (бинарная классификация)

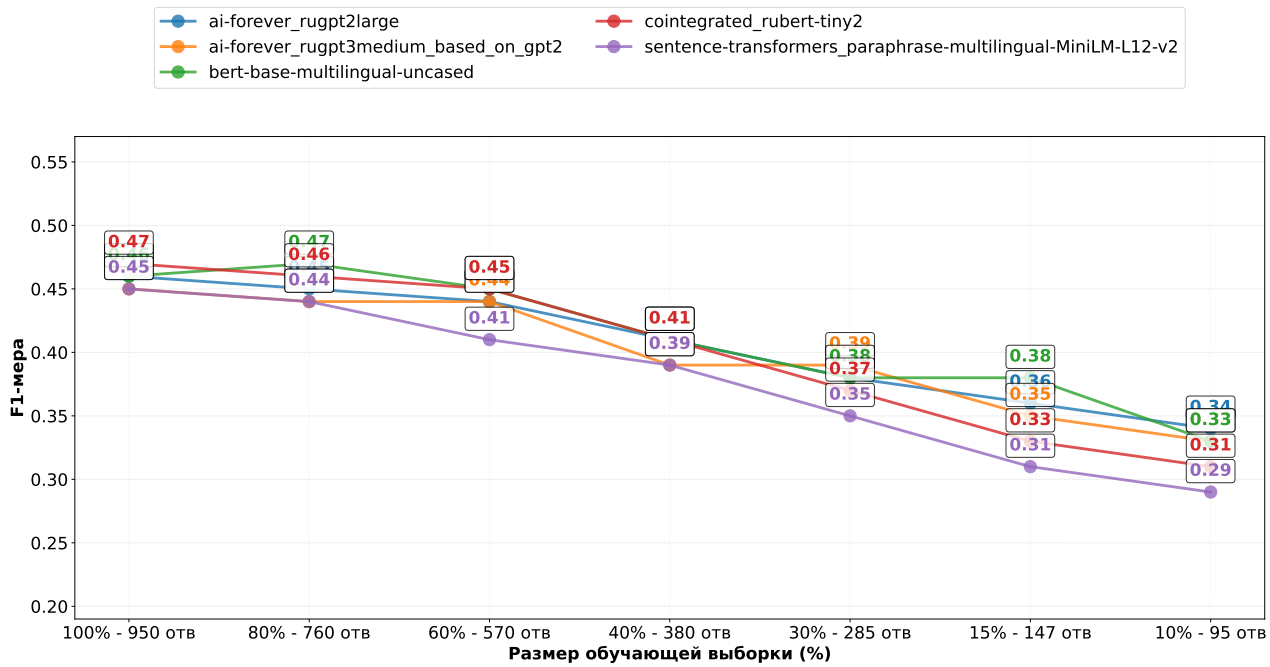

Fig. 4. F1-score on the 8_questions_extended corpus (ternary classification)

Рис. 4. F1-мера на корпусе 8_questions_extended (тернарная классификация)

ных. Размер обучающей выборки на каждом из этапов составлял 1.0, 0.8, 0.6, 0.4, 0.3, 0.15 и 0.1 от объёма исходной обучающей выборки. Тестовая выборка, на которой оценивалось качество классификатора, оставалась неизменной, так же как и гиперпараметры при обучении.

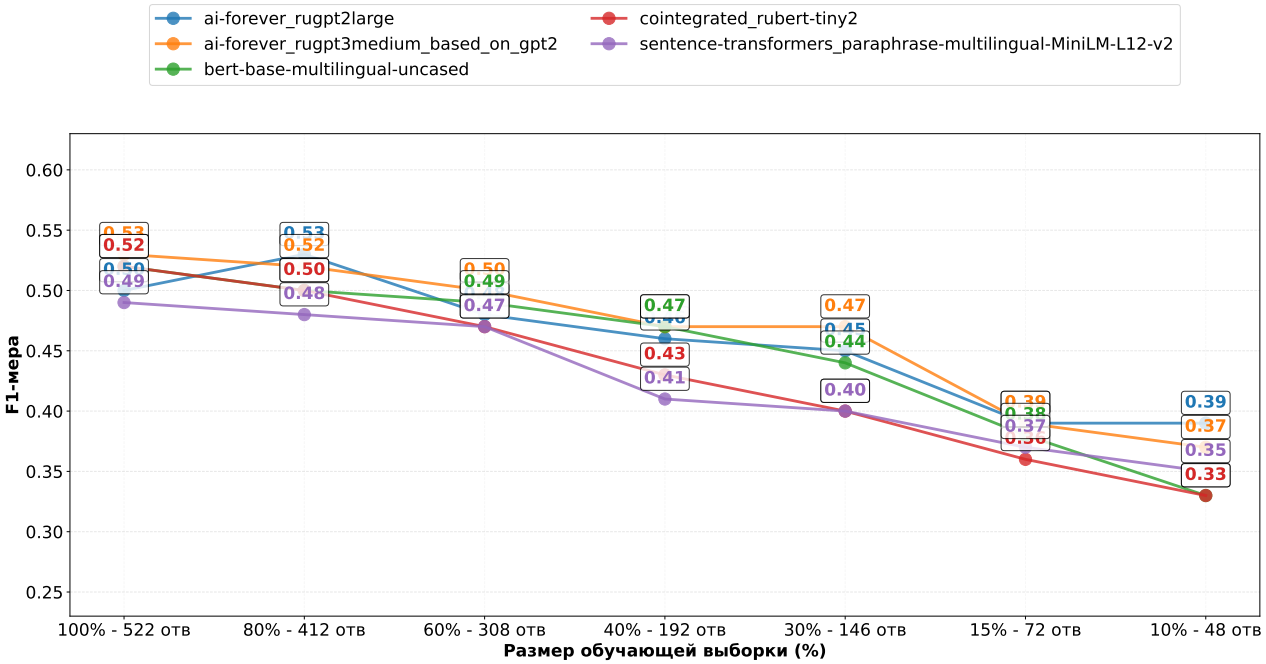

Fig. 5. F1-score on the history_summer_2025 corpus (ternary classification)

Рис. 5. F1-мера на корпусе history_summer_2025 (тернарная классификация)

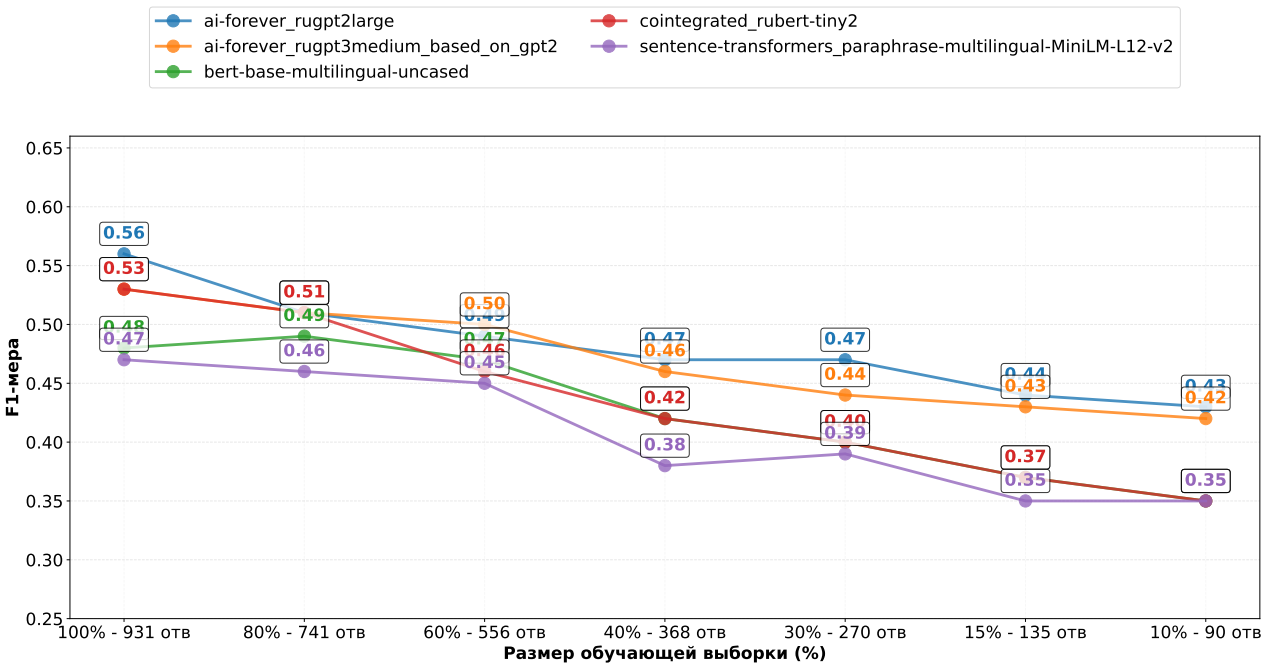

Fig. 6. F1-score on the qt_questions corpus (ternary classification)

Рис. 6. F1-мера на корпусе qt_questions (тернарная классификация)

Проведённый эксперимент показал, что сокращение объёма обучающей выборки закономерно приводит к снижению качества работы моделей. При этом зависимость между качеством классификации и количеством обучающих данных имеет нелинейный характер. Наиболее интенсивное

Table 2. Changes in F1-score when reducing the training set size

Таблица 2. Изменение F1-меры при уменьшении размера обучающей выборки

Набор данных	Классификация	Изменение F-меры при уменьшении размера обучающей выборки	
		до 30 % от исходного	до 10 % от исходного
8_questions_extended	бинарная	-0.04	-0.13
	тернарная	-0.20	-0.30
history_summer_2025	бинарная	-0.03	-0.15
	тернарная	-0.20	-0.31
qt_questions	бинарная	-0.07	-0.15
	тернарная	-0.17	-0.25

падение качества наблюдается при сокращении выборки до 30 % от исходного объёма и ниже; до этого темп снижения метрик значительно меньше, что хорошо заметно по графикам (см. рис. 1–2).

Анализ результатов бинарной классификации демонстрирует, что качество остаётся относительно стабильным вплоть до использования 15–30 % обучающих данных. После этого порога наблюдается резкое ухудшение метрик, что свидетельствует о недостаточном объёме информации для корректного обучения. Особенно сильное снижение наблюдается на корпусах 8_questions_extended и history_summer_2025.

Для задач тернарной классификации картина выглядит несколько иначе. Здесь метрики начинают снижаться уже при уменьшении выборки до 60–80 % от исходного объёма (см. рис. 4–6). Это свидетельствует о более высоких требованиях многоклассовых классификаторов к количеству обучающих примеров: модели требуют более крупных выборок для корректного различения нескольких категорий ответов.

Как показано в таблице 2, процентные изменения F1-меры при уменьшении обучающей выборки с 100 % до 30 % значительно варьируется в зависимости от корпуса и типа классификации. Для корпуса 8_questions_extended в случае бинарной классификации это снижение составило 3.5 %, а для тернарной классификации — 20 %. В то же время, при уменьшении обучающей выборки до 10 % для этого корпуса наблюдается более значительное ухудшение, особенно для тернарной классификации, где снижение F1-меры достигло 30 %. Для корпуса history_summer_2025 снижение составило 2.5 % (бинарная классификация) и до 31 % для тернарной, что также подтверждает значительное ухудшение качества при уменьшении обучающей выборки до 10 %.

Интересным результатом эксперимента стало то, что разные архитектуры моделей демонстрировали неодинаковую устойчивость к уменьшению объёма обучающих данных. Наиболее стабильные результаты показали модели rubert-tiny2 и MiniLM-L12-v2, которые даже при существенном сокращении выборки сохраняли приемлемый уровень F1-меры. Это может объясняться их эффективными архитектурными решениями, обеспечивающими лучшую способность к обобщению при ограниченном количестве данных.

В целом эксперимент показал, во-первых, нелинейность снижения качества оценки правильности коротких ответов при изменении объёма обучающей выборки, во-вторых, что при использовании для получения векторов эмбеддингов одних моделей качество снижается существенно медленнее, чем при использовании других.

5. Влияние аугментации на качество определения оценки

Идея второго эксперимента заключается в проверке возможности синтетического расширения корпуса с помощью аугментации для компенсации нехватки данных. Аугментация выполнялась с помощью генеративной модели DeepSeek версии V3.2-Exp [18]. Все эксперименты с аугментацией

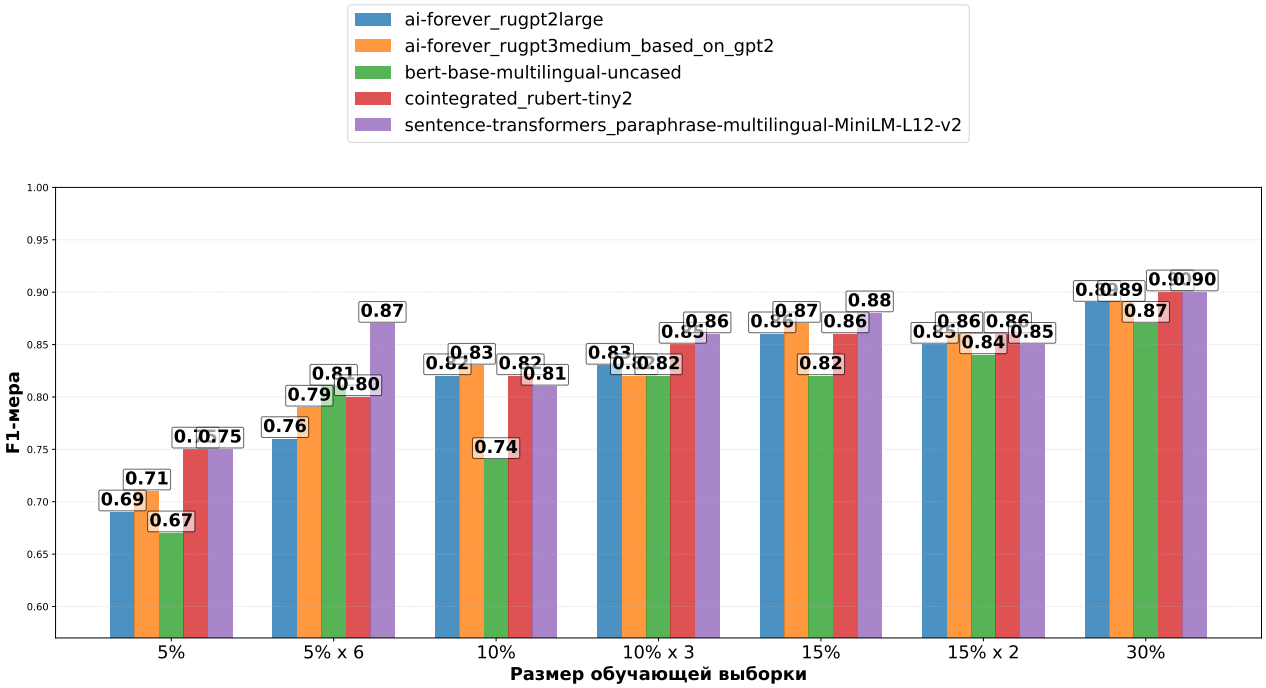

Fig. 7. Comparison of model performance depending on data volume and augmentation: 8_questions (binary classification)

Рис. 7. Сравнение качества моделей в зависимости от объема данных и аугментации: 8_questions (бинарная классификация)

проводились только для бинарной классификации, поскольку влияние аугментации требуется изучить в том числе для крайне малых обучающих выборок, размером до 5 % от исходного объёма, а из-за дисбаланса классов в корпусах при использовании трёх классов для многих вопросов при таком сильном снижении объёма не окажется примеров для одного из классов.

На вход генеративной модели подавались выборки уменьшенного размера, составляющие 5 %, 10 %, 15 % от объёма реальных обучающих выборок, использованных в первом эксперименте. Также на вход DeepSeek был подан промпт, с помощью которого имеющиеся выборки аугментировались до размера в 30 % от размера реальной обучающей выборки (то есть в 6, 2 и 3 раза соответственно), поскольку в ходе первого эксперимента именно при уменьшении выборки менее, чем до 30 % от исходного размера начиналось сильное снижение качества.

Промпт:

> *Я дал тебе файлы с данными коротких ответов. В файлах есть answer — ответ, который тебе в будущем нужно будет генерировать, true_answer — эталонный ответ, с которым необходимо сравнивать, и mark — оценка, где 1 — похож, 0 — нет. Твоя задача — прочитать файлы, проанализировать answer и на основе их составить в n раз больше данных, не забыв в начале написать данные из файла и дополнив их так, чтобы суммарно данных стало ровно в n раз больше.*

Пример вопроса, реального и сгенерированного ответов.

Вопрос: Какие возможности своим наследникам даёт QWidget?’

- **Оригинальный пример:** «Он даёт возможность обрабатывать события, например нажатие мыши, нажатие клавиатуры. Все кнопки и другие части интерфейса сделаны на его основе. Его наследники могут отображаться на экране и реагировать на действия пользователя. Можно изменять их внешний вид».

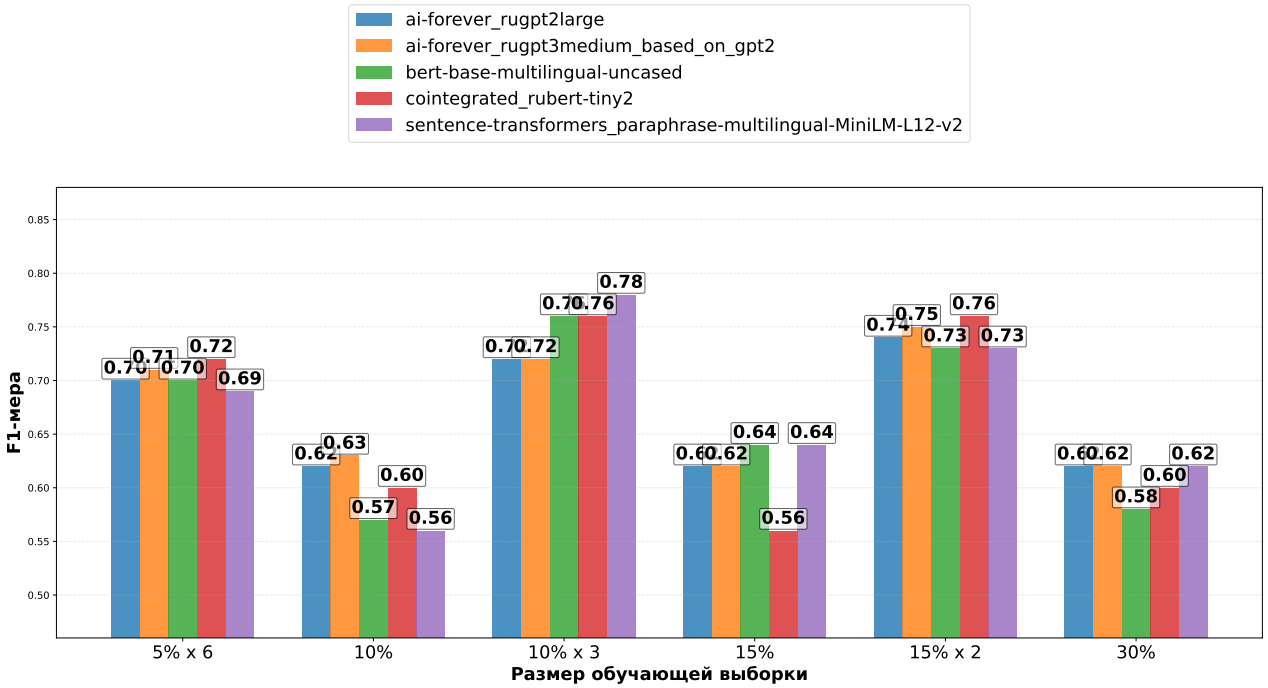

Fig. 8. Comparison of model performance depending on data volume and augmentation: qt_questions (binary classification)

Рис. 8. Сравнение качества моделей в зависимости от объема данных и аугментации: qt_questions (бинарная классификация)

- **Сгенерированный пример DeepSeek:** «Наследники QWidget могут использовать систему искусственного интеллекта: интеграция с ML-моделями, отображение результатов классификации, работа с компьютерным зрением, а также создание интерактивных систем рекомендаций».

Для бинарной классификации на корпусах qt_questions и history_summer_2025 не оценивались результаты при использовании 5 % от обучающей выборки, так как отдельные классы либо полностью отсутствовали, либо были представлены единичными экземплярами, что приводило к нестабильности обучения и фактически случайному характеру предсказаний.

В ходе эксперимента были построены графики зависимости значения F1-меры от размера обучающей выборки (см. рис. 7, 8 и 9). Их анализ показал, что аугментация исходного набора данных повышает качество классификации, особенно при малых объёмах корпуса.

Например, на корпусе 8_questions_extended аугментация данных при использовании 5 % исходной выборки (увеличение в 6 раз) обеспечила наибольший прирост F1-меры. В других случаях добавление новых примеров также дало положительный, хотя и менее выраженный эффект.

Следует отметить, что влияние аугментации данных могло оказаться как положительным, так и отрицательным. На корпусе qt_questions, наблюдался значительный рост качества: любая аугментация обеспечивала увеличение F1-меры примерно в 2 раза по сравнению с исходным объёмом обучающих данных. В то же время на корпусе history_summer_2025 имелись случаи ухудшения качества. Например, при использовании 15 % исходной выборки и её удвоении (15 %×2) некоторые модели продемонстрировали снижение метрики.

Таким образом, результаты эксперимента подтверждают, что аугментация данных может быть крайне полезной при ограниченном объёме обучающих данных.

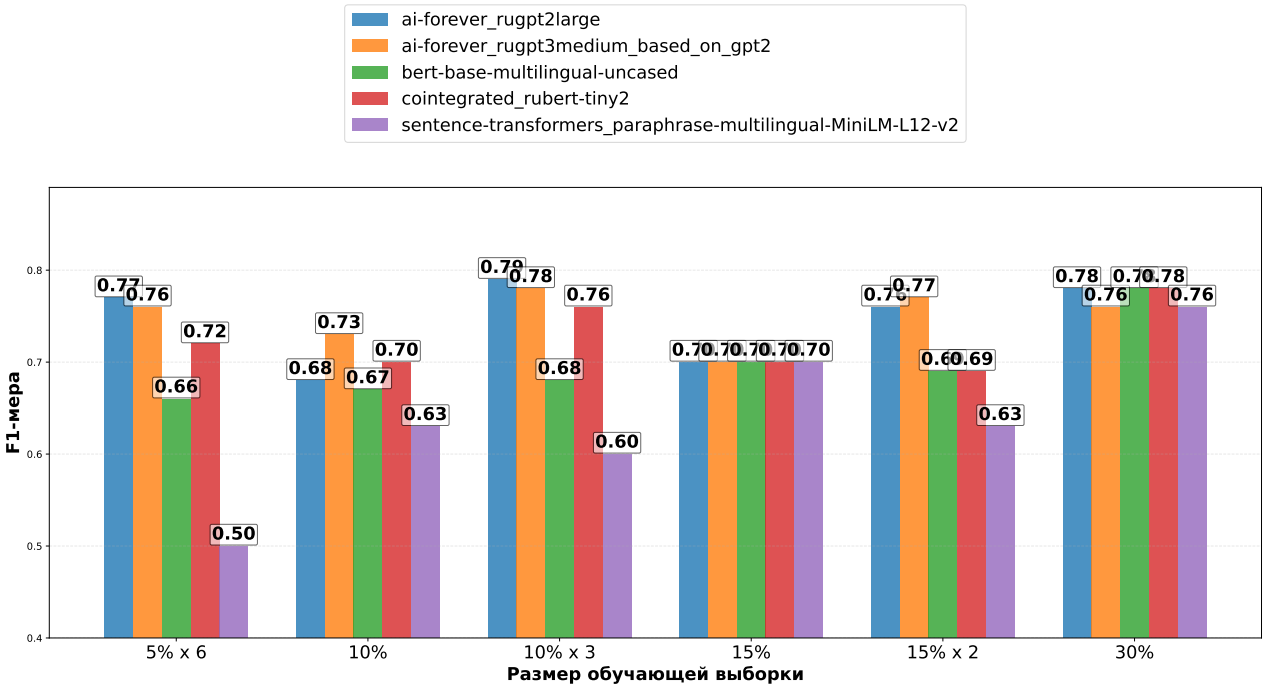

Fig. 9. Comparison of model performance depending on data volume and augmentation: history_summer_2025 (binary classification)

Рис. 9. Сравнение качества моделей в зависимости от объема данных и аугментации: hisory_summer_2025 (бинарная классификация)

Заключение

Эксперименты показали, что качество оценки коротких ответов нелинейно зависит от объёма обучающих данных. В большинстве поставленных экспериментов значение F1-меры медленно снижалось при уменьшении объёма обучающей выборки до 30 % от исходного; и существенно быстрее — при уменьшении до менее чем 30 % от исходного объёма. Наиболее устойчивые и качественные результаты были получены при использовании моделей rubert-tiny2 и MiniLM-L12-v2.

Эксперименты по аугментации обучающей выборки с помощью генеративной модели DeepSeek показали, что синтетическое расширение обучающего корпуса эффективно компенсирует нехватку реальных данных. Эффект от аугментации был (при прочих равных) тем сильнее, чем меньше был объём исходной обучающей выборки, что подчёркивает потенциал генеративных методов при ограниченных ресурсах для создания размеченных корпусов.

Таким образом, результаты работы подтверждают целесообразность использования генеративных методов для дополнения обучающих выборок и показывают, что даже при ограниченном количестве исходных данных возможно достижение высокого качества автоматической оценки коротких ответов.

References

[1] R. Gao, H. E. Merzdorf, S. Anwar, M. C. Hipwell, and A. R. Srinivasa, “Automatic assessment of text-based responses in post-secondary education: A systematic review”, *Computers and Education: Artificial Intelligence*, vol. 6, p. 100 206, 2024. DOI: 10.1016/j.caeai.2024.100206.

[2] N. S. Lagutina and K. V. Lagutina, “A survey of models for automatic assessment of similarity of student’s answer to the reference answer”, *Automatic Control and Computer Sciences*, vol. 59, no. 7, pp. 1152–1169, 2025. DOI: 10.3103/S0146411625700427.

[3] S. Kumar, S. Chakrabarti, and S. Roy, "Earth mover's distance pooling over Siamese LSTMs for automatic short answer grading", in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017, pp. 2046–2052. DOI: 10.24963/ijcai.2017/284.

[4] T. H. Nguyen, H. H. Nguyen, Z. Ahmadi, T.-A. Hoang, and T.-N. Doan, "On the impact of dataset size: A Twitter classification case study", in *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2021, pp. 210–217. DOI: 10.1145/3486622.3493960.

[5] A. R. Nair, R. P. Singh, D. Gupta, and P. Kumar, "Evaluating the impact of text data augmentation on text classification tasks using DistilBERT", *Procedia Computer Science*, vol. 235, pp. 102–111, 2024. DOI: 10.1016/j.procs.2024.04.013.

[6] C. Shorten, T. M. Khoshgoftaar, and B. Furht, "Text data augmentation for deep learning", *Journal of Big Data*, vol. 8, p. 101, 2021. DOI: 10.1186/s40537-021-00492-0.

[7] A. Poulton and S. Eliens, "Explaining transformer-based models for automatic short answer grading", in *Proceedings of the 5th International Conference on Digital Technology in Education*, 2021, pp. 110–116. DOI: 10.1145/3488466.3488479.

[8] G. Kortemeyer, "Performance of the pre-trained large language model GPT-4 on automated short answer grading", *Discover Artificial Intelligence*, vol. 4, no. 1, p. 47, 2024. DOI: 10.1007/s44163-024-00147-y.

[9] C. B. Minnegalieva, I. I. Kashapov, and O. D. Morozova, "Automated grading of students' short answers using language models", *Automatic Documentation and Mathematical Linguistics*, vol. 58, no. Suppl 3, S109–S114, 2024. DOI: 10.3103/S0005105525700177.

[10] H. Tan, C. Wang, Q. Duan, Y. Lu, H. Zhang, and R. Li, "Automatic short answer grading by encoding student responses via a graph convolutional network", *Interactive Learning Environments*, vol. 31, no. 3, pp. 1636–1650, 2023. DOI: 10.1080/10494820.2020.1855207.

[11] J. Lun, J. Zhu, Y. Tang, and M. Yang, "Multiple data augmentation strategies for improving performance on automatic short answer scoring", in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 13 389–13 396. DOI: 10.1609/aaai.v34i09.7062.

[12] M. C. Wijanto and H.-S. Yong, "Combining balancing dataset and SentenceTransformers to improve short answer grading performance", *Applied Sciences*, vol. 14, no. 11, p. 4532, 2024. DOI: 10.3390/app14114532.

[13] S. Zhong, J. Zeng, Y. Yu, and B. Lin, "Clustering algorithms and RAG enhancing semi-supervised text classification with large LLMs", *International Journal of Data Science and Analytics*, vol. 20, pp. 5377–5398, 2025. DOI: 10.1007/s41060-025-00774-3.

[14] R. Li *et al.*, "Retrieval-augmented meta learning for low-resource text classification", in *Proceedings of the International Joint Conference on Neural Networks*, 2024, pp. 1–8. DOI: 10.1109/IJCNN60899.2024.10651119.

[15] Y. Yu, Y. Zhuang, R. Zhang, Y. Meng, J. Shen, and C. Zhang, "ReGen: Zero-shot text classification via training data generation with progressive dense retrieval", in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 11 782–11 805. DOI: 10.18653/v1/2023.findings-acl.748.

[16] H. Dai *et al.*, "AugGPT: Leveraging ChatGPT for text data augmentation", *IEEE Transactions on Big Data*, vol. 11, pp. 907–918, 2025. DOI: 10.1109/TBDATA.2025.3536934.

[17] V. N. Kopnin and N. S. Lagutina, “Analysis of the application of large language models for comparing open student responses with reference answers”, in *Matematicheskoe i Informacionnoe Modelirovanie: Materialy Vserossijskoj Konferencii Molodyh Uchenyh*, in Russian, vol. 23, 2025.

[18] A. Liu *et al.*, *DeepSeek-V3 technical report*, 2025. DOI: 10.48550/arXiv.2412.19437v2. arXiv: 2412.19437 [cs.CL].