Science for Education Today

2025. Том 15. № 6 http://sciforedu.ru ISSN 2658-6762

УДК 004.8+51-77+37.031
DOI: 10.15293/2658-6762.2506.07

Научная статья / Research Full Article
Язык статьи: русский / Article language: Russian

Исследование возможностей генеративного искусственного интеллекта для формирования оценочной обратной связи, близкой к экспертной, при проверке математических задач открытого типа

М. А. Лукоянова[1], А. В. Данилов[1], Р. Р. Зарипова[1], Л. Л. Салехова[1], Н. И. Батрова[1]

[1] Казанский федеральный университет, Казань, Россия

Проблема и цель. *В современной образовательной практике существует противоречие между активным внедрением генеративного искусственного интеллекта и недостаточной изученностью его возможностей для предоставления оценочной обратной связи, направленной на развитие математической грамотности школьников. Цель исследования заключается в выявлении возможностей использования генеративной языковой модели как инструмента педагога для формирования оценочной обратной связи, близкой к экспертной, при проверке математических задач открытого типа.*

Методология. *Исследование основано на системно-деятельностном, критериально-ориентированном и компаративном подходах. Применялись методы теоретического анализа научной литературы, критериального оценивания в сочетании с техниками промпт-инжиниринга, а также количественного и качественного анализа для определения согласованности между оценочной обратной связью, сформированной генеративной языковой моделью, и оценочной обратной связью, предоставленной экспертом. Выборку составили 51 учащийся.*

Результаты. *В результате исследования экспериментально подтверждена возможность применения генеративного искусственного интеллекта для формирования оценочной обратной связи в математическом образовании.*

Разработана и обоснована эффективная стратегия автоматизации процесса проверки математических задач открытого типа на основе критериального оценивания и техник

***Финансирование проекта*:** Исследование выполнено в рамках реализации гранта, предоставленного в 2024 году Академией наук Республики Татарстан на осуществление фундаментальных и прикладных научных работ в научных и образовательных организациях, предприятиях и организациях реального сектора экономики Республики Татарстан № 23/2024-ФИП по теме «Развитие математической грамотности школьников-билингвов на основе методов машинного обучения и искусственного интеллекта».

Библиографическая ссылка: Лукоянова М. А., Данилов А. В., Зарипова Р. Р., Салехова Л. Л., Батрова Н. И. Исследование возможностей генеративного искусственного интеллекта для формирования оценочной обратной связи, близкой к экспертной, при проверке математических задач открытого типа // Science for Education Today. – 2025. – Т. 15, № 6. – С. 151–174. DOI: http://dx.doi.org/10.15293/2658-6762.2506.07

✉ *Автор для корреспонденции: Наиля Ильдусовна Батрова, nibatrova@gmail.com*

промпт-инжиниринга с использованием генеративной языковой модели GigaChat Pro. Эмпирически установлена умеренная согласованность между оценочной обратной связью, сгенерированной GigaChat Pro, и оценкой педагога-эксперта по ключевым метрикам: точность (Accuracy) составила 73 %, коэффициент Коэна (k) достиг 0,57, семантическое соответствие текстовых комментариев (BertScore F1) – 0,614.

Заключение. *Проведенное исследование позволяет сделать вывод, что генеративная языковая модель обладает значительным потенциалом для трансформации практики педагогического оценивания математических задач открытого типа. Выявлены следующие возможности применения генеративной языковой модели: автоматизация процесса оценочной обратной связи экспертного уровня; персонализация оценочной обратной связи; масштабирование практики критериально-ориентированного оценивания. Повышению качества оценочной обратной связи будут способствовать: оптимизация оценочных промптов, внедрение мультиагентной верификации и механизмов выборочного педагогического контроля.*

Ключевые слова: *оценочная обратная связь; генеративные языковые модели; критериальное оценивание; техники промпт-инжиниринга; задачи открытого типа; математическая грамотность.*

Постановка проблемы

Современное образование сталкивается с вызовами в обеспечении эффективной, объективной и своевременной оценочной обратной связи по разнообразным учебным заданиям, особенно в случаях открытых и развернутых ответов, требующих комплексного анализа. Ручное выставление оценок и формирование обратной связи занимает значительное время и подвержено субъективности. Использование искусственного интеллекта, включая генеративные языковые модели и техники промпт-инжиниринга, открывает новые возможности для автоматизации процессов оценки и обратной связи, а также позволяет повысить объективность, оперативность и персонализацию образовательного процесса.

Многие исследователи сходятся во мнении, что генеративный искусственный интеллект (ИИ) является одним из перспективных направлений в области персонализированного обучения и автоматизации рутинных задач. H. Crompton и D. Burke [1] особое внимание уделяют важности новых ИИ-инструментов на примере ChatGPT для автоматизации оценки работ, обратной связи и аналитики. Необходимость адаптации образовательных процессов к новым технологиям, выявления педагогических возможностей генеративного ИИ для развития когнитивной активности обучающихся подчеркивается в исследовании Е. А. Поспеловой с соавторами [2], в котором также даются рекомендации для образовательных организаций по цифровой трансформации и адаптации под персонализированное обучение.

В своем исследовании Т. А. Чекалина [3] приходит к выводу, что применение генеративного ИИ может способствовать персонализации образовательного процесса и развитию новых форм взаимодействия между обучающимися и педагогами. К ключевым возможностям использования генеративного ИИ, по мнению N. S. Alotaibi и A. H. Alshehri [4], относятся персонализированное обучение и адаптация образовательного контента под индивидуальные потребности учащихся, а также автоматизация рутинных задач по оценке и обратной связи.

Возможности и ограничения генеративного ИИ для автоматизации проверки работ учащихся, а именно насколько адекватно и точно такие инструменты могут воспроизводить экспертное мнение преподавателя, рассмотрены в работах I. T. Awidi [5], A. Kinder с соавторами [6]. Как отмечают ученые Университета Дикина (M. Bearman, J. Tai, P. Dawson, D. Boud, R. Ajjawi) [7] и Национального Университета Тайваня (C.-H. Chiang, H.-Y. Lee) [8] важно критически оценивать качество генерируемого текста для обратной связи, поскольку в нем могут содержаться ошибки и устаревшая информация, а также учитывать качество генерации вымышленных или нерелевантных фактов.

При использовании генеративного ИИ в образовании исследователи Е. А. Поспелова с соавторами [2], Т. А. Чекалина [3], M. Bearman с соавторами [7] подчеркивают важность решения вопросов, связанных с соблюдением этики, правового регулирования и методологического сопровождения.

Таким образом, генеративный ИИ является эффективным инструментом педагога, позволяющим обеспечить персонализированное обучение за счет адаптации контента под нужды каждого ученика и автоматизировать рутинные процессы проверки работ и обратной связи, что в совокупности способствует развитию когнитивной активности учащихся, созданию новых форм образовательного взаимодействия и цифровой трансформации учебного процесса в целом.

Как отмечают J. Meyer с соавторами [9], генеративные языковые модели могут быть адаптированы к различным задачам посредством инструкций, тогда как разработка автоматизированных систем для оценочной обратной связи является трудоемкой и дорогостоящей, поскольку требует дополнительного обучения с использованием больших наборов данных, аннотированных человеческими оценками.

В последние годы в обучении иностранным языкам активно используется генеративные ИИ [10], с помощью которых предоставляется обратная связь для проверки письменных работ на иностранном языке. M. G. Hahn с соавторами [11], С. В. Боголепова и М. Г. Жаркова [12] описывают использование языковых моделей для оценки студенческих эссе по иностранному языку, формулировки обратной связи по качеству студенческих работ, снижения предвзятости и сокращения времени, необходимого на проверку преподавателем. O. Zeevy-Solovey [13], T. Kincl и соавторами [14] приходят к выводу, что генеративный ИИ демонстрирует высокую эффективность в обработке больших объемов работ и оценки по формальным критериям, но при этом сталкивается с трудностями в оценивании сложных аспектов, связанных с аргументацией, креативностью и глубиной содержания. Авторы подчеркивают необходимость применения гибридного подхода, где за генеративным ИИ закрепляется роль ассистента педагога [13; 14].

В современных исследованиях описаны направления применения генеративного ИИ для формирования оценочной обратной связи при обучении математике. H. McNichols с соавторами[1] пишет об ограниченной способности языковых моделей к глубокому пониманию математических ошибок учащихся. Дан-

[1] McNichols H., Lee J., Fancsali S., Ritter S., Lan A. Can Large Language Models Replicate ITS Feedback on Open-Ended Math Questions? // Proceedings of the 17th International Conference on Educational Data Mining. – 2024. – P. 769–775. DOI: https://doi.org/10.5281/zenodo.12729946

ная проблема подтверждается в исследованиях S. Baral с соавторами[2], В. К. Колобаева и И. К. Морозовой[3]. Возможности генеративного ИИ для индивидуализации обучения отмечают А. А. Бабкина и Н. А. Андрющечкина[4], Е. И. Маркин с соавторами[5], при этом авторы выделяют ограничения, связанные с недостаточной точностью при оценке глубоких математических рассуждений.

Таким образом, остается малоизученной проблема формирования оценочной обратной связи при решении задач открытого типа по математике. В таких областях, как математика, предоставление эффективной обратной связи может быть более сложным, поскольку языковые модели демонстрируют повышенную тенденцию к системным галлюцинациям[6].

Исследования M. I. Núñez-Peña с соавторами [15], E. R. Fyfe и S. A. Brown [16] подчеркивают важность оценочной обратной связи от педагога при обучении математике, так как она снижает математическую тревожность и развивает способность к рассуждению.

Существующее противоречие между быстрым развитием ИИ-инструментов для оценивания и недостаточной изученностью их применения к решениям математических задач открытого типа актуализирует проблему соответствия автоматизированной оценочной обратной связи экспертной оценке педагога.

В связи с этим цель исследования заключается в выявлении возможностей использования генеративной языковой модели как инструмента педагога для формирования оценочной обратной связи, близкой к экспертной, при проверке математических задач открытого типа.

Методология исследования

Исследование осуществлялось на основе совокупности взаимодополняющих методологических подходов: системно-деятельностного подхода, который приобретает особую значимость при работе с задачами открытого типа, поскольку именно в процессе их решения формируется способность к применению математических знаний в нестандартных ситуациях, что составляет основу математической грамотности; критериально-ориентированного подхода, составляющего основу для формализации критериев оценки решений математических задач открытого типа; компаративного подхода, который используется для анализа согласованности оценочной обратной связи, генерируемой языковой моделью, с экспертной оценкой.

[2] Baral S., Worden E., Lim W.-C., Luo Z., Santorelli C., Gurung A., Heffernan N. Automated Assessment in Math Education: A Comparative Analysis of LLMs for Open-Ended Responses // Proceedings of the 17th International Conference on Educational Data Mining. – 2024. – P. 732–737. DOI: https://doi.org/10.5281/zenodo.12729932

[3] Колобаев В. К., Морозова И. К. Возможности использованная искусственного интеллекта в обучении математике // Мир педагогики и психологии. – 2024. – № 8. – С. 39–43.

[4] Бабкина А. А., Андрющечкина Н. А. Применение искусственного интеллекта в математике // Международный журнал гуманитарных и естественных наук. – 2023. – № 11-2. – С. 178–181.

[5] Маркин Е. И., Зупарова В. В., Панфилова М. И. Разработка системы LMS с использованием больших языковых моделей для автоматизации проверки программных заданий // XXI век: итоги прошлого и проблемы настоящего плюс. – 2024. – Т. 13, № 4. – С. 90–94.

[6] Capellini R., Atienza F., Sconfield M. Knowledge accuracy and reducing hallucinations in LLMs via dynamic domain knowledge injection // Research Square. – 2024. – P. 1–8. DOI: https://doi.org/10.21203/rs.3.rs-4540506/v1

В качестве методов исследования применялся теоретический анализ актуальных проблем использования генеративных языковых моделей в образовательном процессе, включающий сравнительно-аналитический обзор источников по теме исследования. Как показал анализ литературы, применение метода критериального оценивания в сочетании с техниками промпт-инжиниринга позволяет обеспечить гибкую адаптацию генеративной языковой модели в формировании эффективной, объективной и своевременной оценочной обратной связи при проверке математических задач открытого типа, решение которых требует глубокого понимания контекста, логических рассуждений и оценочных суждений. Количественные методы применялись для определения точности и согласованности между оценочной обратной связью, генерируемой языковой моделью, и предоставленной педагогом-экспертом; качественные методы обработки эмпирических данных – для определения их семантического соответствия.

Для оценки возможностей применения генеративной языковой модели как инструмента педагога для формирования оценочной обратной связи, близкой к экспертной, при проверке математических задач открытого типа была проведена экспериментальная работа, в ходе которой группе из 51 учащегося 5-х классов была предложена математическая задача открытого типа, направленная на развитие математической грамотности.

Результаты исследования

Описание объекта исследования и методики оценивания

В рамках исследования была выбрана эталонная задача открытого типа по математике [7] на примере ситуации «Парусники», направленная на развитие математической грамотности учащихся 5-х классов. Согласно методическим рекомендациям по формированию функциональной грамотности обучающихся 5–9-х классов во внеурочной деятельности, разработанным Г. С. Ковалевой с соавторами[8], математическая грамотность школьника включает развитие следующих умений:

– математические знания, которые необходимы для повседневной практической деятельности, восприятия и интерпретации разнообразной информации;

– математический стиль мышления, который проявляется в определенных приемах и методах мышления (например, анализ и синтез, классификация и систематизация), логическое мышление, обеспечивающее возможность формулировать, обосновывать и доказывать суждения;

– понимание особенностей применения математики для решения научных и прикладных задач.

При решении задачи «Парусники» осуществляется оценка способности учащихся планировать ход решения задачи, извлекать и соотносить информацию в тексте и таблице, выполнять прикидку результата действия с ве-

[7] Ковалёва Г. С. Математическая грамотность. Сборник эталонных задач / Г. С. Ковалёва, Л. О. Рослова, К. А. Краснянская, О. А. Рыдзе, Е. С. Квитко. – М.; СПб.: Просвещение, 2020. – 79 с. URL: https://edu.tatar.ru/upload/storage/org1866/files/книга.pdf

[8] Методические рекомендации по формированию функциональной грамотности обучающихся 5-9 классов во внеурочной деятельности (с использованием открытого банка заданий на основе программы курса внеурочной деятельности «Функциональная грамотность: учимся для жизни»). 5 класс / [Г. С. Ковалева, А. А. Бочихина, Ю. Н. Гостева и др.]; научн. ред. Г.С. Ковалева. М.: ФГБНУ «Институт стратегии развития образования», 2023. – 197 с. URL: https://edsoo.ru/wp-content/uploads/2024/01/metod_rek_fg_5-klass_2023.pdf

личинами и делать вывод, находить и учитывать в ходе решения все условия учебной задачи, соотносить результаты действий с указанными условиями, выдвигать и обосновывать гипотезу (ответ).

Содержание задачи «Парусники» представлено в виде таблицы с указанием названий судов, года спуска на воду, длины судна с бушпритом, ширины судна, класса судна, количества экипажа (чел.). Учащимся предлагается проанализировать ответы на вопрос Федора и Иры, решить правы ли они и выбрать соответствующий ответ (да или нет), далее объяснить свой выбор.

Критерии оценивания решения задачи «Парусники»[2] представлены в таблице 1.

Таблица 1

Критерии оценки для проверки математической задачи открытого типа по ситуации «Парусники»

Table 1

Evaluation criteria for checking an open-ended mathematical problem on the situation “Sailboats”

Критерии оценки решения задачи	Балл
Верно выбраны ответы и даны объяснения в обоих случаях	2
Выбран ответ «Да» и дано объяснение, что Федор прав, потому что судно, у которого длина менее 109 м и меньшая численность экипажа (51 чел), – это «Паллада», оно было спущено на воду в 1989 году. Комментарий. В таблице приведены годы спуска на воду двух судов, у которых длина меньше 109 м – «Паллада» (1989) и «Херсонес» (1988). Но у «Паллады» экипаж составляет 51 чел., а у «Херсонеса» – 55 чел. Значит, Федор прав. ИЛИ Выбран ответ «Нет», в объяснении говорится, что Ира не права, так как она не учла условие, что длина судна должна быть меньше 109 м. Комментарий: Ира не права, потому что не выполнено условие задания: «имеющий длину с бушпритом меньшую 109 м». Она указала год спуска на воду судна «Надежда» – 1992, у которого длина с бушпритом 109 м 40 см.	1
Дан другой ответ ИЛИ ответ отсутствует	0

В исследовании принял участие 51 пятиклассник, решивший данную математическую задачу открытого типа. С целью сравнительного анализа каждое решение ученика было подвергнуто двум независимым процедурам оценивания: экспертной оценке педагога по математике и автоматизированной оценке с использованием генеративной языковой модели GigaChat Pro. Критериальной основой для обоих видов оценки выступили критерии, разработанные для оценки эталонной задачи «Парусники» (табл. 1).

Результатом оценки являлась структурированная оценочная обратная связь, включающая количественную оценку (балл) и качественный анализ полноты и обоснованности объяснения решения задачи (текстовый комментарий). Данный подход к оценке решений предложенной задачи позволяет понять,

насколько оценочная обратная связь, полученная от генеративной языковой модели, является близкой к экспертной оценке.

Формирование обратной связи, сгенерированной языковой моделью GigaChat Pro как инструмента автоматизированной оценки, осуществлялось на основе разработанной авторами комплексной стратегии. Она основана на адаптации метода критериального оценивания LLM-as-a-Judge в сочетании с современными техниками промпт-инжиниринга Few-shot prompting, Role prompting and Chain-of-Thought prompting, что позволяет формализовать и стандартизировать процесс экспертной оценки средствами генеративной языковой модели GigaChat Pro.

Выбор языковой модели GigaChat Pro, разработанной компанией Сбербанк, обусловлен общедоступностью (имеет пробный бесплатный лимит); простотой использования и способностью обрабатывать и генерировать тексты на русском языке. Модель GigaChat Pro – версия для ресурсоемких задач, она обеспечивает максимальную эффективность в обработке данных, креативность и соблюдение инструкций [9, 10][17].

В зарубежных исследованиях для автоматизации процесса оценки текстов, близкой к экспертной, на основе заданных критериев используется метод LLM-as-a-Judge [11, 12, 13, 14]. Предлагаются различные варианты составления промпта для оценки при реализации LLM-as-a-Judge: direct scoring (прямое оценивание), pairwise comparison (парное сравнение), criteria-based evaluation (оценка по критериям), self-consistency и CoT [15, 16]. Важно отметить, что для оценки возможно настроить несколько параметров, например, логичность, точность, полнота, токсичность, корректность, релевантность, согласованность, краткость, креативность, наличие выдуманной информации.

Разработанная авторами стратегия включает реализацию нескольких ключевых механизмов, характерных для метода LLM-as-a-

[9] Ениватов А. С. Интеграция нейронной сети Gigachat в систему поддержки образовательного процесса // Новые информационные технологии в научных исследованиях: Материалы XXIX Всероссийской научно-технической конференции студентов, молодых ученых и специалистов, Рязань, 27–29 ноября 2024 года. – Рязань: Рязанский государственный радиотехнический университет им. В.Ф. Уткина, 2024. – С. 17–18.

[10] Косов Н. В., Краснова А. Д., Тарханова О. В. Сравнительный анализ точности результатов систем искусственного интеллекта “GIGACHAT” и “YANDEXGPT4-pro” // Приоритетные направления экономического, социального и политического развития информационного общества (ПН-2024): Материалы Международной научно-практической конференции, Тюмень, 13 декабря 2024 года. – Тюмень: Тюменский индустриальный университет, 2025. – С. 183–187.

[11] LLM-as-a-judge: a complete guide to using LLMs for evaluations. URL: https://www.evidentlyai.com/llm-guide/llm-as-a-judge

[12] Li Q., Dou S., Shao K., Chen C., Hu H. Evaluating Scoring Bias in LLM-as-a-Judge // ArXiv. – 2025. – P. 1–8. DOI: https://doi.org/10.48550/arXiv.2506.22316

[13] Gu J., Jiang X., Shi Z., Tan H., Zhai X., Xu C., Li W., Shen Y., Ma S., Liu H., et al. A survey on LLM-as-a-Judge // ArXiv. – 2024. – P. 1–52. DOI: https://doi.org/10.48550/arXiv.2411.15594

[14] Dasgupta S., Shankar H. AI Agents-as-Judge: Automated Assessment of Accuracy, Consistency, Completeness and Clarity for Enterprise Documents // ArXiv. – 2023. – P. 1–17. DOI: https://doi.org/10.48550/arXiv.2506.22485 https://arxiv.org/pdf/2506.22485

[15] Kennedy S., Seth D., Subramani D. Dissecting Physics Reasoning in Small Language Models: A Multi-Dimensional Analysis from an Educational Perspective Nicy Scaria // ArXiv. – 2025. – P. 1–22. DOI: https://doi.org/10.48550/arXiv.2505.20707

[16] Langfuse. URL: https://cloud.langfuse.com/project/clkpwwm0m000gmm094odg11gi/evals/templates?search=context&peek=cmal6wart009lyn-rdtpv6olve

Judge: оценку по критериям выставления баллов 0, 1, 2 с детализированными условиями для проверки (например, название и длина судна, экипаж, год спуска на воду) и параметры оценки «точность и полнота ответа». Таким образом, генеративная языковая модель действует как строгий эксперт, согласно заданным критериям.

Few-shot prompting применяется к сложным и чувствительным к контексту задачам [18]. Role prompting позволяет активировать специфические знания и стиль модели через назначение ей определенной роли, что формирует стиль, глубину и область знаний генерируемого ответа, делая его более релевантным и адаптированным под конкретную ситуацию или область знаний [19]. Метод Chain-of-Thought (CoT) улучшает прозрачность и точность в сложных задачах, требующих рассуждений, за счет генерации промежуточных логических шагов, минимизирует риск ошибок, позволяет получить проверяемый ход решения, структурировать оценку ответов по четким критериям [20].

В исследовании Chain-of-Thought (CoT) реализуется через структурированный пошаговый анализ решений учеников, который разбит на логические блоки.

1. Формирование контекста. Промпт начинается с явного указания условия задачи (task_description), эталонного ответа (golden_answer), критериев оценки (criteria_rubrics: баллы 0, 1, 2 с четкими условиями). Таким образом модель сначала получает всю необходимую информацию – это аналог понимания задачи.

2. Развернутые правила оценки. Критерии оценки прописаны в формате «если – то» с детализацией:

– балл 0: четко перечислены условия (нет ответа, оба неверны и т. д.);

– балл 1: разделены на два независимых сценария для Федора и Иры с конкретными требованиями к объяснениям (название судна, длина, экипаж и т. д.);

– балл 2: условие комбинации двух корректных ответов.

Это соответствует «разбиению задачи на подшаги» в CoT.

3. Примеры с пояснениями. Приведены несколько примеров с ответами учеников и обоснованием выставленного балла (например, «Федор не указал длину судна – значит, балл «0»). Это обучает модель рассуждать вслух, как в CoT, показывая цепочку: решение ученика → проверка критериев → вывод балла и текстового комментария к нему.

4. Структура запроса на оценку. Финальный запрос направляет модель сначала проверить ответ Федора, затем проверить ответ Иры и объединить результаты. Это имитирует человеческое мышление: «Сначала проверь X, потом Y, затем объединить результаты».

Таким образом, формализуется роль генеративной языковой модели с учетом заданных параметров точности и полноты, настраивается связь абстрактного названия с конкретными правилами через промпт, что позволяет расширять систему с новыми критериями без переписывания логики. В итоге модель не «думает», а применяет заранее заданные критерии, как судья. В структуре промпта criteria_name задает точность и полноту как параметр оценщика (LLM-as-a-Judge) для его реализации (табл. 2).

Таблица 2

Реализация метода LLM-as-a-Judge для автоматизированной критериальной оценки

Table 2

Using of the LLM-as-a-Judge method for automated criteria-based assessment

Рубрики оценки LLM-as-a-Judge	**Характеристика**
Фокус на конкретный критерий	criteria_name сужает роль языковой модели до оценки строго одного аспекта – точности (правильности ответа) и полноты (наличия всех требуемых деталей). Это имитирует реального судью, который проверяет работу по заранее определенному чек-листу, а не оценивает ее «в целом»
Чек-лист для языковой модели	criteria_name косвенно преобразуется в конкретные пункты для проверки внутри промпта: – точность: ответы Федора/Иры должны совпадать с golden_answer («Да» / «Нет»); – полнота: объяснения должны включать все указанные параметры (длина, экипаж и т. д.). Модель последовательно проверяет каждый пункт, как судья с чек-листом, и не отклоняется от заданного критерия
Связь с criteria_rubrics	criteria_name – это «заголовок» для рубрики, а criteria_rubrics – ее детализация. В других сценариях criteria_rubrics мог бы динамически подставлять правила оценки под выбранный критерий
Ограничение субъективности	criteria_name явно исключает другие аспекты (например, «креативность» или «грамотность»). Языковая модель действует как узкоспециализированный судья, что повышает согласованность оценок. Без четкого criteria_name модель могла бы учитывать посторонние факторы (например, длину текста или эмоциональность). Здесь же оценка привязана к точности и полноте
Интеграция с системой оценок	Параметр позволяет расширять возможности стратегии: – добавить другие параметры оценщика (например, criteria_name = «Креативность»); – для каждого параметра – свой промпт и правила. Языковая модель становится «модульным судьей», переключаемым между разными чек-листами

На основе разработанной стратегии автоматизации оценочной обратной связи с использованием генеративной языковой модели GigaChat Pro был реализован оценочный промпт для проверки решения математической задачи открытого типа «Парусники» с текстовым комментарием:

Ты учитель математики. Для учащихся 5 классов была предложена задача {problem} по математической грамотности.
Условие задачи: {problem}
Оцени решения учеников {stud_solution} по следующим критериям {criteria_rubrics} в соответствии с эталонным решением {golden_answer}
Решение ученика: {stud_solution}
Параметры оценщика: {criteria_name}
Критерии оценки: {criteria_rubrics}
Эталонное решение: {golden_answer}.
Используй только варианты «0», «1» или «2» в качестве вердикта {score}.
Оценочная обратная связь:
Балл за решение {score}
Текстовый комментарий {comment}

В процессе реализации разработанной стратегии автоматизации оценочной обратной связи, интегрирующей автоматическую критериальную оценку и техники промпт-инжиниринга, была получена оценочная обратная связь при проверке математической задачи открытого типа по ситуации «Парусники» от генеративной языковой модели GigaChat Pro с последующей экспертной валидацией.

В таблице 3 представлены примеры оценочной обратной связи, включающей оценку по шкале от 0 до 2 баллов и текстовый комментарий, при проверке генеративной языковой моделью GigaChat Pro и педагогом-экспертом решения учениками математической задачи открытого типа.

Таблица 3

Примеры оценочной обратной связи по решению задачи по математике открытого типа «Парусники»

Table 3

Examples of evaluative feedback on solving an open-ended math problem “Sailboats”

Ответ ученика	Балл (GigaChat Pro)	Комментарий (генеративная языковая модель GigaChat Pro)	Балл (педагог-эксперт)	Комментарий (педагог-эксперт)
1	2	3	4	
1) да. потому что судно которое меньше 109 см и меньшую численность экипажа это Паллада, она выпущена в 1989; 2) нет. потому в 1992 году был выпущен корабль Надежда, она не подходит к описанию	2	✅ Ты – молодец! Все верно. Разберем твое решение: Федор правильно указал, что подходящим судном является «Паллада», спущенная на воду в 1989 году, длиной 108 м 60 см и экипажем 51 человек; Ира также была точна, отметив, что в 1992 году спустили на воду «Надежду», которая не соответствует условиям задачи по длине судна	2	Верно выбраны ответы в обоих случаях и даны верные объяснения в обоих случаях

Продолжение таблицы 3

1	2	3	4	
1) да, он написал верно, ведь меньше 109 м; 2) нет, она написала неверно, ведь меньше 109 м, а не 109м	1	⚠ Хорошая работа, но есть ошибка. Разберем твое решение, Федор правильно указал год спуска на воду корабля «Паллада», удовлетворяющего условиям задачи как по длине, так и по численности экипажа. Ира указала корабль «Надежда», однако его длина составляет ровно 109 м 40 см, тогда как в условии требуется длина судна строго меньше 109 метров. Правильным решением будет указать корабль «Паллада»	1	Верно выбраны ответы в обоих случаях, но в первом случае – объяснение неполное
1) да. потому что в таблице так написано; 2) нет. потому что в таблице другой ответ	0	✗ К сожалению, решение неверное. Разберем твое решение: не обоснован выбор критерия у Федора; не приведены конкретных данных из таблицы, подтверждающих утверждение Иры, и не учтены оба условия задачи одновременно. Правильное решение: из таблицы видно, что только у корабля «Паллада» длина с бушпритом составляет менее 109 метров (108 м 60 см), а численность экипажа равна 51 человеку, что является наименьшей среди судов с длиной менее 109 метров. Таким образом, правильный ответ – 1989 год	0	Верно выбраны ответы в обоих случаях, но даны неверные объяснения
1) да, Федор прав. Он прав, в этом году длина судна с бушпритом меньше 109 м; 2) нет, Ира не права, в этом году больше чем 109 м	1	⚠ Хорошая работа, но есть ошибка. Разберем твое решение, Ответ Федора верен, он правильно указал год спуска на воду корабля «Паллада», который имеет длину с бушпритом меньше 109 метров, но не указан наименьший экипаж среди подходящих по длине судов. Ира указала неверный год, поскольку длина судна «Надежда» составляет ровно 109 м 40 см, а значит, не удовлетворяет условию задачи о длине судна строго меньше 109 метров. Правильное решение: выбрать из таблицы судно длиной менее 109 метров («Паллада») и указать его год спуска на воду – 1989	1	Верно выбраны ответы в обоих случаях, но в первом случае – объяснение неверно

Окончание таблицы 3

1	2	3	4	
1) да, по таблице можно увидеть, что в 1989 году был спущен Паллада и ее длина судна с бушпритом 108 м 60 см; 2) нет. в 1992 году был спущен корабль Надежда, его длина судна с бушпритом 109 м 40 см и численность экипажа 50	2	✓ Ты – молодец! Все верно. Разберем твое решение: Федор правильно указал, что в 1989 году была спущена «Паллада», длина которой составляет 108 м 60 см, а численность экипажа равна 51 человеку; Ира же верно заметила, что в 1992 году спустили на воду «Надежду», однако не учла, что по условию задачи требуется найти именно корабль длиной менее 109 метров	2	Верно выбраны ответы в обоих случаях

Количественный и качественный анализ оценочной обратной связи

В результате эмпирического исследования определялась согласованность оценочной обратной связи, сгенерированной языковой моделью GigaChat Pro, с экспертной оценкой на основе количественного и качественного анализа полученных данных.

Общая доля совпадения оценок, сгенерированной языковой моделью GigaChat Pro, и экспертной оценкой определялась с помощью метрики Accuracy (точность)[17], стандартной для задач классификации. В условиях многоклассовой классификации (0, 1 и 2) проведен детальный анализ ошибок, допущенных генеративной языковой моделью с использованием метрики F1-score. Данный показатель позволяет оценить баланс между точностью (Precision) и полнотой (Recall), учитывать ложные срабатывания (FP) и пропуски (FN), что повышает адекватность интерпретации результатов по сравнению с общей точностью Accuracy.

Оценка уровня согласия между обратной связью, сгенерированной языковой моделью GigaChat Pro, и экспертной обратной связью измерялась с помощью коэффициента каппа Коэна. Данная метрика была выбрана в качестве основного показателя, поскольку она учитывает возможность случайного совпадения результатов, обеспечивая тем самым статистическую обоснованную оценку согласованности.

В целях верификации смысловой близости и оценочных конструкций в текстовых комментариях, сгенерированной языковой моделью GigaChat Pro, и оценочной обратной связи, предоставленной педагогом-экспертом, проводилось измерение уровня их семантического соответствия с использованием метрики BERTScore[18].

Количественный анализ показал, что доля совпадающих оценок, полученных от языковой модели GigaChat Pro, с оценками педагога-эксперта достигает 73 % (по метрике Accuracy), что является высоким показателем для оценки математических задач открытого типа и указывает на практическую применимость генеративной языковой модели GigaChat Pro (рис. 1).

[17] Hu T., Zhou X.-H. Unveiling LLM Evaluation Focused on Metrics: Challenges and Solutions // ArXiv. – 2024. – P. 1–45. DOI: https://doi.org/10.48550/arXiv.2404.09135

[18] Бручес Е. П., Батурова Д. Т., Бондаренко И. Ю. BERTScore для русского языка // Труды Института системного программирования РАН. – 2025. – Т. 37, № 3. – С. 147–158.

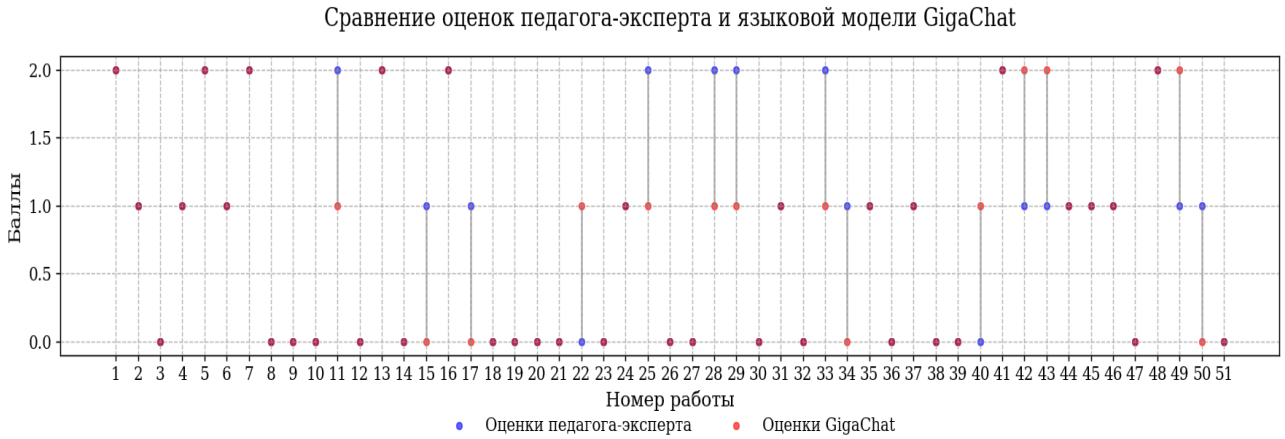

Рис. 1. Распределение оценок, полученных от педагога-эксперта и GigaChat Pro

Fig. 1. Distribution of scores received from an expert teacher and GigaChat Pro

Поскольку критериальная шкала содержит три значения баллов (0, 1 и 2), то для многоклассовой классификации проведен подробный анализ ошибок языковой модели (F1-score), который позволяет оценить баланс между точностью (Precision) и полнотой (Recall), учитывать ложные срабатывания (FP) и пропуски (FN), что делает его более надежным (табл. 4).

Таблица 4

Сравнение согласованности оценок, полученных от языковой модели GigaChat Pro, с экспертными оценками

Table 4

Comparison of the consistency between assessments generated by GigaChat Pro and expert evaluations

Класс (балл)	Precision	Recall	F1-score	Support
0 (неверно)	0,83	0,91	0,87	22
1 (частично)	0,59	0,59	0,59	17
2 (верно)	0,70	0,58	0,64	12

Для углубленного анализа структуры расхождений была построена и проанализирована матрица ошибок (рис. 2). Ее анализ выявил системные ошибки модели (занижение/завышение) по определенным критериальным категориям, позволяющие идентифицировать конкретные типы оценок, для которых требуется настройка промпта с целью минимизации выявленных смещений. Таким образом, было обнаружено 14 расхождений в баллах из 51 проверенной задачи:

– языковая модель поставила «0» вместо «1» в 3 случаях → занижение;

– языковая модель поставила «2» вместо «1» в 4 случаях → завышение;

– наибольшие проблемы с баллом «1» (занижение или завышение).

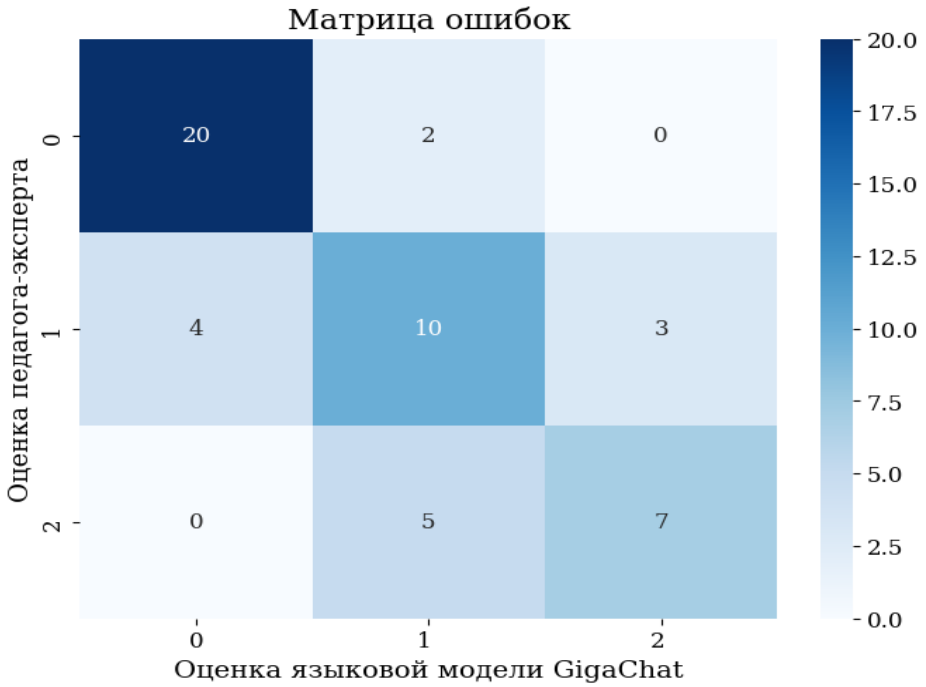

Рис.2. Матрица ошибок в оценках GigaChat Pro и педагога-эксперта
Fig. 2. Confusion matrix in GigaChat Pro and the expert teacher assessments

В исследовании выявлено, что классы примерно сбалансированы, поскольку MacroAvg (0,70) ≈ Weighted Avg (0,72). Языковая модель лучше предсказывает частые оценки (Weighted Avg = 0,72), чем редкие (MacroAvg = 0,70), что является неплохим результатом и указывает на необходимость настройки промпта с критериями для баллов 0 и 2. Следовательно, языковую модель можно использовать для проверки решения задач по математике открытого типа с дополнительной проверкой педагогом-экспертом около 30 % спорных случаев.

Уровень согласия между оценочной обратной связью**,** сгенерированной языковой моделью GigaChat Pro, и экспертной оценочной обратной связью измерялся с помощью коэффициента каппа Коэна. Согласно шкале Ландиса и Коха значение между 0,01 и 0,20 показывает незначительное согласие, между 0,21 и 0,40 – справедливое, 0,41–0,60 – умеренное, между 0,61 и 0,80 – существенное и диапазон 0,81–1 указывает на почти идеальное согласие [21]. По данным, полученным в исследовании, коэффициент каппа Коэна равен 0,57, что означает умеренное согласие между оценками языковой модели и педагога-эксперта и является статистически значимым. Это подтверждает возможность использования генеративной языковой модели GigaChat Pro в качестве эффективного инструмента автоматизированной оценки, но с необходимостью дополнительного контроля в спорных случаях.

Автоматическая оценочная обратная связь генеративной языковой модели содержит не только балл за проверку математической задачи открытого типа, но и текстовый комментарий к нему. Анализ семантического сходства между текстовыми комментариями, полученными от генеративной языковой мо-

дели GigaChat Pro, и представленными педагогом-экспертом, показал следующие результаты.

1. Семантическое совпадение наблюдается в 57,6 % случаев (Average Precision (P) ≈ 0,576), что указывает на необходимость настройки оценочного промпта.

2. В комментариях языковой модели нашли отражение 65,9 % смысловых единиц из комментариев педагога-эксперта (Average Recall (R) ≈ 0,659). Полученное значение показывает, что комментарии языковой модели умеренно соответствуют комментариям педагога-эксперта. Recall выше Precision (0,659 > 0,576) – это значит, что языковая модель неплохо покрывает смысл комментариев педагога-эксперта, но при этом включает лишние и неточные комментарии (потеря точности). 34 % упущенного смысла – возможно, языковая модель пропускает ключевые замечания.

Обобщенную оценку качества семантического соответствия отражает гармоническое среднее между Precision и Recall (BertScore F1 = 0,614), которое чем ближе к 1, тем лучше. F1 = 0,614 – типичный для задач, где есть компромисс между точностью и полнотой (например, суммаризация, машинный перевод). Для задач по математике F1 обычно ниже, чем для неспециализированных текстов, из-за специфичной терминологии, особенностей языка изложений в решениях задач учениками 5-го класса (неполные ответы, использование разговорных слов, грамматические ошибки и др.).

Заключение

Проведенное исследование доказало возможность эффективного использования генеративных языковых моделей как инструмента педагога для автоматизации процесса формирования оценочной обратной связи, близкой к экспертной, при проверке математических задач открытого типа, решение которых требует глубокого понимания контекста, логических рассуждений и оценочных суждений.

Экспериментально установлено, что оценочная обратная связь, сгенерированная языковой моделью GigaChat Pro, демонстрирует статистически значимую согласованность с экспертной оценкой.

Научная новизна и авторский вклад работы заключаются в следующем.

1. Авторами апробирована и доказана эффективность разработанной стратегии, основанной на методе критериального оценивания (LLM-as-a-Judge) в сочетании с техниками промпт-инжиниринга (Few-shot prompting, Role prompting and Chain-of-Thought), что позволяет обеспечить согласованность автоматизированной и экспертной оценки математических задач открытого типа.

2. Установлены объективные количественные показатели согласованности оценочной обратной связи, сгенерированной языковой моделью, и предоставленной педагогом-экспертом. Экспериментально подтверждено, что оценочная обратная связь, сгенерированная языковой моделью демонстрирует высокий уровень точности (73 %), что соответствует диапазону межэкспертной согласованности. Статистически значимое значение коэффициента каппа Коэна (k = 0.57) свидетельствует об умеренном уровне согласия, что подтверждает способность модели воспроизводить экспертные стратегии оценивания.

3. В результате проведенного анализа авторами выявлен характерный паттерн систематических ошибок языковой модели, позволяющий установить, что языковой модели свойственна устойчивая тенденция к занижению оценок для частично верных ответов

(балл 1). Это проявляется в значительном снижении метрики Precision и Recall для частично верного ответа по сравнению с верным (балл 2) и неверным ответом (балл 0). Данное наблюдение указывает на фундаментальную сложность формализации для генеративной языковой модели «частичной правильности» решения.

4. На основе анализа семантического сходства доказано, что генеративная языковая модель GigaChat Pro способна не только формально классифицировать ответ, но и генерировать текстовые комментарии, семантически близкие к формулировкам педагога-эксперта, что является ключевым условием для персонализации обратной связи.

5. Определены критические условия и ограничения эффективного применения генеративной языковой модели GigaChat Pro для оценивания. Для минимизации выявленных системных ошибок необходима точечная настройка промптов для коррекции смещения в оценке 1 балл. Реализация гибридного подхода в оценивании с выборочной верификацией педагогом-экспертом 20–30 % позволит нивелировать текущие погрешности языковой модели и повысить итоговую надежность системы оценивания.

Повышению качества оценочной обратной связи будет способствовать внедрение мультиагентной системы с привлечением нескольких педагогов-экспертов по математике и дополнительных генеративных языковых моделей для верификации спорных случаев.

Таким образом, вклад исследования заключается в получении новых знаний о возможностях и ограничениях генеративной языковой модели GigaChat Pro в контексте развития математической грамотности учащихся: установлены объективные метрики эффективности, выявлены специфические паттерны ошибок оценивания и определены направления для их коррекции. Полученные результаты составляют основу для дальнейших исследований в области использования генеративного ИИ как инструмента педагога и вносят вклад в развитие цифровой дидактики, предлагая научно обоснованные решения для автоматизации проверки математических задач открытого типа с сохранением качества оценочной обратной связи. Разработанная стратегия открывает перспективы для масштабирования практики детального, критериально-ориентированного оценивания, обеспечивая его доступность для неограниченного числа обучающихся без потери качества и персонализированного характера оценочной обратной связи.

СПИСОК ЛИТЕРАТУРЫ

1. Crompton H., Burke D. Artificial intelligence in higher education: the state of the field // International Journal of Educational Technology in Higher Education. – 2023. – Vol. 20. – P. 1–22. DOI: https://doi.org/10.1186/s41239-023-00392-8
2. Поспелова Е. А., Отоцкий П. Л., Горлачева Е. Н., Файзуллин Р. В. Генеративный искусственный интеллект в образовании: анализ тенденций и перспектив // Профессиональное образование и рынок труда. – 2024. – Т. 12, № 3. – С. 6–21. URL: https://www.elibrary.ru/item.asp?id=69176655 DOI: https://doi.org/10.52944/PORT.2024.58.3.001
3. Чекалина Т. А. ИИ-дидактика: новый тренд или эволюция процесса обучения? // Вестник Мининского университета. – 2025. – Т. 13, № 2. – С. 5. URL: https://elibrary.ru/item.asp?id=82539976 DOI: https://doi.org/10.26795/2307-1281-2025-13-2-5

4. Alotaibi N. S., Alshehri A. H. Prospers and Obstacles in Using Artificial Intelligence in Saudi Arabia Higher Education Institutions. The Potential of AI-Based Learning Outcomes // Sustainability. – 2023. – Vol. 15 (13). – P. 10723. DOI: https://doi.org/10.3390/su151310723
5. Awidi I. T. Comparing expert tutor evaluation of reflective essays with marking by generative artificial intelligence (AI) tool // Computers and Education: Artificial Intelligence. – 2024. – Vol. 6. – P. 100226. DOI: https://doi.org/10.1016/j.caeai.2024.100226
6. Kinder A., Briese F. J., Jacobs M., Dern N., Glodny N., Jacobs S., Leßmann S. Effects of adaptive feedback generated by a large language model: A case study in teacher education // Computers and Education: Artificial Intelligence. – 2025. – Vol. 8. – P. 100349. DOI: https://doi.org/10.1016/j.caeai.2024.100349
7. Bearman M., Tai J., Dawson P., Boud D., Ajjawi R. Developing evaluative judgement for a time of generative artificial intelligence // Assessment & Evaluation in Higher Education. – 2024. – Vol. 49 (6). – P. 893–905. DOI: https://doi.org/10.1080/02602938.2024.2335321
8. Chiang C.-H., Lee H.-y. Can Large Language Models Be an Alternative to Human Evaluations? // Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. – 2023. – Vol. 1. – P. 15607–15631. DOI: https://doi.org/10.18653/v1/2023.acl-long.870
9. Meyer J., Jansen T., Schiller R., Liebenow W., Steinbach M., Horbach A., Fleckenstein J. Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions // Computers and Education: Artificial Intelligence. – 2024. – Vol. 6. – P. 100199. DOI: https://doi.org/10.1016/j.caeai.2023.100199
10. Пак Л. Е., Крюкова А. А. О возможностях использования программ с искусственным интеллектом в обучении иностранному языку // Территория новых возможностей. Вестник Владивостокского государственного университета. – 2024. – Т. 16, № 2. – С. 81–95. URL: https://elibrary.ru/item.asp?id=67900721 DOI: https://doi.org/10.24866/VVSU/2949-1258/2024-2/081-095
11. Hahn M. G., Navarro S. M. B., De La Valentín L., Burgos D. A systematic review of the effects of automatic scoring and automatic feedback in educational settings // Institute of Electrical and Electronics Engineers Access. – 2021. – Vol. 9. – P. 108190–108198. DOI: https://doi.org/10.1109/ACCESS.2021.3100890
12. Боголепова С. В., Жаркова М. Г. Исследование потенциала генеративных моделей для оценивания эссе и обеспечения обратной связи // Отечественная и зарубежная педагогика. – 2024. – Т. 1, № 5. – С. 123–137. URL: https://elibrary.ru/item.asp?id=73431773
13. Zeevy-Solovey O. Comparing peer, ChatGPT and teacher corrective feedback in EFL writing: Students' perceptions and preferences // Technology in Language Teaching & Learning. – 2024. – Vol. 6 (3). – P. 1482. DOI: https://doi.org/10.29140/tltl.v6n3.1482
14. Kincl T., Gunina D., Novák M., Pospíšil J. Comparing Human and AI-based Essay Evaluation in the Czech Higher Education: Challenges and Limitations // Trendy v podnikání - Business Trends. – 2024. – Vol. 14 (2). – P. 25–34. DOI: https://doi.org/10.24132/jbt.2024.14.2.25_34
15. Núñez-Peña M. I., Bono R., Suárez-Pellicioni M. Feedback on students' performance: A possible way of reducing the negative effect of math anxiety in higher education // International Journal of Educational Research. – 2015. – Vol. 70. – P. 80–87. DOI: https://doi.org/10.1016/j.ijer.2015.02.005
16. Fyfe E. R., Brown S. A. Feedback influences children's reasoning about math equivalence: A meta-analytic review // Thinking & Reasoning. – 2017. – Vol. 24 (2). – P. 157–178. DOI: https://doi.org/10.1080/13546783.2017.1359208

17. Kouzminov Y., Kruchinskaia E. The Evaluation of GenAI Capabilities to Implement Professional Tasks // Foresight and STI Governance. – 2024. – Vol. 18 (4). – P. 67–76. https://elibrary.ru/item.asp?id=75194200 DOI: https://doi.org/10.17323/2500-2597.2024.4.67.76
18. Schorcht S., Buchholtz N., Baumanns L. Prompt the problem – investigating the mathematics educational quality of AI-supported problem solving by comparing prompt techniques // Frontier Education. – 2024. – Vol. 9. – P. 1–15. DOI: https://doi.org/10.3389/feduc.2024.1386075
19. Qian Y. Prompt Engineering in Education: A Systematic Review of Approaches and Educational Applications // Journal of Educational Computing Research. – 2025. – Vol. 0 (0). – P. 1–37. DOI: https://doi.org/10.1177/07356331251365189
20. Lee G. G., Latif E., Wu X., Liu N., Zhai X. Applying large language models and chain-of-thought for automatic scoring // Computers and Education: Artificial Intelligence. – 2024. – Vol. 6. – P. 100213. DOI: https://doi.org/10.1016/j.caeai.2024.100213
21. Albakkosh I. Using Fleiss' kappa coefficient to measure the intra and inter- rater reliability of three AI software programs in the assessment of EFL learners' story writing // International Journal of Educational Sciences and Arts. – 2024. – Vol. 3 (1). – P. 69–96. DOI: https://doi.org/10.59992/IJESA.2024.v3n1p4

Поступила: 19 августа 2025 Принята: 11 ноября 2025 Опубликована: 31 декабря 2025

Заявленный вклад авторов:

М.А. Лукоянова: обзор литературы, разработка стратегии AI-промптинга, проведение экспериментов с языковой моделью и анализ результатов, структурирование учебного контента, подготовка окончательного варианта рукописи для подачи.

А.В. Данилов: организация исследования, разработка дизайна исследования.

Р.Р. Зарипова: обработка данных, выполнение статистических процедур, проведение экспертной оценки.

Л.Л. Салехова: обзор литературы, координация процедуры экспертной оценки.

Н.И. Батрова: разработка концепции исследования, разработка стратегии AI-промптинга, проведение экспериментов с языковой моделью, интерпретация результатов, подготовка окончательного варианта рукописи для подачи.

Все авторы ознакомились с результатами работы и одобрили окончательный вариант рукописи.

Информация о конфликте интересов:

Авторы декларируют отсутствие явных и потенциальных конфликтов интересов в связи с публикацией данной статьи

Информация об авторах

Лукоянова Марина Александровна
кандидат педагогических наук, доцент,
кафедра билингвального и цифрового образования,
Казанский федеральный университет,
Кремлёвская, д. 18, 420021, г. Казань, Россия.
ORCID ID: https://orcid.org/0000-0002-5630-0876
SPIN-код: 5983-3077
E-mail: marina-lkn@yandex.ru

Данилов Андрей Владимирович
кандидат педагогических наук, доцент,
кафедра билингвального и цифрового образования,
Казанский федеральный университет,
Кремлёвская, д. 18, 420021, г. Казань, Россия.
ORCID ID: https://orcid.org/0000-0002-2358-1157
SPIN-код: 8525-5480
E-mail: tukai@yandex.ru

Зарипова Рината Раисовна
кандидат педагогических наук, доцент,
кафедра билингвального и цифрового образования,
Казанский федеральный университет,
Кремлёвская, д. 18, 420021, г. Казань, Россия.
ORCID ID: https://orcid.org/0000-0003-4514-5513
SPIN-код: 4115-9105
E-mail: rinata-z@yandex.ru

Салехова Ляйля Леонардовна
доктор педагогических наук, профессор,
кафедра билингвального и цифрового образования,
Казанский федеральный университет,
Кремлёвская, д. 18, 420021, г. Казань, Россия.
ORCID ID: https://orcid.org/0000-0002-8177-3739
SPIN-код: 7607-9153
E-mail: salekhova2009@gmail.com

Батрова Наиля Ильдусовна
кандидат педагогических наук, доцент,
кафедра билингвального и цифрового образования,
Казанский федеральный университет,
Кремлёвская, д. 18, 420021, г. Казань, Россия.
ORCID ID: https://orcid.org/0000-0002-1945-3507
SPIN-код: 2342-5952
E-mail: nibatrova@gmail.com

Science for Education Today

2025, vol. 15, issue 6 http://en.sciforedu.ru/ ISSN 2658-6762

DOI: 10.15293/2658-6762.2506.07 Research Full Article / Article language: Russian

Research on the potential of generative artificial intelligence for providing expert-level evaluative feedback in open-ended mathematical problems assessment

Marina A. Lukoyanova[1], Andrey V. Danilov[1], Rinata R. Zaripova[1], Leila L. Salekhova[1], Nailya I. Batrova [1]

[1] Kazan (Volga Region) Federal University, Kazan, Republic of Tatarstan, Russian Federation

Abstract

Introduction. *Modern education faces a contradiction between the active integration of generative artificial intelligence and its underexplored potential for providing evaluative feedback in development students' mathematical literacy. The purpose of the article is to identify the potential of using a generative language model as a teacher's tool for generating expert-level evaluative feedback when assessing open-ended mathematical problems*

Materials and Methods. *The research is based on systemic-activity, criteria-oriented, and comparative approaches. Methods employed included theoretical analysis of scholarly literature, criteria-based assessment combined with prompt engineering techniques, as well as quantitative and qualitative analysis to determine the agreement between the evaluative feedback generated by the language model and that provided by a human expert. The sample consisted of 51 students*

Results. *The research experimentally confirmed the feasibility of using generative artificial intelligence for providing evaluative feedback in mathematics education. An effective strategy for automating the assessment of open-ended mathematical problems was developed and substantiated, based on criteria-based assessment and prompt engineering techniques using GigaChat Pro language model. Empirical data revealed a moderate agreement between the evaluative feedback generated by GigaChat Pro and that provided by an expert teacher: accuracy reached 73%, Cohen's coefficient (k) was 0,57, and the semantic similarity of textual comments (BertScore F1) was 0,614.*

Acknowledgments

The study was financially supported by the Academy of Sciences of the Republic of Tatarstan by a grant, provided in 2024 for fundamental and applied research in scientific and educational institutions, enterprises, and organizations of the real sector of the Republic of Tatarstan's economy. Project No. 23/2024-ФИП ("Development of Mathematical Literacy in Bilingual Schoolchildren Using Machine Learning and Artificial Intelligence Methods").

For citation

Lukoyanova M. A., Danilov A. V., Zaripova R. R., Salekhova L. L., Batrova N. I. Research on the potential of generative artificial intelligence for providing expert-level evaluative feedback in open-ended mathematical problems assessment. *Science for Education Today*, 2025, vol. 15 (6), pp. 151–174. DOI: http://dx.doi.org/10.15293/2658-6762.2506.07

Corresponding Author: Nailya I. Batrova, nibatrova@gmail.com

Conclusions. *The research concludes that generative language model holds significant potential for transforming assessment practice of open-ended mathematical problems. Key applications include automating and personalizing expert-level evaluative feedback, and scaling criteria-based assessment. Feedback quality is enhanced by optimizing assessment prompts, implementing multi-agent verification, and introducing selective assessment.*

Keywords

Evaluative feedback; Generative language model; Criteria-based assessment; Prompt engineering techniques; Open-ended problems; Mathematical literacy.

REFERENCES

1. Crompton H., Burke D. Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education*, 2023, vol. 20, pp. 1-22. DOI: https://doi.org/10.1186/s41239-023-00392-8
2. Pospelova E. A., Ototsky P. L., Gorlacheva E. N., Faizullin R. V. Generative artificial intelligence in education: Analysis trends and prospects. *Vocational Education and Labour Market*, 2024, vol. 12 (3), pp. 6-21. (In Russian) URL: https://www.elibrary.ru/item.asp?id=69176655 DOI: https://doi.org/10.52944/PORT.2024.58.3.001
3. Chekalina T. A. AI-didactics: A new trend or evolution of the learning process? *Vestnik of Minin University*, 2025, vol. 13 (2), pp. 5. (In Russian) URL: https://elibrary.ru/item.asp?id=82539976 DOI: https://doi.org/10.26795/2307-1281-2025-13-2-5
4. Alotaibi N. S., Alshehri A. H. Prospers and obstacles in using artificial intelligence in Saudi Arabia higher education institutions - The potential of AI-based learning outcomes. *Sustainability*, 2023, vol. 15 (13), pp. 10723. DOI: https://doi.org/10.3390/su151310723
5. Awidi I. T. Comparing expert tutor evaluation of reflective essays with marking by generative artificial intelligence (AI) tool. *Computers and Education: Artificial Intelligence*, 2024, vol. 6, pp. 100226. DOI: https://doi.org/10.1016/j.caeai.2024.100226
6. Kinder A., Briese F. J., Jacobs M., Dern N., Glodny N., Jacobs S., Leßmann S. Effects of adaptive feedback generated by a large language model: A case study in teacher education. *Computers and Education: Artificial Intelligence*, 2025, vol. 8, pp. 100349. DOI: https://doi.org/10.1016/j.caeai.2024.100349
7. Bearman M., Tai J., Dawson P., Boud D., Ajjawi R. Developing evaluative judgement for a time of generative artificial intelligence. *Assessment & Evaluation in Higher Education*, 2024, vol. 49 (6), pp. 893-905. DOI: https://doi.org/10.1080/02602938.2024.2335321
8. Chiang C.-H., Lee H.-y. Can large language models be an alternative to human evaluations? *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023, vol. 1, pp. 15607-15631. DOI: https://doi.org/10.18653/v1/2023.acl-long.870
9. Meyer J., Jansen T., Schiller R., Liebenow W., Steinbach M., Horbach A., Fleckenstein J. Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, 2024, vol. 6, pp. 100199. DOI: https://doi.org/10.1016/j.caeai.2023.100199
10. Pak L. E., Kryukova A. A. Capabilities of artificial intelligence programs in teaching a foreign language. *The Territory of New Opportunities: The Herald of Vladivostok State University*, 2024, vol. 16 (2), pp. 81-95. (In Russian) URL: https://elibrary.ru/item.asp?id=67900721
11. Hahn M. G., Navarro S. M. B., De La Valentín L., Burgos D. A systematic review of the effects of automatic scoring and automatic feedback in educational settings. *Institute of Electrical and*

Electronics Engineers Access, 2021, vol. 9, pp. 108190-108198. DOI: https://doi.org/10.1109/ACCESS.2021.3100890

12. Bogolepova S. V., Zharkova M. G. Researching the potential of generative language models for essay evaluation and feedback provision. *Domestic and Foreign Pedagogy*, 2024, vol. 1 (5), pp. 123-137. (In Russian) URL: https://elibrary.ru/item.asp?id=73431773
13. Zeevy-Solovey O. Comparing peer, ChatGPT and teacher corrective feedback in EFL writing: Students' perceptions and preferences. *Technology in Language Teaching & Learning*, 2024, vol. 6 (3), pp. 1482. DOI: https://doi.org/10.29140/tltl.v6n3.1482
14. Kincl T., Gunina D., Novák M., Pospíšil J. Comparing human and ai-based essay evaluation in the Czech higher education: Challenges and limitations. *Trendy v Podnikání - Business Trends*, 2024, vol. 14 (2), pp. 25-34. DOI: https://doi.org/10.24132/jbt.2024.14.2.25_34
15. Núñez-Peña M. I., Bono R., Suárez-Pellicioni M. Feedback on students' performance: A possible way of reducing the negative effect of math anxiety in higher education. *International Journal of Educational Research*, 2015, vol. 70, pp. 80-87. DOI: https://doi.org/10.1016/j.ijer.2015.02.005
16. Fyfe E. R., Brown S. A. Feedback influences children's reasoning about math equivalence: A meta-analytic review. *Thinking & Reasoning*, 2017, vol. 24 (2), pp. 157-178. DOI: https://doi.org/10.1080/13546783.2017.1359208
17. Kouzminov Y., Kruchinskaia E. The Evaluation of GenAI Capabilities to Implement Professional Tasks. *Foresight and STI Governance*, 2024, vol. 18 (4), pp. 67-76. DOI: https://doi.org/10.17323/2500-2597.2024.4.67.76
18. Schorcht S., Buchholtz N., Baumanns L. Prompt the problem – investigating the mathematics educational quality of AI-supported problem solving by comparing prompt techniques. *Frontier Education*, 2024, vol. 9, pp. 1-15. DOI: https://doi.org/10.3389/feduc.2024.1386075
19. Qian Y. Prompt engineering in education: A systematic review of approaches and educational applications. *Journal of Educational Computing Research*, 2025, vol. 63 (7–8), pp. 1782-1818. DOI: https://doi.org/10.1177/07356331251365189
20. Lee G. G., Latif E., Wu X., Liu N., Zhai X. Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, 2024, vol. 6, pp. 100213. DOI: https://doi.org/10.1016/j.caeai.2024.100213
21. Albakkosh I. Using Fleiss' kappa coefficient to measure the intra and inter- rater reliability of three AI software programs in the assessment of EFL learners' story writing. *International Journal of Educational Sciences and Arts*, 2024, vol. 3 (1), pp. 69-96. DOI: https://doi.org/10.59992/IJESA.2024.v3n1p4

Submitted: 19 August 2025 Accepted: 11 November 2025 Published: 31 December 2025

The authors' stated contribution:

Marina A. Lukoyanova
Contribution of the co-author: literature review, development of the AI prompting strategy, conducting experiments with the language model and analysis of results, structuring of the educational content, preparation of the final manuscript for submission.

Andrey V. Danilov
Contribution of the co-author: organization of the study, design of the study.

Rinata R. Zaripova
Contribution of the co-author: data processing, statistical analysis, conducting an expert assessment.

Leila L. Salekhova
Contribution of the co-author: literature review, coordination of the expert evaluation procedure.

Nailya I. Batrova
Contribution of the co-author: concept of the study, development of the AI prompting strategy, conducting experiments with the language model, interpretation of the results, preparation of the final manuscript for submission.

All authors reviewed the results of the work and approved the final version of the manuscript.

Information about competitive interests:

The authors declare no apparent or potential conflicts of interest in connection with the publication of this article

Information about the Authors

Marina Alexandrovna Lukoyanova
Candidate of Sciences (Education), Associate Professor,
Department of Bilingual and Digital Education,
Kazan Federal University,
Kremlyvskaya st. 18, 420021 Kazan, Russian Federation.
ORCID ID: https://orcid.org/0000-0002-5630-0876
E-mail: marina-lkn@yandex.ru

Andrew Vladimirovich Danilov
Candidate of Sciences (Education), Associate Professor,
Department of Bilingual and Digital Education,
Kazan Federal University,
Kremlyvskaya st. 18, 420021 Kazan, Russian Federation.
ORCID ID: https://orcid.org/0000-0002-2358-1157
E-mail: tukai@yandex.ru

Rinata Raisovna Zaripova
Candidate of Sciences (Education), Associate Professor,
Department of Bilingual and Digital Education,
Kazan Federal University,
Kremlyvskaya st. 18, 420021 Kazan, Russian Federation.
ORCID ID: https://orcid.org/0000-0003-4514-5513
E-mail: rinata-z@yandex.ru

Leila Leonardovna Salekhova
Doctor of Sciences (Education), Professor,
Department of Bilingual and Digital Education,
Kazan Federal University,
Kremlyvskaya st. 18, 420021 Kazan, Russian Federation.
ORCID ID: https://orcid.org/0000-0002-2358-1157
E-mail: salekhova2009@gmail.com

Nailya Ildusovna Batrova
Candidate of Sciences (Education), Associate Professor,
Department of Bilingual and Digital Education,
Kazan Federal University,
Kremlyvskaya st. 18, 420021 Kazan, Russian Federation.
ORCID ID: https://orcid.org/0000-0002-1945-3507
E-mail: nibatrova@gmail.com