DOI: 10.15514/ISPRAS-2026-38(3)-30

Применение языковых моделей для автоматической проверки открытых ответов учащихся

В.Н. Копнин, ORCID: 0009-0007-5451-775X <vlad.kopnen@mail.ru>
К.В. Лагутина, ORCID: 0000-0002-1742-3240 <lagutinakv@mail.ru>
А.Ю. Полетаев, ORCID: 0000-0003-0116-4739 <anatoliy-poletaev@mail.ru>
Н.С. Лагутина, ORCID: 0000-0002-6137-8643 <lagutinans@rambler.ru>

Ярославский государственный университет им. П.Г. Демидова,
Россия,150003, г. Ярославль, ул. Советская, д. 14.

Аннотация. Автоматическая оценка коротких открытых ответов обучающихся упрощает работу преподавателя и позволяет быстро и эффективно оценить работу студента. Целью данной работы является сравнение методов классификации русскоязычных коротких ответов в зависимости от оценки. Анализируется применение нейросетевых языковых моделей и методов машинного обучения. Оценивание происходит на основе эталонного ответа по двум классам: верный/неверный, или трем: верный/частично верный/неверный ответ. Для проведения экспериментов авторы собрали четыре корпуса ответов на вопросы из различных дисциплин и предметных областей: корпус общих вопросов по ИТ-дисциплинам и высшей математике, корпус вопросов по базам данных, корпус вопросов по истории и корпус вопросов по разработке с помощью программного инструмента Qt. В процессе экспериментов с данными текстами сравнивались 11 предобученных языковых моделей, 2 способа обучения, 2 способа разбиения на обучающую и тестовую выборки и 7 классификаторов, чтобы проанализировать различные способы векторного представления и классификации русскоязычных текстов. Анализ результатов бинарной классификации показал, что не существует доминирующей пары «модель + классификатор», которая бы стабильно превосходила остальные на всех корпусах. F-меру более 0.9 показывали BERT-модели в комбинациях с центроидным классификатором, логистической регрессией или многослойным перцептроном. Для тернарной классификации лучшими комбинациями оказались модели rugpt3m, MiniLM-L12 и rubert-tiny2 в сочетании с категориальным бустингом и центроидным классификатором, F-мера составила 0.58. Повысить качество F-меры до 0.96 для бинарной и до 0.91 для тернарной классификации помогла аугментация на основе правил для рекомбинации реальных данных. Анализ ошибок показал, что основную сложность представляет отделение полностью верных ответов от частично верных. На основе результатов экспериментов была разработана и опубликована программная система для проведения контрольных мероприятий среди учащихся.

Ключевые слова: обработка естественного языка; оценка ответов учащихся; классификация текстов; нейросетевые языковые модели; искусственный интеллект в образовании.

Для цитирования: Копнин В.Н., Лагутина К.В., Полетаев А.Ю., Лагутина Н.С. Применение языковых моделей для автоматической проверки открытых ответов учащихся. Труды ИСП РАН, том 38, вып. 3, часть 2, 2026 г., стр. 197–214. DOI: 10.15514/ISPRAS–2026–38(3)–30.

Благодарности: Исследование выполнено при поддержке Российского научного фонда (проект № 25-21-00196).

Applying language models to automatically check students' open-ended answers

V.N. Kopnin, ORCID: 0009-0007-5451-775X <vlad.kopnen@mail.ru>
K.V. Lagutina, ORCID: 0000-0002-1742-3240 <lagutinakv@mail.ru>
A.Y. Poletaev, ORCID: 0000-0003-0116-4739 <anatoliy-poletaev@mail.ru>
N.S. Lagutina, ORCID: 0000-0002-6137-8643 <lagutinans@rambler.ru>

P.G. Demidov Yaroslavl State University,
Russia, 150003, Yaroslavl, Sovetskaya St., 14.

Abstract. Automatic grading of short open-ended student answers simplifies teachers' work and allows for quick and effective assessment. The goal of this study is to compare methods for classifying Russian-language short answers depending on the assessment. The authors analyzed the application of neural network language models and machine learning methods. Evaluation is based on a reference answer. The student' answer is categorized in two classes: correct/incorrect, or three: correct/partially correct/incorrect. For the experiments, the authors collected four corpora of answers to questions from various disciplines and subject areas: a corpus of general questions on IT disciplines and higher mathematics, a corpus of questions on databases, a corpus of questions on history, and a corpus of questions on Qt development. During the experiments with these texts, 11 pre-trained language models, 2 training methods, 2 methods of splitting training and test sets, and 7 classifiers were compared to analyze various methods of vector representation and classification of Russian-language texts. An analysis of binary classification results revealed that there is no dominant model + classifier pair that consistently outperforms others across all corpora. BERT models in combination with a centroid classifier, logistic regression, or multilayer perceptron demonstrated the F-measure greater than 0.9. For ternary classification, the best combinations were rugpt3m, MiniLM-L12, and rubert-tiny2 models in combination with categorical boosting and a centroid classifier, with the F-measure of 0.58. Augmentation based on rules for recombination of real data helped to improve the F-measure to 0.96 for binary classification and to 0.91 for ternary classification. Error analysis revealed that the main difficulty is separating completely correct answers from partially correct ones. Based on the experimental results, a software system for conducting assessments among students was developed and published.

Keywords: natural language processing; assessing students' answers; text classification; neural network language models; artificial intelligence in education.

For citation: Kopnin V.N., Lagutina K.V., Poletaev A.Y., Lagutina N.S. Applying language models to automatically check students' open-ended answers. Trudy ISP RAN/Proc. ISP RAS, vol. 38, issue 3, part 2, 2026, pp. 197-214 (in Russian). DOI: 10.15514/ISPRAS-2026-38(3)-30.

Acknowledgements. This work was supported by a grant from the Russian Science Foundation, project no. 25-21-00196.

1. Введение

Одним из основных компонентов современного обучения является оценка знаний. Она не только определяет уровень понимания учебного материала и владения практическими навыками, но и развивает мышление учащихся, навыки решения проблем, планирования и самооценки [1]. Качественная оценка выполненных учащимися заданий является крайне трудоёмкой задачей, из-за чего для проведения контрольных работ часто используются закрытые тестовые вопросы с несколькими вариантами ответов без обратной связи. Чтобы обучение и оценивание были масштабируемыми и персонализированными без потери качества, необходимо развивать системы, автоматизирующие отдельные аспекты работы преподавателя [2]. Создание таких систем стало возможно в рамках применения искусственного интеллекта в образовании, в том числе в связи с совершенствованием больших языковых моделей для автоматической обработки текста [3]. Базовые алгоритмы решения поставленной задачи лежат в областях автоматической оценки коротких ответов (automatic short answer grading, ASAG) и определения семантического сходства текстов.

Задача оценки открытых ответов часто формулируется как задача классификации текстов на естественном языке по оценкам, при этом тексты моделируются как вектора чисел [4]. Первые методы такого моделирования использовали простейшие характеристики текста: частоту символов и слов, среднюю длину предложения и т. п. В последние годы самыми часто используемыми моделями стали эмбеддинги (англ. embeddings, вектора чисел, полученные на основе языковой модели), построенные с помощью методов глубокого обучения и языковых моделей, в том числе больших языковых моделей LLaMA и GPT [5].

Сравнение сходства ответов учащихся с эталонными лежит в основе многих исследований в области ASAG [6]. Качество решения этой задачи сильно различается в разных работах. Например, использование RoBerta показало F-меру в диапазоне 0.80-0.9 для разных корпусов [7]. А комбинация BERT-эмбеддингов с классификатором RandomForest позволила достичь F-меры всего 0.57 [8]. Следует отметить, что подавляющее большинство исследований выполнено для английского языка, в то время как русскоязычные тексты остаются малоизученными [9]. Поэтому применение и сравнение методов автоматической оценки открытых ответов на новых корпусах текстов является актуальной задачей.

Цель данной работы – сравнение методов для автоматической оценки русскоязычных коротких открытых ответов учащихся. Тексты моделируются как эмбеддинги при помощи современных языковых моделей. Методы включают в себя как стандартные алгоритмы машинного обучения для бинарной и мультиклассификации, так и классификационные алгоритмы на основе сравнения с эталоном. Каждому вопросу соответствует один эталонный ответ. Баллы за ответ выставляются по шкале от -2 до 1:

- 1 – ответ полностью верный;
- 0 – ответ частично верный;
- -1 – ответ неверный, но по теме;
- -2 – ответ не по теме.

Классификационная задача ставится в двух вариантах:

1. Бинарная классификация на верный/неверный ответ, где баллы 0 и 1 обозначают общий класс верного ответа, а баллы -2 и -1 – неверного ответа. При объединении баллов 0 и 1 в общий класс моделируется такой подход к оценке ответов, когда студенту достаточно продемонстрировать минимальный уровень знаний.
2. Тернарная классификация на верный/частично верный/неверный ответ, где баллы -2 и -1 обозначают общий класс неверного ответа. Ответы не по теме в рассматриваемых корпусах встречались редко, поэтому не были выделены в отдельный класс.

Таким образом, два варианта классификационной задачи позволяют проанализировать автоматизацию как строгих, так и нестрогих подходов к оценке открытых ответов.

Для оценки устойчивости и обобщающей способности методов, помимо стандартной кросс-валидации, использовалась схема LOCO (Leave One Corpus Out), моделирующая проверку ответов на новые, незнакомые системе вопросы.

2. Аналогичные работы

Задача автоматической оценки коротких ответов обучающихся наиболее часто и успешно решается при помощи предобученных языковых моделей [9]. Одни из лучших результатов демонстрируют языковые модели, являющиеся вариациями BERT. Авторы исследования [10] классифицировали с их помощью тексты из корпуса SemEval-2013 на правильные, неправильные и противоречивые и получили значение F-меры от 0.67 до 0.79 для различных вопросов.

В обзоре [11] сравнение моделей на основе эмбеддингов для корпусов текстов Mohler и SemEval-2013 показывает, что BERT-модели превосходят остальные подходы к автоматической оценке коротких ответов, основанные на нейросетях, однако в целом F-мера не превосходит 0.83.

Высокие значения F-меры могут быть достигнуты комбинацией эмбеддингов и других векторных моделей. Например, авторы работы [12] объединили n-граммы и эмбеддинги FastText для решения задачи, аналогичной оценке коротких ответов на основе эталона: определение близости текстов. На собственном корпусе метод достиг F-меры около 0.93.

В последние годы методы и подходы на основе языковых моделей развиваются в сторону использования генеративного искусственного интеллекта. В фреймворке GradeHITL [13] скомбинированы большие языковые модели и дообучение с помощью SBERT и RoBERTa, что привело к доле правильных ответов 0.72-0.91 для разных вопросов. GPT 3.5 [14] без дообучения достиг точности в оценке эссе 0.29-0.35, сильно разойдясь с эталонными оценками. GPT-4 [15] показал более высокую точность 0.716. Однако исследователи отмечают, что GPT может часто выставлять разные оценки одним и тем же ответам при повторном запросе [14, 15].

В статье [5] применялись две большие языковые модели: общедоступная LLaMA 3.2 и премиальная GPT-4o, а также анализировалась согласованность как между человеческими и автоматическими оценками, так и оценками от нескольких преподавателей. Авторы собрали собственный корпус текстов на португальском языке и автоматически оценивали тексты по набору критериев, где шкала варьировалась от нуля до нескольких десятков баллов в зависимости от вопроса. Тест Манна-Уитни подтвердил близость оценок обеих моделей к оценкам, выставляемым людьми. Коэффициент корреляции Спирмена составил примерно 0.96 как между показателями LLM и преподавателей, так и среди самих преподавателей.

В исследовании [16] проведён сравнительный анализ двух подходов к автоматической оценке коротких ответов: использование генеративных больших языковых моделей без предварительного обучения и использование классических языковых моделей с обучением на тренировочной выборке. Авторы экспериментировали на открытых наборах данных: Mohler с англоязычными текстами и PT_ASAG с португалоязычными текстами. Подход с генеративным искусственным интеллектом с применением модели GPT-4o показал значения RMSE в диапазоне 1.48-1.75. Второй подход с экспериментами на традиционных языковых моделях Glove, BERT, FastText обеспечил значительно лучшие значения RMSE: 0.67-1.02. Несмотря на превосходство традиционных подходов по качеству, авторы подчеркнули, что большие языковые модели не нуждаются в дополнительном обучении, что упрощает работу с ними.

Авторы исследования [17] также сопоставляли генеративный искусственный интеллект, а именно модель GPT-4, и традиционные эмбеддинги, включая BERT и RoBERTa-large. Сравнение выполнялось на англоязычных корпусах SciEntsBank и Beetle как для бинарной классификации ответов по оценкам, так и для тернарной. Модель GPT-4 демонстрирует F-меру около 0.74 для обеих видов классификации на корпусе SciEntsBank и менее стабильный результат 0.52-0.61 на корпусе Beetle. Предварительно обученные эмбеддинги достигают F-меры порядка 0.73-0.81 на SciEntsBank и 0.73-0.91 на Beetle. Бинарная классификация выполняется ожидаемо лучше тернарной. Предобученные языковые модели превосходят методы генеративного искусственного интеллекта. Кроме того, авторы оценивали тексты с помощью GPT-4 как с применением эталона, так и без него, но не получили статистически значимого отличия в качестве.

Для экспериментов с национальными языками авторы создают собственные корпуса. Например, для финского языка был собран корпус из 2000 ответов на 100 вопросов, и GPT-4 достаточно успешно справился с их оценкой, показав QWK 0.6-0.8 [18]. Тексты на польском языке оценивались нейросетевыми алгоритмами и алгоритмом, измеряющим количество одинаковых слов или синонимов [19], F-мера составила 0.98. Для русского языка подход с учетом синонимов, антонимов и прочих лингвистических особенностей текста продемонстрировал F-меру 0.70 [20].

В работе [21] представлен метод автоматической оценки ответов на русском языке, основанный на вычислении косинусного сходства BERT-эмбеддингов студента и эталонного ответа и на

проверке ключевых слов. Экспертная оценка показала высокое качество метода: максимальная доля правильных ответов составляет 0.90, средняя – 0.77.

Подходы, показавшие лучшие значения F-меры, сгруппированы в табл. 1. Среди всех метрик была выбрана именно F-мера, так как авторы используют её как основную метрику качества (см. разделы 4 и 6). Лучшие результаты составляют >0.79 для нейросетевых подходов.

Обзор работ показывает перспективность использования языковых моделей для автоматической оценки ответов учащихся, но ставит вопрос о способах их применения. Другим недостатком исследований является то, что они часто ограничиваются одним корпусом текстов [22]. Таким образом, системное исследование моделей русского языка и сбор корпусов ответов является актуальной задачей.

Табл. 1. Подходы с лучшей F-мерой.
Table 1. Approaches with the best F-measure.

Статья	Язык	Подход	F-мера
Camus и Filighera [10]	Английский	BERT	0.79
Ahmed и др. [11]	Английский	BERT	0.83
Shashavali и др. [12]	Английский	n-граммы и FastText	0.93
Kortemeyer [17]	Английский	GPT-4	0.91
Bani Saad [19]	Польский	NN и синонимы	0.98
Леонов и др. [20]	Русский	Лингвистические характеристики	0.70

3. Метод оценки ответов учащихся

3.1 Корпуса ответов учащихся

Для проверки эффективности и универсальности изучаемых методов автоматической оценки требовались разнообразные данные. Были собраны четыре оригинальных корпуса ответов студентов бакалавриата и магистратуры на русском языке. Ответы студентов оценивались преподавателями вручную. Для корпусов databases-exam-2025, history-2025-summer и qt-questions оценка производилась одним преподавателем, который вёл соответствующий курс. Ответы из general-questions-masters-2025 оценивались тремя преподавателями, в качестве итоговой оценки выбиралась наиболее часто встречающаяся, если оценки оценки преподавателей оказывались разными, для итоговой оценки привлекался четвёртый эксперт.

Вопросы в корпусах охватывают различные предметные области и типы формулировок. Это позволяет более объективно оценить, насколько хорошо методы справляются с задачами разной сложности: от воспроизведения точных фактов и определений до аргументации и развернутых рассуждений. Каждый корпус содержит несколько вопросов, эталонный ответ на каждый вопрос и ответы студентов. Все корпуса опубликованы: [23].

3.1.1 Корпус general-questions-masters-2025

Корпус general-questions-masters-2025 служит базовым полигоном для тестирования методов, объединяя вопросы из разных областей компьютерных наук, математики и общих знаний. Примеры вопросов:

1. Что такое модель в архитектуре модель-вид-контроллер?
2. Что такое производная функции?
3. Что такое файл в области канцелярских принадлежностей?
4. Что такое файл в операционной системе компьютера?

Вопросы намеренно подобраны так, чтобы проверить базовые знания, полученные в разных курсах. Ответы ограничены тремя предложениями, что подталкивает студентов давать сжатые, содержательные формулировки.

Как видно из табл. 2, корпус демонстрирует ярко выраженный дисбаланс в сторону правильных ответов (класс 1). Это объясняется тем, что вопросы касаются фундаментальных понятий. При этом наличие значительного числа частично верных ответов (класс 0) указывает на то, что студенты часто вспоминают материал неточно или неполно, что является типичной ситуацией в образовательном процессе.

Табл. 2. Количество вопросов и ответов в корпусах.
Table 2. Number of questions and answers in corpora.

Корпус	Всего вопросов	Оценка				Всего ответов
		-2	-1	0	1	
general-questions-masters-2025	8	2	43	118	384	547
databases-exam-2025	3	0	18	31	24	73
history-2025-summer	12	0	43	120	359	522
qt-questions	12	2	44	330	555	931

3.1.2 Корпус databases-exam-2025

Корпус databases-exam-2025 представляет собой пример специализированного экзаменационного тестирования и состоит из ответов студентов на вопросы, сфокусированные на практических аспектах курса о базах данных:

1. В чём преимущества SQL в качестве языка для обработки данных?
2. С помощью каких средств можно поддерживать целостность информации в базе данных?
3. В каких случаях использование ORM принесёт пользу, а в каких – нет?

Сбор данных проводился в условиях реального экзамена. Это объясняет две ключевые особенности корпуса (таблица 2):

1. Отсутствие ответов «не по теме» (класс -2), так как студенты в экзаменационной ситуации стараются дать любой релевантный ответ.
2. Сильный дисбаланс классов, потому что студенты целенаправленно готовились к экзамену, что привело к малому количеству полностью неверных ответов (-1) и преобладанию частично верных (0) и правильных (1) ответов.

Небольшой объём корпуса делает его сложной, но интересной задачей для классификации.

3.1.3 Корпус history-2025-summer

Корпус history-2025-summer репрезентирует гуманитарную дисциплину и требует от студентов работы с историческими концепциями, датами и причинно-следственными связями. Вопросы посвящены истории России конца XIX – XX века. Примеры вопросов:

1. Как бы Вы обосновали необходимость СССР подписание Пакта о ненападении с Германией в 1939 г.?
2. Приведите НЕ МЕНЕЕ 3 причин, обусловивших распад СССР.
3. ДВУМЯ ключевыми идейными направлениями в общественной мысли России в 1830-1840-е гг. были...

Форматы вопросов разнообразны: от простого перечисления пунктов до развернутых объяснений и аргументации. Они подразумевают как точное знание фактов, так и

сформированное историческое мышление, что делает оценку особенно сложной для автоматизации. Как и в предыдущих корпусах, наблюдается дисбаланс в сторону правильных ответов (табл. 2), что характерно для экзамена, где студенты стараются продемонстрировать наилучшие результаты.

3.1.4 Корпус qt-questions

Корпус qt-questions отражает практику текущего контроля знаний в рамках технической дисциплины о разработке приложений на Qt и отражает работу студентов с материалом, изученным на предыдущих занятиях. Корпус состоит из ответов студентов второго курса. Форматы вопросов варьируются от простого перечисления до объяснения концепций. Примеры вопросов:

1. Что такое делегат в архитектуре модель-вид-делегат?
2. Что можно узнать о файле с помощью QFileInfo? Перечислите хотя бы 3 свойства.

Вопросы задавались в ходе регулярных занятий и были сфокусированы на материале, пройденном на предыдущих лекциях. Это создает условия, в которых студенты опираются на свежие, но еще не закрепленные долгосрочной практикой знания.

Как видно из таблицы 2, это самый большой по объёму корпус (931 ответ). В нем сохраняется общая тенденция к дисбалансу в сторону правильных ответов, однако для некоторых вопросов наблюдается значительное количество частично верных ответов, что может указывать на сложность усвоения данной концепции студентами.

Этот корпус ценен тем, что моделирует ситуацию «промежуточного» контроля, где ответы могут быть менее отрепетированными, чем на экзамене, и чаще содержат частично верные или неточные формулировки. Это позволяет исследовать работу моделей в условиях, более близких к повседневному учебному процессу.

Собранные корпуса образуют репрезентативную базу для исследования. Их ключевые общие черты – разнообразие предметных областей, форматов вопросов и выраженный дисбаланс классов в сторону правильных ответов. Последнее является не недостатком данных, а отражением реальной образовательной практики и формирует основную методологическую трудность, которую необходимо преодолевать при построении моделей автоматической оценки.

3.2 Метод классификации

Ответы на вопросы классифицировались по оценкам с помощью нейросетевых языковых моделей и методов машинного обучения. Было проведено сравнение 11 предобученных языковых моделей, 7 классификаторов, 2 способа разбиения на обучающую и тестовую выборки и 2 способов обучения, чтобы охватить различные способы векторного представления и классификации текстов на русском языке.

Для всестороннего сравнения различных подходов к автоматической оценке использовалась унифицированная методология экспериментов. Её цель – оценить универсальность, стабильность и способность к обобщению различных комбинаций моделей и классификаторов, а также выявить лучшие по качеству результаты.

Каждый эксперимент следовал единой схеме:

1. Разбиение корпуса на обучающую и тестовую выборки одним из трёх способов (см. 3.2.2). Корпус – это набор вопросов, каждому вопросу соответствует один эталонный ответ и несколько ответов обучающихся. У каждого ответа обучающегося есть оценка: 0 и 1 для бинарной классификации, -1, 0 и 1 для тернарной.
2. Векторизация текстов: преобразование текста ответа учащегося и эталонного ответа в векторные представления (эмбеддинги) с помощью одной из предобученных языковых моделей (см. 3.2.3). Каждый текст эталонного ответа и ответа обучающегося независимо от остальных преобразуется в вектор чисел.
3. Обучение классификаторов на основе эмбеддингов и оценки эксперта с использованием выбранного алгоритма (см. 3.2.4). В зависимости от входных параметров классификатора эмбеддинг ответа обучающегося либо конкатенировался с эмбеддингом эталонного ответа в общий вектор, либо эти два эмбеддинга подавались парой отдельных векторов.
4. Валидация: проверка качества предсказаний на тестовой выборке с расчетом стандартных метрик (F-мера, точность, полнота, AUC-ROC, для мультиклассовой классификации использовались макро-метрики).

3.2.1 Способы обучения

При обучении моделей применялось два основных подхода: обучение на каждом вопросе корпуса по отдельности и обучение на объединенном наборе данных всех вопросов корпуса. Данный подход позволил оценить, насколько методы автоматической оценки способны к обобщению.

Первый подход, при котором каждый вопрос рассматривался как самостоятельный мини-корпус, предполагает, что модель обучается и тестируется только на ответах, относящихся к одному конкретному вопросу. Это моделирует ситуацию узкоспециализированной системы, настроенной на проверку ответов по строго определенной теме. Такой способ позволяет измерить «идеальную» производительность модели в условиях, когда контекст и терминология максимально однородны.

Второй подход, при котором все вопросы в рамках одного корпуса объединялись в единый набор данных, направлен на проверку универсальности и способности модели к переносу знаний. Модель, обученная таким образом, должна выявлять общие паттерны «правильности» и «неправильности» ответов, абстрагируясь от конкретной предметной области отдельного вопроса. Это моделирует создание более гибкой системы, способной работать с разнообразными темами в рамках одной дисциплины или типа заданий без необходимости перенастройки.

Сравнение результатов, полученных этими двумя способами, позволяет сделать выводы о том, насколько успешно те или иные комбинации языковых моделей и классификаторов справляются с задачей обобщения, и в каких практических сценариях (узконаправленная или универсальная проверка) их применение наиболее целесообразно.

3.2.2 Способы разбиения выборки

Чтобы оценить модель в разных сценариях, применялись две стратегии валидации, каждая из которых отвечает на свой исследовательский вопрос:

- Стратифицированная 5-кратная кросс-валидация, стандартный способ сравнения производительности моделей;
- LOCO (Leave One Corpus Out) – один вопрос корпуса тестовый, остальные обучающие; позволяет посмотреть, как себя ведет модель на абсолютно новых данных, оценить способность к обобщению и переносимость.

Дополнительно для стратифицированной кросс-валидации применялся подход, когда эксперимент проводился отдельно для каждого вопроса корпуса. Это позволяет выявить, насколько качество модели зависит от конкретной формулировки и темы задания.

Метод LOCO позволяет оценить, насколько модель, обученная на одних вопросах, способна корректно классифицировать ответы на другой, исключённый из обучения вопрос, что моделирует сценарий реального использования системы для проверки новых заданий.

3.2.3 Используемые модели эмбеддингов

Было протестировано 11 предобученных языковых моделей, чтобы охватить различные методы и подходы к пониманию текста на русском языке. Все модели были взяты с платформы HuggingFace.

Энкодерные модели: ai-forever/ruBert-large, bert-base-multilingual-cased, bert-base-multilingual-uncased, cointegrated/rubert-tiny2, s-nlp/ruRoberta-large-paraphrase-v1, sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2. Преимущества моделей данного типа: контекстуализированные эмбеддинги для каждого слова/токена, идеально подходят для задач понимания текста. Были выбраны как модели среднего размера, так и облегченные модели.

Энкодер-декодерные модели: cointegrated/rut5-base, google/byt5-small, google/mt5-small. Преимущества: универсальные архитектуры, предобученные на задачах, близких к перефразированию и генерации, что может быть полезно для анализа семантического сходства ответов.

Декодерные модели: ai-forever/rugpt3medium_based_on_gpt2, alenusch/rugpt3-paraphraser. Преимущество: сильны в генерации, их эмбеддинги потенциально могут содержать информацию о правдоподобности продолжения текста.

3.2.4 Используемые классификаторы

Для классификации ответов применялся набор из 7 классификаторов, представляющих разные семейства алгоритмов: логистическая регрессия, случайный лес, градиентный бустинг (CatBoost, ансамблевый метод, последовательно обучающий деревья решений), метод опорных векторов (SVM), многослойный перцептрон, решающее дерево, центроидный классификатор.

Центроидный классификатор основан на идее интерпретации метрики близости как принадлежности ответа к классу с помощью центроидных значений. Авторы применяли его следующим образом. При обучении для каждого класса (-2, -1, 0, 1) вычислялся центроид – среднее значение косинусного сходства между всеми парами «ответ-эталон» этого класса. При прогнозировании для новой пары вычислялось косинусное сходство. Этой паре присваивался тот класс, центроид которого находился ближе всего к вычисленному значению. Например, если центроиды классов (-1, 0, 1) были равны (-0.1; 0.6; 0.9), а сходство новой пары было равно 0.8, ответ был отнесен к классу 1.

Полученные эмбеддинги ответа и эталона подавались на вход классификатора либо как конкатенированный вектор, либо как пара векторов, если классификатор поддерживал такой формат (например, центроидный классификатор и CatBoost).

3.3 Метод аугментации данных

Для преодоления дисбаланса классов и увеличения объёма обучающих данных был разработан метод аугментации, основанный на предметно-ориентированных правилах рекомбинации существующих ответов. Данный подход позволяет целенаправленно генерировать синтетические примеры для всех классов, моделируя наиболее разнообразные сочетания вариантов ответов.

Правила аугментации формулируются на уровне пар «ответ учащегося – другой ответ учащегося» и включают следующие этапы:

1. Генерация примеров класса «полностью верный ответ» (метка 1): создаются все возможные пары между разными текстами ответов, которые были оценены экспертом как полностью верные (исходная метка 1). Это позволяет модели обучаться на разнообразии корректных формулировок.
2. Генерация примеров класса «частично верный ответ» (метка 0): создаются пары, где первый текст – частично верный ответ (исходная метка 0), а второй – полностью верный ответ (метка 1).
3. Генерация примеров класса «неверный, но по теме ответ» (метка -1): формируются пары между неверными ответами (исходная метка -1) и, соответственно, верными (метка 1) или частично верными (метка 0) ответами. Таким образом генерируются релевантные ответы на текущий вопрос, но содержащие ошибочные утверждения относительно верного ответа.
4. Генерация примеров класса «ответ не по теме» (метка -2): для моделирования семантически нерелевантных ответов создаются пары, где первый текст взят из текущего корпуса, а второй – из ответов на любой другой вопрос (или из другого корпуса). Данный способ позволяет сгенерировать ответы, не имеющие семантически ничего общего с ожидаемым ответом на текущий вопрос.

В экспериментах аугментация применялась только к обучающей выборке, чтобы избежать «утечки» информации из тестовых данных. Данный метод позволил увеличить объём обучающих данных на несколько порядков (например, более чем в 10 000 раз для корпуса history-2025-summer, подробнее в разделе 4.3), что существенно улучшило качество обучения моделей, особенно в сценарии тернарной классификации.

Важно отметить, что сгенерированные аугментированные данные с четырьмя исходными метками (-2, -1, 0, 1) впоследствии преобразовывались в целевые классы, соответствующие конкретному эксперименту (бинарному или тернарному), как описано ниже в разделах 4.1 и 4.2.

4. Эксперименты с автоматической оценкой ответов

4.1 Эксперименты с бинарной классификацией

Для задачи бинарной классификации ответы учащихся были разделены на два класса: класс 0 (неправильный ответ) объединил исходные классы -2 и -1; класс 1 (правильный ответ) объединил исходные классы 0 и 1. Описанное преобразование имеет практический смысл: класс 0 соответствует оценке «незачет», а класс 1 – «зачет», когда даже частично верный ответ считается удовлетворительным.

Было проведено более 950 экспериментов с различными комбинациями моделей, классификаторов и методов разбиения данных. В табл. 3 представлены 5 конфигураций экспериментов (по 1-2 на каждый корпус), на которых достигалось наилучшее качество (максимальная F-мера или AUC).

В столбцах таблицы 3-6 представлены: пара «модель + классификатор», F-мера и ее стандартное отклонение, точность (P), полнота (R), метрика AUC-ROC и ее стандартное отклонение, способ разбиения данных на обучающую и тестовую выборки (сплиттер) и корпус, на котором были получены данные результаты. Лучшие результаты по каждому показателю метрик выделены полужирным шрифтом. Помимо результатов 5-кратной кросс-валидации, в таблице представлены данные экспериментов по схеме LOCO (Leave One Corpus Out). В этой схеме каждый вопрос поочерёдно исключается из обучающей выборки и используется для тестирования, что позволяет оценить устойчивость моделей к новым, незнакомым формулировкам в рамках одной предметной области.

Анализ всех проведенных экспериментов, включая результаты LOCO, подтвердил, что не существует доминирующей пары «модель + классификатор», которая бы стабильно превосходила остальные на всех корпусах и для всех способов валидации. Однако в сценарии LOCO удается достичь высокой полноты (1.00 с точностью до округления) и F-меры (до 0.97). Это говорит о том, что некоторые комбинации сохраняют высокую чувствительность к правильным ответам даже при обучении на данных с других вопросов, хотя их общая стабильность (низкий AUC и высокое стандартное отклонение) остаётся проблемой.

Табл. 3. Наиболее показательные эксперименты с бинарной классификацией.
Table 3. The most revealing experiments with binary classification.

Модель + классификатор	F (std_F)	P	R	AUC (std_{auc})	Сплиттер (корпус)
rubert-tiny2 + центр. класс.	0.92 (0.02)	**0.97**	0.87	0.87 (0.11)	5-Fold (hist-exm, вопросы вместе)
byt5-small + лог. рег.	0.87 (0.02)	0.76	**1.00**	0.87 (0.12)	5-Fold (db-exam, вопросы вместе)
bert-cased + лог. рег.	0.91 (0.12)	0.91	0.93	**0.95** (0.14)	5-Fold (gen-qsts, воп. раздельно)
byt5-small + лог. рег.	**0.97 (0.00)**	0.95	**1.00**	0.56 (0.16)	5-Fold (qt-qsts, вопросы вместе)
ruBert-large + многосл. перц.	0.91 (0.12)	0.91	0.92	0.94 (0.16)	5-Fold (gen-qsts, воп. раздельно)
MiniLM-L12 + центр. класс.	0.91 (0.05)	0.88	0.94	0.96 (**0.03**)	LOCO (gen-qsts, вопросы вместе)
mt5-small + кат. буст.	**0.97** (0.04)	0.95	**1.00**	0.78 (0.15)	LOCO (hist-exm, вопросы вместе)
bert-uncased + многосл. перц.	0.83 (0.17)	0.77	0.95	0.82 (0.16)	LOCO (db-exam, вопросы вместе)
bert-uncased + кат. буст.	**0.97** (0.04)	0.94	**1.00**	0.82 (0.19)	LOCO (qt-qsts, вопросы вместе)

Большинство моделей демонстрируют высокие значения F-меры: более 0.80, что соответствует лучшим результатам в предметной области (таблица 1). Это подтверждает принципиальную возможность автоматической бинарной оценки. Однако низкая стабильность метрики AUC (стандартное отклонение >10%) указывает на чувствительность моделей к конкретной выборке и их недостаточную обобщающую способность.

Некоторые конфигурации (например, byt5-small на корпусе qt-questions) показывают исключительно высокие F-меру и полноту, но низкий AUC. Это говорит о том, что модель практически всегда корректно находит правильные ответы (полнота равна 1.00), но её способность к классификации на всех порогах (AUC) ограничена.

Основными причинами выявленной нестабильности, вероятно, являются дисбаланс классов в исходных корпусах (правильных ответов значительно больше) и их относительно малый объём для обучения.

4.2 Эксперименты с тернарной классификацией

Для задачи тернарной классификации ответы учащихся были распределены по трем классам, отражающим степень правильности: класс -1 (неверный ответ) объединил исходные классы -2 и -1; классы 0 (частично верный ответ) и 1 (полностью верный ответ) сохранены без изменений.

Данное разбиение позволяет сохранить более тонкую градацию оценки по сравнению с бинарным. Было проведено более 960 экспериментов. В табл. 4 представлены наиболее репрезентативные результаты.

Переход к тернарной классификации, как и ожидалось, привел к снижению абсолютных значений всех метрик качества по сравнению с бинарным случаем. Наивысшие значения F-меры не превышают 0.58, что указывает на возросшую сложность задачи различения трех, часто семантически близких, классов.

Табл. 4. Наиболее показательные эксперименты с тернарной классификацией.
Table 4. The most revealing experiments with ternary classification.

Модель + классификатор	F (std_F)	P	R	AUC (std_{auc})	Сплиттер (корпус)
rugpt3m + кат. буст.	0.50 (0.12)	**0.64**	0.54	**0.84** (0.04)	5-Fold (db-exam, вопросы вместе)
mt5-small + кат. буст.	**0.58** (0.14)	**0.64**	**0.61**	0.78 (0.08)	5-Fold (db-exam, вопросы вместе)
rugpt3m + центр. класс.	0.51 (0.20)	0.53	0.56	0.74 (0.14)	5-Fold (hist-exm, вопросы вместе)
rubert-tiny2 + центр. класс.	0.42 (**0.06**)	0.45	0.49	0.66 (**0.03**)	5-Fold (gen-qsts, вопросы вместе)
rugpt3m + центр. класс.	0.48 (0.24)	0.49	0.55	0.65 (0.14)	5-Fold (qt-qsts, вопросы вместе)
MiniLM-L12 + центр. класс.	0.42 (0.08)	0.45	0.48	0.64 (0.10)	LOCO (gen-qsts, вопросы вместе)
MiniLM-L12 + кат. буст.	0.54 (0.19)	0.63	0.56	0.71 (0.17)	LOCO (hist-exm, вопросы вместе)
rugpt3m + центр. класс.	0.49 (0.07)	0.55	0.54	0.72 (0.08)	LOCO (db-exam, вопросы вместе)
bert-cased + кат. буст.	0.39 (0.22)	0.41	0.46	0.58 (0.04)	LOCO (qt-qsts, вопросы вместе)

Сильный дисбаланс классов в корпусах (доминирование правильных ответов) и их относительно малый объём для столь тонкой задачи могут быть основными факторами, ограничивающими качество. Моделям может не хватать данных для надежного обучения паттернам, характерным для миноритарных классов (-1 и 0).

Несмотря на низкие абсолютные показатели, комплексный анализ всех экспериментов, включая LOCO, позволил выявить наиболее устойчивые пары. Модели rugpt3m и MiniLM-L12 в сочетании с категориальным бустингом или центроидным классификатором демонстрируют относительно предсказуемую работу не только при кросс-валидации, но и в более сложном сценарии LOCO, показывая одни из лучших и стабильных результатов на каждом корпусе в отдельности (стандартное отклонение F-меры от 0.07 до 0.19). Именно идентификация этих стабильных, хоть и не самых эффективных в абсолютных цифрах, комбинаций «модель + классификатор», в том числе и в условиях проверки на обобщение (LOCO), стала ключевым фактором для их последующего выбора в экспериментах с аугментацией данных, направленных на преодоление проблемы дисбаланса и недостатка обучающих данных (см. раздел 4.4).

Таким образом, задача тернарной классификации является существенно более сложной по сравнению с бинарной в условиях заданных корпусов. Хотя текущие результаты не позволяют говорить о высокой готовности модели к практическому применению для тонкой оценки, был идентифицирован ряд стабильных методов. Это создает основу для дальнейшей работы, в первую очередь, через применение методов аугментации данных и работы с дисбалансом для улучшения качества распознавания частично верных и неверных ответов.

4.3 Эксперименты с аугментацией корпусов

Низкие результаты мультиклассовой классификации позволили выдвинуть гипотезу о недостаточном объёме данных для обучения моделей. Для проверки этой гипотезы был разработан и реализован специализированный аугментатор, описанный в разделе 3.3.

Данный аугментатор позволил увеличить объём обучающей выборки более чем в 10 000 раз для корпуса history-2025-summer и более чем в 9 000 раз для корпуса general-questions-masters-2025.

Для экспериментов были выбраны две наиболее качественные модели и два лучших классификатора. Полученные результаты демонстрируют значительное улучшение метрик (табл. 5).

Табл. 5. Результаты бинарной и тернарной классификаций с аугментацией данных.
Table 5. Results of binary and ternary classifications with data augmentation.

Модель + классификатор	F (std_F)	P	R	AUC (std_{auc})	Сплиттер (корпус)
Бинарная классификация					
rubert-tiny2 + кат. буст.	0.98 (0.04)	**0.98**	0.98	0.93 (0.18)	5-Fold (hist-exm, воп. отдельно)
MiniLM-L12 + кат. буст.	0.96 (0.05)	**0.96**	**0.97**	0.95 (0.10)	5-Fold (gen-qsts, воп. отдельно)
Тернарная классификация					
rubert-tiny2 + центр. класс.	0.91 (0.11)	**0.96**	**0.88**	0.86 (0.22)	5-Fold (hist-exm, воп. отдельно)

Наиболее показательным является эксперимент с тернарной классификацией на корпусе general-questions-masters-2025. Обучение проводилось на каждом вопросе отдельно с использованием 5-кратной кросс-валидации, при этом аугментация применялась только к обучающей выборке. Это позволило увеличить объём обучающих данных в несколько раз.

Применение аугментации позволило достичь значений F-меры, близких к 0.98, что свидетельствует о практически пригодном для реального использования качестве в сценарии «зачет/незачет». При этом сохраняется высокая стабильность F-меры, точности и полноты.

Полученный результат для тернарной классификации (F-мера 0.91) представляет собой значительный прорыв по сравнению с предыдущими экспериментами без аугментации, где лучший результат не превышал 0.58. Это доказывает, что проблема была не в принципиальной нерешаемости задачи, а в недостатке данных.

Разработанный подход к аугментации, основанный на логических правилах предметной области, показал свою высокую эффективность. Он позволяет целенаправленно генерировать синтетические примеры для редких и сложных классов (-2, -1, 0), тем самым решая ключевую проблему дисбаланса.

Несмотря на значительное улучшение F-меры, точности и полноты, стабильность метрики AUC осталась низкой (стандартное отклонение 0.10-0.22), что указывает на необходимость дальнейшей работы по улучшению обобщающей способности модели и, возможно, по донастройке процесса аугментации и применению иных способов (например, генеративные нейронные сети).

Таким образом, эксперименты с аугментацией данных убедительно доказали, что объём и баланс размеченных данных являются критическим фактором для задач автоматической оценки. Применение специализированного аугментатора позволило радикально улучшить качество моделей, особенно в мультиклассовых сценариях, и открывает путь к созданию практических систем автоматизированной оценки образовательных результатов. Однако, для достижения высокой стабильности, особенно в мультиклассовой классификации, требуются дальнейшие исследования.

5. Система автоматической оценки ответов студентов

На основе результатов проведенных экспериментов была разработана программная система для автоматической оценки открытых ответов в рамках образовательного проекта «Абрис». Данный проект представляет собой веб-платформу, предназначенную для поддержки учебного процесса, и включает модуль для проведения контрольных мероприятий.

Система обеспечивает работу с академическими группами и потоками студентов, позволяя преподавателям создавать и назначать тестирования. Для типов заданий «открытый ответ» и «заполнение пропусков» реализована функция автоматической проверки на основе семантического сходства с эталонным ответом.

Ядро модуля автоматической проверки реализует лучшую из апробированных в ходе исследования конфигураций: для векторного представления текстов используется модель sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2, а для классификации – центроидный классификатор. Данный выбор обусловлен стабильно высокими показателями качества, продемонстрированными этой парой в экспериментах как для бинарной, так и для тернарной классификации, а также ее эффективностью.

В настоящий момент проект находится на финальной стадии разработки и проходит этап бета-тестирования. Параллельно ведутся работы по дальнейшему повышению качества автоматической проверки. Для ознакомления с функциональностью системы доступно демонстрационное веб-приложение [24].

6. Обсуждение результатов

В результате экспериментов с бинарной классификацией было получено достаточно высокое качество определения того, был бы за ответ поставлен «зачёт» или нет, с F-мерой 0.87 и выше, достаточно стабильной вне зависимости от конкретного разбиения на обучающую и тестовую выборки. Это показывает высокие возможности для автоматизации проверки коротких ответов обучающихся, выполняемой по принципу «зачтено»/«не зачтено». В то же время, для различных корпусов наивысшая достигнутая F-мера различалась достаточно сильно, от 0.87 (корпус вопросов про базы данных) до 0.97 (корпус вопросов про фреймворк Qt), и это различие не является следствием вариации из-за разделения данных на обучающую и тестовую выборки. Следовательно, даже для относительно простой задачи качество может существенно изменяться в зависимости от особенностей предметной области. Эксперименты с разбиением LOCO подтвердили, что для бинарной классификации высокие показатели полноты (до 1.00) и F-меры (до 0.97) достижимы даже при тестировании модели на вопросах, не участвовавших в её обучении. Однако при этом сохраняется высокая дисперсия ROC-AUC, указывающая на возможный недостаток данных.

Качество тернарной классификации ответов на полностью верные, частично верные и неверные оказалось существенно ниже, чем качество бинарной классификации: наивысшая достигнутая F-мера для различных корпусов составила от 0.46 (корпус общих вопросов по ИТ) до 0.58 (корпус вопросов про базы данных). Задача отделения полностью верного ответа от частично верного оказалась достаточно сложной, причем нужно отметить, что и эксперты-преподаватели часто испытывают с ней затруднения. Таким образом, результаты экспериментов показывают, что при попытке автоматизировать оценку коротких ответов по более детальной шкале, чем «зачтено» / «не зачтено» шкале основную сложность представляет именно отделение полностью верных ответов от частично верных. Применение LOCO в тернарной классификации выявило дополнительную сложность: модели демонстрируют ещё больший разброс результатов (стандартное отклонение F-меры до 0.24), что подчёркивает высокую зависимость качества от конкретной формулировки вопроса и необходимость существенного увеличения данных для повышения стабильности.

В ходе экспериментов наилучшие результаты были получены при использовании для представления текста энкодерных моделей семейства BERT в паре с категориальным бустингом. То, что этот способ представления текста и классификатор показали себя примерно одинаково хорошо при оценке ответов на вопросы как по ИТ-дисциплинам, так и по истории, дает основания предполагать, что их можно так же успешно использовать и для автоматизации проверки ответов на любую тематику в рамках учебного процесса.

Эксперимент с аугментацией данных показал, что с помощью относительно несложного и нетребовательного к вычислительным ресурсам набора правил можно значительно увеличить объём обучающей выборки, что позволяет получить существенный прирост качества, подняв качество тернарной классификации до F-меры 0.90 и выше.

Табл. 6. Сравнительная таблица экспериментов с аугментацией и без нее.
Table 6. Comparative table of experiments with and without augmentation.

Модель + классиф.	F (std_F)	AUC (std_{auc})	Аугментация	Сплиттер
Бинарная классификация				
general-questions-masters-2025				
MiniLM-L12 + кат. буст.	0.96 (0.05)	0.95 (0.10)	Да	5-Fold (отдельно)
rubert-tiny2 + SVM	0.94 (0.10)	0.95 (0.15)	Нет	5-Fold (отдельно)
history-2025-summer				
rubert-tiny2 + кат. буст..	0.98 (0.04)	0.93 (0.18)	Да	5-Fold (отдельно)
bert-uncased + SVM	0.97 (0.01)	0.78 (0.14)	Нет	5-Fold (вместе)
Тернарная классификация				
general-questions-masters-2025				
MiniLM-L12 + кат. буст.	0.82 (0.17)	0.97 (0.05)	Да	5-Fold (отдельно)
ruBert-large + центр. класс.	0.47 (0.19)	0.51 (0.28)	Нет	5-Fold (отдельно)

Заключение

Авторами была проведена комплексная исследовательская работа по оценке методов автоматической классификации коротких открытых ответов на русском языке. В рамках исследования были созданы и опубликованы в открытом доступе четыре оригинальных корпуса русскоязычных ответов учащихся, охватывающих различные предметные области и учебные контексты (от экзаменов до текущего контроля).

Проведено масштабное всестороннее сравнение различных методов решения задачи. Установлена принципиальная возможность высококачественной автоматической оценки для бинарного сценария («зачет/незачет»), где были достигнуты значения F-меры до 0.97. Показано, что для этой задачи не существует единственной доминирующей пары «модель + классификатор», но многие комбинации демонстрируют высокую эффективность.

Выявлена и проанализирована ключевая проблема, ограничивающая качество в мультиклассовых сценариях – сильный дисбаланс классов в корпусах и недостаточный объём данных для обучения моделей распознаванию тонких различий между ответами различной степени правильности. Наивысшие результаты для тернарной классификации без аугментации достигли F-меры, равной 0.58. Эксперименты с использованием схемы валидации LOCO выявили дополнительные сложности, связанные с переносом моделей на новые формулировки вопросов, особенно для тернарной классификации. Это указывает на важность разработки методов, устойчивых к изменению контекста задания.

Предложен и успешно апробирован метод предметно-ориентированной аугментации данных, который позволил радикально улучшить качество моделей. В результате его применения существенно улучшилось качество: в бинарной классификации F-мера возросла до 0.98, а в тернарной – до 0.91, что подтверждает гипотезу о значительной роли объёма и репрезентативности данных. Несмотря на впечатляющее улучшение метрик F-меры, точности и

полноты после аугментации, сохраняющаяся низкая стабильность метрики AUC-ROC (стандартное отклонение 0.10–0.22) указывает на необходимость дальнейшей работы.

Перспективы дальнейших исследований видятся в следующих направлениях. Это эксперименты с другими методами аугментации, такими как генеративные нейронные сети (DeepSeek, ChatGPT, QWEN), обратный перевод и модификация текста (удаление/добавление стоп-слов); комбинирование нескольких методов аугментации для синергетического эффекта; увеличение объёма и разнообразия исходных размеченных корпусов; исследование возможностей few-shot и zero-shot обучения с использованием самых современных больших языковых моделей.

Проведенная работа является основой для создания практических систем автоматизированной оценки образовательных результатов на русском языке, способных эффективно работать в реальных учебных процессах.

Список литературы / References

[1]. Azevedo R., Bouchet F., Duffy M., Harley J., Taub M., Trevors G., Cerezo R. Lessons learned and future directions of MetaTutor: Leveraging multichannel data to scaffold self-regulated learning with an intelligent tutoring system. Frontiers in Psychology, vol. 13, 2022, pp. 813632. DOI: 10.3389/fpsyg.2022.813632.

[2]. Tisha S. M., Oregon R. A., Baumgartner G., Alegre F., Moreno J. An automatic grading system for a high school-level computational thinking course. In Proc. of the 4th International Workshop on Software Engineering Education for the Next Generation, 2022, pp. 20-27. DOI: 10.1145/3528231.3528357.

[3]. Gao R., Merzdorf H.E., Anwar S., Hipwel M. C., Srinivasa A.R. Automatic assessment of text-based responses in post-secondary education: A systematic review. Computers and Education: Artificial Intelligence, vol. 6, 2024, pp. 100206. DOI: 10.1016/j.caeai.2024.100206.

[4]. Han M., Zhang X., Yuan X., Jiang J., Yun W., Gao C. A survey on the techniques, applications, and performance of short text semantic similarity. Concurrency and Computation: Practice and Experience, 33 (5), 2021, pp. e5971. DOI: 10.1002/cpe.5971.

[5]. Mendonça P.C., Quintal F., Mendonça F. Evaluating LLMs for automated scoring in formative assessments. Applied Sciences, 15(5), 2025, pp. 2787. DOI: 10.3390/app15052787.

[6]. Леонов А.Г., Мартынов Н.С., Мащенко К.А., Холькина А.А., Шляхов А.В. Автоматизация проверки семантической составляющей текстовых ответов обучающихся в цифровой образовательной платформе. Программные продукты и системы, том 37, вып. 3, 2024 г., стр. 440–452. DOI: 10.15827/0236-235X.142.440-452 / Leonov A.G., Martynov N.S., Mashchenko K.A., Kholkina A.A., Shlyakhov A.V. Automation of semantic analysis for textual responses of studentsfor textual responses of students in a digital educational platform. Software & Systems, 2024, vol. 37, issue 3, pp. 440-452 (in Russian). DOI: 10.15827/0236-235X.142.440-452.

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

[8]. Weegar R., Idestam-Almquist P. Reducing workload in short answer grading using machine learning. International Journal of Artificial Intelligence in Education, 34(2), 2024, pp. 247-273. DOI: 10.1007/s40593-022-00322-1.

[9]. Лагутина Н.С., Лагутина К.В. Обзор моделей автоматической оценки сходства ответа учащегося с эталонным ответом. Моделирование и анализ информационных систем, том 32, вып. 1, 2025 г., стр. 42-65. DOI: 10.18255/1818-1015-2025-1-42-65 / Lagutina N.S., Lagutina K.V. A survey of models for automatic assessment of similarity of student's answer to the reference answer. Modeling and Analysis of Information Systems, 2025, vol. 32, issue 1, pp. 42-65 (in Russian). DOI: 10.18255/1818-1015-2025-1-42-65.

[10]. Camus L., Filighera A. Investigating transformers for automatic short answer grading. In Proc. of the Artificial Intelligence in Education: 21st International Conference, 2020, pp. 43-48. DOI: 10.1007/978-3-030-52240-7_8.

[11]. Ahmed A., Joorabchi A., Hayes M. J. On Deep Learning Approaches to Automated Assessment: Strategies for Short Answer Grading, vol. 2, 2022, pp. 85-94. DOI: 10.5220/0011082100003182.

[12]. Shashavali D., Vishwjeet V., Kumar R., Mathur G., Nihal N., Mukherjee S., Patil S. V. Sentence similarity techniques for short vs variable length text using word embeddings. Computación y Sistemas, 23(3), 2019, pp. 999-1004. DOI: 10.13053/cys-23-3-3273.

[13]. Li H., Chu Y., Yang K., Copur-Gencturk,Y., Tang J. (2025) LLM-based Automated Grading with Human-in-the-Loop. arXiv preprint arXiv:2504.05239 (online). Доступно по ссылке: https://arxiv.org/abs/2504.05239, 01.10.2025.
[14]. Flodén J. Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. British educational research journal, 51(1), 2025, pp. 201-224. DOI: 10.1002/berj.4069.
[15]. Jauhiainen J.S., Garagorry Guerra A. Generative AI in education: ChatGPT-4 in evaluating students' written responses. Innovations in Education and Teaching International, 62(4), 2025, pp. 1377-1394. DOI: 10.1080/14703297.2024.2422337.
[16]. Ferreira M.R., Pereira J.C., Rodrigues L., Pereira F.D., Cabral L., Costa N., Ramalho G., Gasevic D. Automatic Short Answer Grading in the LLM Era: Does GPT-4 with Prompt Engineering beat Traditional Models? In Proc. of the 15th international learning analytics and knowledge conference, 2025, pp. 93-103. DOI: 10.1145/3706468.3706481.
[17]. Kortemeyer G. Performance of the pre-trained large language model GPT-4 on automated short answer grading. Discover Artificial Intelligence, 4(1), 2024, pp. 47. DOI: 10.1007/s44163-024-00147-y
[18]. Chang L.-H., Ginter F. Automatic Short Answer Grading for Finnish with ChatGPT. In Proc. of the AAAI Conference on Artificial Intelligence, 38(21), 2024, pp. 23173-23181. DOI: 10.1609/aaai.v38i21.30363
[19]. Bani Saad M., Jackowska-Strumillo L., Bieniecki, W. Hybrid ANN-Based and Text Similarity Method for Automatic Short-Answer Grading in Polish. Applied Sciences, 15(3), 2025, pp. 1605. DOI: 10.3390/app15031605.
[20]. Леонов А.Г., Мартынов Н.С., Мащенко К.А., Холькина А.А., Шляхов А.В. Автоматизация проверки семантической составляющей текстовых ответов обучающихся в цифровой образовательной платформе. Программные продукты и системы, том 37, вып. 3, 2024, стр. 440-452. DOI: 10.15827/0236-235X.142.440-452 / Leonov A.G., Martynov N.S., Mashchenko K.A., Kholkina A.A., Shlyakhov A.V. Automation of semantic analysis for textual responses of students in a digital educational platform. Software & Systems, 2024, vol. 37, Issue 3, 2024, pp. 440–452 (in Russian). DOI: 10.15827/0236-235X.142.440–452.
[21]. Minnegalieva, C.B., Kashapov, I.I., Morozova, O.D. Automated Grading of Students' Short Answers Using Language Models. Automatic Documentation and Mathematical Linguistics, 58 (Suppl 3), 2024, pp. S109-S114. DOI: 10.3103/S0005105525700177.
[22]. Del Gobbo E., Guarino A., Cafarelli B., Grilli L., Limone P. Automatic evaluation of open-ended questions for online learning. A systematic mapping. Studies in Educational Evaluation, vol. 77, 2023, pp. 101258. DOI: 10.1016/j.stueduc.2023.101258.
[23]. Наборы коротких ответов с оценками. Доступно по адресу: https://gitverse.ru/Shtepser/asag-datasets, дата обращения 17.03.2026.
[24]. Проект «Абрис», образовательный портал факультета ИВТ. Доступно по адресу: https://abris.yarsu.ru, дата обращения: 17.03.2026. Для входа в демонстрационный режим использовать логин 'demouser', пароль: 'demouser'.

Информация об авторах / Information about authors

Владислав Николаевич Копнин – помощник исследователя в научно-исследовательской лаборатории FRUCT-YSU, студент ЯрГУ. Сфера научных интересов: компьютерная лингвистика, моделирование текстов, нейронные сети, искусственный интеллект.

Vladislav Nikolaevich Kopnin – research assistant at the FRUCT-YSU research laboratory, student at YarSU. Research interests: computational linguistics, text modeling, neural networks, artificial intelligence.

Ксения Владимировна Лагутина – научный сотрудник в научно-исследовательской лаборатории FRUCT-YSU, доцент, кандидат технических наук, доцент кафедры вычислительных и программных систем ЯрГУ. Сфера научных интересов: компьютерная лингвистика, моделирование текстов, нейронные сети, искусственный интеллект.

Ksenia Vladimirovna Lagutina – Cand. Sci. (Tech.), Assoc. Prof., researcher at the FRUCT-YSU research laboratory, associate professor at the Department of computing and software systems at YarSU. Research interests: computational linguistics, sentiment analysis, applied statistics.

Анатолий Юрьевич Полетаев – кандидат технических наук, младший научный сотрудник в научно-исследовательской лаборатории FRUCT-YSU, старший преподаватель кафедры компьютерных сетей ЯрГУ. Сфера научных интересов: компьютерная лингвистика, анализ тональности, прикладная статистика.

Anatoly Yuryevich Poletaev – Cand. Sci. (Tech.), Assoc. Prof., junior researcher at the FRUCT-YSU research laboratory, senior lecturer at the Department of computer networks at YarSU. Research interests: computational linguistics, text modeling, neural networks, artificial intelligence.

Надежда Станиславовна Лагутина – кандидат физико-математических наук, доцент, старший научный сотрудник в научно-исследовательской лаборатории FRUCT-YSU, доцент кафедры вычислительных и программных систем ЯрГУ. Сфера научных интересов: компьютерная лингвистика, моделирование текстов, нейронные сети, искусственный интеллект.

Nadezhda Stanislavovna Lagutina – Cand. Sci. (Phis.-Math.), senior researcher at the FRUCT-YSU research laboratory, associate professor at the Department of computing and software systems at YarSU. Research interests: computational linguistics, text modeling, neural networks, artificial intelligence.