Анализ согласованности экспертных и нейросетевых оценок при проверке заданий открытого типа

Ч. Б. Миннегалиева

Институт вычислительной математики и информационных технологий,
Казанский федеральный университет

mchulpan@gmail.com

Аннотация. **В работе рассматриваются результаты тестирования знаний при помощи заданий открытого типа. Ответы обучающихся были сформулированы в виде текста длиной до 300 символов. Проанализирована согласованность экспертных оценок с полученными при помощи модели ИИ и определением косинусного сходства. Выявлено влияние условий проведения проверки знаний, формулировок вопросов на согласованность оценок.**

Ключевые слова: большая языковая модель; векторизация текста; косинусное сходство; модель ИИ; контроль знаний; открытый вопрос; экспертная оценка

I. Введение

Оценка является компонентом процесса обучения, оказывает на него значительное влияние. Современные взгляды на цели и содержание образования меняют точку зрения и на оценивание: значительно расширяются его функции и внедряются новые инструменты [1]. Сама система оценивания является достаточно проблемной и сложной областью. На объективность оценок влияет увеличение числа обучающихся, использующих при прохождении процедур проверки знаний вспомогательные средства. Это обусловлено расширением спектра доступных по стоимости технических устройств передачи данных, открытостью образовательного процесса, а также избирательным снижением личностной готовности обучающихся к обучению [2]. Сложность оценивания также связана с тем, что проверка знаний является трудоёмким процессом. Для того, чтобы получить объективные результаты, необходима большая работа экспертов-преподавателей как при составлении, так и при оценивании заданий. Поэтому в связи с ростом достижений в области обработки естественного языка и искусственного интеллекта расширяется исследование проблемы автоматизации систем оценивания знаний [3]. Изучается правильное внедрение генеративного искусственного интеллекта в оценивание. Выявлено, что использование искусственного интеллекта для получения обратной связи помогает в определении формирующей оценки [4]. Исследовалась согласованность между экспертными оценками и оценками на основе искусственного интеллекта (ИИ) планов занятий. Результаты показали, что ИИ, особенно при использовании структурированных подсказок, может предоставлять надежные и последовательные оценки, которые тесно соответствуют экспертным заключениям. Также отмечено, что системы искусственного интеллекта должны скорее дополнять, а не заменять экспертное мнение [5]. Использование возможностей искусственного интеллекта может помочь в комплексной оценке программ высшего образования за счет автоматизированного анализа больших объемов данных в эффективные сроки [6].

Предлагаются разные подходы для решения проблемы согласованности, когда знания или качество обучения оцениваются несколькими экспертами-преподавателями. На оценку могут повлиять личные предпочтения проверяющего, необходимость уложиться в сжатые сроки, загруженность преподавателя. В работе [7] представлены методы корректировки проблемы несогласия экспертной группы.

Экспертное оценивание применяется в различных областях, например, в медицине, в управлении персоналом, в управлении проектами. Для измерения степени согласия между экспертами обычно используются Каппа Коэна и Каппа Флейса [8, 9].

В данной работе приведены результаты анализа согласованности оценок, поставленных экспертами и полученных с использованием возможностей современных технологий. Проверялись ответы обучающихся, сформулированные в свободной форме.

II. Данные и методы

A. Данные

Согласно учебному плану, студенты, обучающиеся по направлению «Информационные системы и технологии», во 2 семестре 3 курса изучают основы компьютерной графики и мультимедиа технологий. В феврале 2026 года в ходе входного контроля 57 студентов ответили на 4 вопроса, при помощи которых проверялись знания ряда основных понятий. 2 вопроса были одинаковыми для всех, 2 вопроса были заменены для 29 из 57 отвечающих в целях повышения объективности контроля, т.к. у студентов разных групп опрос проводился в разное время. Ответы были собраны через Яндекс.Формы. Студенты отвечали в аудиториях при преподавателе, были ознакомлены с требованием недопущения использования справочной литературы, поисковых систем, возможностей генеративного искусственного интеллекта, других вспомогательных инструментов. На ответы было отведено 20 минут, большинство студентов закончили работу раньше. Обучающиеся отвечали на вопросы своими словами, длина ответа ограничивалась 300 символами. Данное условие было введено в связи с тем, что в предыдущих работах было выявлено, что длина ответа может влиять на точность автоматизированного оценивания [10].

Например, надо было объяснить своими словами, принцип технологии MIDI, и почему MIDI-файл музыкального произведения будет меньше, чем её WAV-запись. Были получены верные ответы: «насколько знаю, MIDI это не аудиозапись, а запись индексов проигрываемых нот/клавиш/струн определенных инструментов во времени», «WAV-файл хранит каждую звуковую волну, сам звук, его можно прослушать, а MIDI-файл – это как будто файл инструкция, для музыкальных программ (он зашифрован). Он намного легче чем WAV, т.к. WAV хранит в себе больше волн и звуков (по сравнению с ним MIDI это набор чертежей)», «MIDI это протокол, который передаёт не сам звук, а инструкцию о том, какую ноту сыграть, на каком инструменте, с каком громкостью и длительностью. MIDI-файл записывает только ноты и параметры, поэтому они в сотни раз меньше особенно для сложных произведений, таких как симфонии», «MIDI не содержит звук, а только команды для инструментов» и другие. Примеры неверных ответов: «Так как MIDI хранит файл в плохом качестве», «Потому что WAV-запись предназначена для максимального качества звука и обладает возможностью дальнейшего редактирования звука», «не знаю», «Потому что MIDI хранит запись в более плохом качестве, чем WAV». В работе ответы представлены в том виде, в каком получены от студентов, без исправления речевых, стилистических и других ошибок.

Полученные баллы добавлялись как дополнительные к баллам за семестр. То есть студенты были заинтересованы ответить верно, в то же время неверные ответы никак не влияли на дальнейшие оценки.

Также в работе проанализированы ответы, полученные и изученные ранее [11]. В них требовалось дать описание одного понятия или процесса. В данном случае ответы были короче, большинство из них сформулированы в виде одного предложения.

B. Методы

Первоначально ответы были оценены чатом DeepSeek и при помощи косинусного сходства с использованием языковых моделей [12, 13]. При обращении к DeepSeek был использован, например, промпт: «В файле ответы студентов на вопрос “Объясните разницу между понятиями «контейнер» и «кодек» для видеофайлов”. Оцени каждый из них отдельно. Если верно, поставь 1 балл, если нет – 0 баллов. Ответ выведи в виде такой же таблицы, добавь справа столбец с оценкой. Если встретятся ответы, сгенерированные при помощи ИИ, отметь их». При одновременном оценивании множества ответов модель DeepSeek может повышать или занижать оценку в зависимости от общего уровня ответов, поэтому в промпт было включено выражение о необходимости оценивать каждый ответ отдельно. В дальнейшем планируется изучить результаты оценивания отдельными запросами для каждого ответа.

Для определения косинусного сходства рассматривались вектор ответа студента и вектор шаблонного ответа, также вектор, являющийся средним векторов всех верных ответов, полученных от студентов. Все вопросы были открытыми, то есть обучающиеся могли выразить мысли своими словами. Поэтому рассмотрение среднего вектора позволяет учесть всю вариативность формулировок верных ответов [14].

Далее преподаватель проверил те ответы, результаты оценивания по которым различались. Студентам была предоставлена возможность ознакомиться с оценками и получить по ним пояснения. После этого преподаватель проанализировал результаты и определил все окончательные оценки. Для определения согласованности экспертных оценок и оценок, полученных от DeepSeek, были вычислены коэффициент Каппы Коэна, коэффициент корреляции Мэттьюса, оценка Брайера [15]. Согласованность всех трех оценок измерялась при помощи Каппы Флейса.

III. Результаты

В табл. 1 приведены количество ответов, полученных от студентов, количество верных и неверных ответов (оценка эксперта и модели DeepSeek). Рисунок 1 показывает количество ответов по каждому вопросу, оцененных как верные и неверные экспертом.

Как видно из таблицы и по рисунку, только в наборах ответов на вопросы 3 и 4 не наблюдается явного дисбаланса. Среди ответов на вопросы 1 и 2 преобладают верные, на последние 2 вопроса – неверные. Поэтому планируется повторный анализ ответов после добавления результатов новых тестирований знаний.

В табл. 2 показаны вычисленные коэффициенты, показывающие согласованность экспертных и нейросетевых оценок. Для первых двух коэффициентов наилучшим значением является 1, значение 0 соответствует случайному совпадению. Для оценки Брайера наилучшее значение равно 0.

Например, по Каппе Коэна для вопроса 1 наблюдается высокое согласие оценок, по другим вопросам – полное согласие. Отвечая на первый вопрос, студенты должны были объяснить понятие интерактивности.

ТАБЛИЦА I.

Номер вопроса	Всего ответов	Оценены как верные (эксперт)	Оценены как верные (модель DeepSeek)
1	57	50	48
2	57	56	56
3	28	15	15
4	28	18	17
5	29	8	8
6	29	5	5

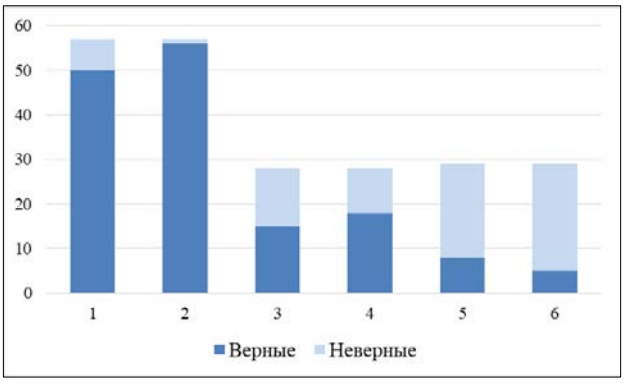

Рис. 1. Количество ответов по вопросам

ТАБЛИЦА II.

Номер вопроса	Каппа Коэна	Коэффициент корреляции Мэттьюса	Оценка Брайера
1	0.71	0.72	0.07
2	1	1	0
3	1	1	0
4	0.92	0.93	0.04
5	1	1	0
6	1	1	0

Вопрос несложный, но описать его можно разными формулировками. 4 ответа из 57 чатом DeepSeek были выделены как сомнительные, они похожи на формулировку, которую дают поисковая система, системы генеративного ИИ. Возможно, обучающиеся смогли воспользоваться дополнительными источниками информации, несмотря на то, что преподаватель был в аудитории. Преподавателем были засчитаны верными ответы, проверяя которые он посчитал, что студент понял основную мысль. Например, достаточно длинный ответ: «Взаимодействие между звуком и картинкой. То есть скажем при презентации в неё можно добавить кликабельные кнопки при нажатии на которые будет производиться звук, или выскакивать картинка, или ещё как либо действие. Должны производиться какие-либо действия в презентации». Он сформулирован не очень четко, но демонстрирует понимание обучающимся сути понятия. Преподаватель поставил 1 балл, чат DeepSeek – 0 баллов.

Чат DeepSeek оценивал как верные более точные формулировки. В то же время, моделью был оценен как верный ответ «разнообразные идеи, использование множества звуковых эффектов», который далек от описания понятия интерактивности. Разработка критериев оценивания и детальных требований к формулировке ответа повысит точность оценивания. Например, можно определить: 1 балл – верно и полностью раскрыт термин, есть пояснения про взаимодействие, реакцию системы на действия пользователя, могут быть допущены незначительные неточности; 0 баллов – неверное определение, упущено пояснение про взаимодействие, реакцию системы на действия пользователя и т.д.

Продумывание детализированного промпта можно выполнить при помощи систем генеративного искусственного интеллекта, но на данном этапе развития технологий преподаватель должен будет их перепроверить.

В случае, если разработка критериев не планируется, при автоматизированном оценивании ответов необходимо предусмотреть дополнительный этап проверки. В качестве него может быть использовано определение косинусного сходства. Были рассмотрены вопросы 3 и 4 (количество ответов приведено в таблице выше). В табл. 3 приведены вычисленные значения каппы Флейса, показывающего степень согласия между оценками, поставленными экспертом, чатом DeepSeek и определенными через косинусное сходство. Как видно из таблицы, наблюдается хорошее и отличное согласие. Рис. 2 и 3 иллюстрируют согласованность оценок с экспертными по вопросам 3 и 4.

Дополнительно были рассмотрены ответы, полученные при опросах, проведенных ранее.

ТАБЛИЦА III.

Номер вопроса	Всего ответов	Каппа Флейса
3	28	0.8068
4	28	0.7271

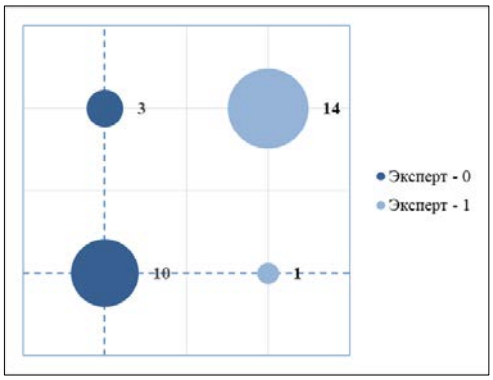

Рис. 2. Согласованность оценок по вопросу 3

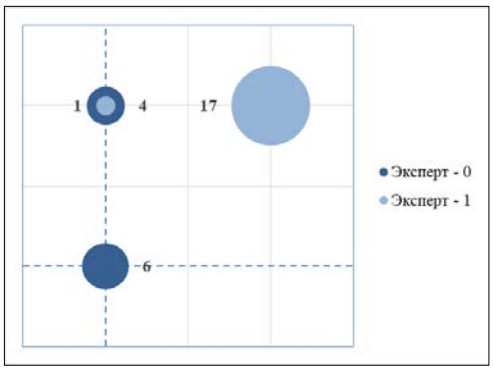

Рис. 3. Согласованность оценок по вопросу 4

На вопрос «Что называют картой высот?» было получено 46 ответов, значение каппы Флейса оказалось равным 0.4933, хорошее согласие, но меньше, чем предыдущие значения. Вариант верного ответа: «текстура в оттенках серого, яркость каждого пиксела показывает высоту». Например, значение косинуса между вектором ответа «текстура в серых оттенках» и вектором верного ответа получилось равным 0.658, так как в ответах часть слов повторяется. На самом деле ответ неполный, поэтому преподавателем был оценен как неверный.

Оценивание с использованием косинусного сходства показало хорошие результаты при проверке ответов, когда студент должен был дать короткое определение явлению или процессу и когда ответ содержал специальные термины. При более длинных ответах чаще наблюдается некорректное оценивание. Например, когда требовалось описать разницу между глобальным и локальным освещением в трехмерной графике, получены ответы «Если не путаю, глобальное освещение – свет на всей карте, на всех объектах, а локальное – подсветка определенного объекта, например того, над которым ведутся работы» (неверный) и «Глобальное освещение учитывает все источники света и их отражения от всех поверхностей. Локальное освещение учитывает только прямой свет от источника до объекта» (верный). Данные ответы близки по значению косинуса (использовалась модель sentence-transformers/distiluse-base-multilingual-cased-v2), часть слов в ответах повторяют слова вопроса.

IV. Обсуждение

Одним из важных условий корректного использования современных технологий в контроле знаний является четкая формулировка заданий. Вопрос должен быть понят обучающимися однозначно, сформулирован коротко, при этом не допускать двусмысленности и в ходе традиционного контроля знаний. При автоматизации процесса проверки требования к формулировке не ниже, т.к. в данном случае у обучающегося нет возможности уточнить задание у преподавателя.

Для текущего контроля возможно применение алгоритма, когда первоначальные оценки определяются при помощи систем генеративного искусственного интеллекта. Дополнительным этапом может быть определение косинусного сходства. Далее ответы, по которым оценки различаются, проверяет преподаватель, что существенно сократит время проверки. Обсуждение со студентами оценок, определенных моделью ИИ и при помощи косинусного сходства, повышает интерес к изучаемой дисциплине, учит критически оценивать возможности инструментов искусственного интеллекта. Для итогового контроля перепроверка всех ответов преподавателем на данном этапе развития технологий обязательна. Но анализ оценок, полученных при помощи моделей, повысит объективность контроля.

Использование детализированных промптов, указание условий, в каком случае выставляется тот или иной балл, повышает точность оценивания при помощи моделей искусственного интеллекта. Также необходимо каждый ответ оценивать отдельно.

Оценивание при помощи косинусного сходства точнее, когда получен короткий ответ и он содержит специальные термины. Если проводится тестирование знаний по недавно пройденным темам, студенты чаще употребляют слова и формулировки из материала лекций. В этом случае определение значения косинуса между векторами ответов также даёт результаты лучше, чем при проверке остаточных знаний и при входном контроле. Оценивание только при помощи косинусного сходства не может быть точным, но этот метод будет полезен как дополнительный этап проверки. Его преимущество в том, что преподаватель знает, по каким принципам работает метод.

В работе анализированы ответы, полученные от студентов бакалавриата по основам компьютерной графики и мультимедиа технологий. Результаты оценивания знаний обучающихся образовательных учреждений других уровней и по другим областям необходимо изучить отдельно.

V. Заключение

Развитие языковых моделей, инструментов искусственного интеллекта открывает новые возможности для оптимизации процесса проверки знаний. Современные технологии и инструменты могут помочь как при разработке многовариантных заданий, так и при оценивании ответов, сформулированных в свободной форме.

В работе проанализированы результаты оценивания знаний моделью DeepSeek и при помощи косинусного сходства с использованием языковой модели на основе трансформеров. При корректной формулировке вопросов согласованность полученных оценок с экспертными высокая. Данные методы можно применять в текущем контроле знаний.

При проведении промежуточной и итоговой аттестации рассмотренные подходы могут дополнять экспертные оценки. Это повысит объективность контроля, так как позволит исправить неточности оценивания, вызванные субъективным мнением эксперта.

Список литературы

[1] Шмигирилова И.Б., Рванова А.С., Григоренко О.В. Оценивание в образовании: современные тенденции, проблемы и противоречия (обзор научных публикаций) // Образование и наука. 2021. Т. 23. № 6. С. 43-83.

[2] Барашкова С.А., Березнева Е.Ю., Авдеев Д.Б. Повышение объективности оценивания результатов обучения // Наука о человеке: гуманитарные исследования. 2024. Т.18. № 2. С. 95-105.

[3] Ayaan A., Ng K.Y. Automated grading using natural language processing and semantic analysis // MethodsX. Vol.14, 103395, 2025.

[4] Louatouate H., Zeriouh M. Formative Assessment Using a Personalized Bot Developed and Equipped with Educator-Specific Data: System and Assessment Feedback Quality // In: Lahby M. (eds) Innovative Educational Assessment with Generative AI: Opportunities, Challenges, and Practical Case Studies. Information Systems Engineering and Management. 2026. Vol 70. Springer, Cham. pp. 153-170.

[5] Coşkun T.K., Altan E.B. Comparative analysis of AI and expert evaluations in engineering design pedagogy // PLOS ONE. 2025. Vol. 20(9). e0332715.

[6] Davlatova M., Shalom K., Kobtseva A. Designing an AI Expert of Educational Programmes in Higher Education // 2025 5th International Conference on Artificial Intelligence and Education (ICAIE), Suzhou, China, 2025. pp. 11-14.

[7] Liu X. -q., Chen X. College Teaching Quality Assessment Based on Group Consensus // 2014 International Conference on Virtual Reality and Visualization, Shenyang, China. 2014. pp. 12-17.

[8] Warrens M.J. Kappa coefficients for dichotomous-nominal classifications // Advances in Data Analysis and Classification, 2021. Vol. 15. pp. 193–208.

[9] 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 (101). С. 123–137.

[10] Minnegalieva C., Ziyatdinova S. Automated Analysis of Short Answers Using Text Vectorization // In Proceedings of 2025 28th International Conference on Soft Computing and Measurements, SCM 2025. Saint Petersburg, Russian Federation, 2025. pp.182-185.

[11] Minnegalieva C.B, Kashapov I.I, Morozova O.D, Automated Grading of Students' Short Answers Using Language Models // Automatic Documentation and Mathematical Linguistics. 2024. Vol.58, Is.SUPPL3. pp.109-S114.

[12] DeepSeek (Chat model), V3.2. URL: https://chat.deepseek.com/ (Дата обращения: 02.03.2026).

[13] Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) Hong Kong, China Association for Computational Linguistics. 2019. pp. 3982–3992

[14] Minnegalieva C., Using Text Vectorization to Analyze Students' Answers Formulated in Different Languages // Proceedings - 4th International Conference on Technological Advancements in Computational Sciences, ICTACS 2024. Tashkent, 2024. pp.1262-1266.

[15] Chicco D., Warrens M. J., Jurman G. The Matthews Correlation Coefficient (MCC) is More Informative Than Cohen’s Kappa and Brier Score in Binary Classification Assessment // IEEE Access. 2021. Vol. 9. pp. 78368-78381.