УДК 004.85:373.3/.5:37.015.3
МРНТИ 28.23.37

https://doi.org/10.55452/1998-6688-2026-23-3-243-254

[1]Утебаев К.А.,
магистрант, ORCID ID: 0009-0004-3032-5049,
e-mail: utebayevkuat@gmail.com
[2]*Молдагулова А.Н.,
к.ф.-м.н., профессор, ORCID ID: 0000-0002-1596-561X,
*e-mail: a.moldagulova@satbayev.university
[2]Касенхан А.М.,
PhD, профессор, ORCID ID: 0000-0002-6355-9544,
e-mail: a.kassenkhan@satbayev.university

[1]Алматинский филиал Национального исследовательского ядерного университета «МИФИ»,
г. Алматы, Казахстан
[2]Казахский национальный исследовательский технический университет им. К.И. Сатпаева,
г. Алматы, Казахстан

ПРИМЕНЕНИЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В РАЗРАБОТКЕ АДАПТИВНОЙ СИСТЕМЫ ОБУЧЕНИЯ ДЛЯ РАЗВИТИЯ НАВЫКОВ КРИТИЧЕСКОГО МЫШЛЕНИЯ У СТУДЕНТОВ

Аннотация

В статье представлены результаты разработки и пилотной апробации адаптивной обучающей системы, предназначенной для формирования и оценивания критического мышления обучающихся с использованием элементов искусственного интеллекта. Научная новизна относительно существующих исследований по автоматизированной оценке эссе и оценке на основе больших языковых моделей (LLM) заключается в многокомпонентной оценке критического мышления, каскадном механизме ИИ-оценивания и механизме апелляции, что позволяет более точно определять слабые места студента в развитии критического мышления, обеспечивать прозрачную оценку критического мышления студента, устойчивость к сбоям системы и возможность оспаривания оценки студентом в спорных ситуациях. Система оценивает эссе и решение задачи по 6 компонентам критического мышления, а именно: анализу, интерпретации, оценке, аргументации, логической связности и рефлексии. Для каждого из компонентов были предложены формулы вычисления, которые вычисляются с помощью языковых моделей, семантических эмбеддингов преобразователей предложений и резервного TF-IDF-механизма. Апробация проведена на выборке из 20 студентов, являвшихся студентами КазНИТУ им. К.И. Сатпаева (Satbayev University). В экспертном оценивании участвовали три преподавателя в лице авторов. А экспериментальные данные включали 40 студенческих работ: 20 эссе и 20 решений задач. Каждая работа была оценена независимо тремя преподавателями, а также обучающей системой. Эталонная оценка была выбрана как усредненная оценка трех преподавателей. Корреляция автоматизированной оценки с усредненной экспертной оценкой составила r = 0,84, а среднее отклонение – 1,0 балла по десятибалльной шкале. Показанные результаты позволяют использовать систему как инструмент поддержки преподавателя, но также требуют дальнейшего исследования на большой выборке. Элементы геймификации, визуализация прогресса и механизм апелляции повышают вовлеченность обучающихся и обеспечивают баланс между автоматизированным и экспертным оцениванием.

Ключевые слова: адаптивная обучающая система, критическое мышление, искусственный интеллект, языковые модели, автоматизированное оценивание, многофакторная модель, геймификация, механизм апелляции, образовательная платформа.

Введение

В нынешних реалиях развитые навыки критического мышления являются необходимым условием для успешной адаптации личности в обществе со все большим доступом к различ-

ной информации. Это подтверждает исследование [1]. Критическое мышление позволяет отличать истину от лжи, формировать обоснованные суждения, смотреть на проблему с разных точек зрения и принимать взвешенное решение. Данные навыки востребованы не только в научной среде, но и в профессиональной деятельности [2]. Традиционные методы оценивания критического мышления, как правило, основаны на субъективной экспертизе преподавателя. Проверка работ обучающегося (эссе, решение задач) требует значительных временных затрат и неизбежно сказывается влияние человеческого фактора: усталость, личные предпочтения и неоднородность критериев оценивания приводят к неверным результатам. Также при большом количестве обучающихся преподаватель не сможет дать обратную связь, что, несомненно, скажется на образовательном процессе.

Искусственный интеллект, в частности большие языковые модели, открывают новые методы решения для этих проблем. Языковые модели достаточно хорошо анализируют текст на семантическом уровне [3]. Они позволяют выявлять структуру аргументации, оценивать логическую согласованность рассуждений и сравнивать с эталонным решением. Это позволяет создать автоматизированные системы оценивания, которые также оценивают качество суждений обучающегося.

Механизм апелляции нужен тогда, когда обучающийся не согласен с выставленной оценкой в спорных ситуациях или когда искусственный интеллект галлюцинирует. Поэтому было принято использовать гибридный подход, то есть возможность в случае несогласия обучающегося с оценкой передать его работу на рассмотрение преподавателю.

Система основана на многофакторной модели, включающей шесть критериев: анализ, интерпретация, оценка, аргументация, рефлексия и логика. Эта модель была принята как базовая модель оценивания критического мышления [4].

Для вычисления каждого критерия были разработаны формулы, которые помогают нейросетям приближенно оценивать по 6 критериям и, как выяснилось, достаточно хорошо оценивают критическое мышление обучающегося, так как есть сильная корреляция между оцениванием с помощью формул и оцениванием преподавателя.

Для повышения заинтересованности обучающегося в систему были добавлены элементы геймификации: система опыта, уровней, бейджи и обезличенных рейтингов. Также была добавлена функция получения рекомендаций (книги) на основе ответов обучающегося, чтобы улучшить критическое мышление обучающегося. Масштабирование и интеграция в существующие платформы образования в будущем обеспечивается с помощью модульного принципа построения архитектуры, а эта архитектура разделена на такие слои, как уровень интерфейса, уровень бизнес-логики и базы данных.

Научная новизна исследования состоит в многокомпонентной модели оценивания критического мышления. В существующих исследованиях автоматизированного оценивания основное внимание уделяется согласованию машинного балла и оценки эксперта, повышению точности нейросетевых моделей и генерации обратной связи по письменной работе. Но обычно это рассматривалось как отдельные проблемы. Например, Pack et al. [3] анализируют валидность и надежность больших языковых моделей при оценивании эссе, Uto [5] систематизирует архитектуры глубоких нейросетевых моделей автоматизированного оценивания, Lee et al. [6] предлагают multi-trait specialization для zero-shot оценивания эссе, Stahl et al. [7] рассматривают совместную генерацию оценки и обратной связи, а Chu et al. [8] усиливают multi-trait оценивание за счет рациональных объяснений, генерируемых LLM. В данной работе сделан акцент на построении полноценной платформы оценки и развития критического мышления, интерпретируемости критического мышления в виде вычисления формул компонентов критического мышления, модерации спорных случаев с помощью механизма апелляции и последующей индивидуальной обратной связи. Такое инженерное решение также позволяет уменьшить субъективность проверки, снизить нагрузку преподавателя и обеспечить возможность автономной работы системы при ограниченной технической инфраструктуре.

Материалы и методы

Общая архитектура системы

Обучающая система была реализована на языке программирования Python 3.11 с использованием графической библиотеки CustomTkinter для построения пользовательского интерфейса. Архитектура разделена на четыре слоя: уровень представления (Presentation Layer), уровень бизнес-логики (Application Layer), уровень доступа к данным (Persistence Layer) и внешний ИИ-сервис (External AI Service) (рисунок 1).

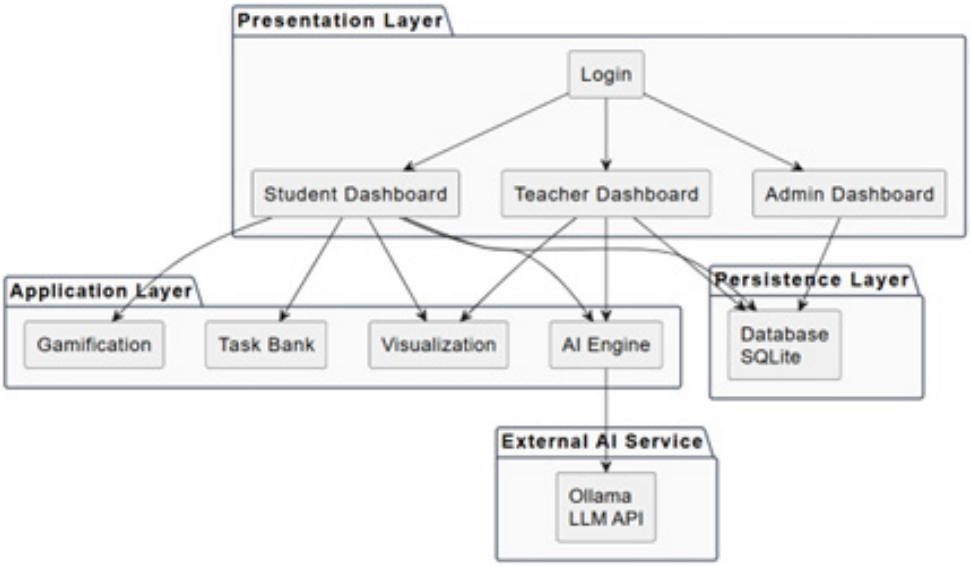

Рисунок 1 – Архитектура обучающей платформы

Диаграмма компонентов (рисунок 2) показывает взаимодействие между уровнями. Слои состоят из модуля входа в систему (login), панели студента (StudentDashboard), преподавателя (TeacherDashboard), администратора (AdminDashboard). Уровень бизнес-логики содержит модули AI Engine, Gamification, TaskBank, Recommendations и Visualization. Уровень данных обеспечивает доступ к реляционной базе данных SQLite через модуль Database.

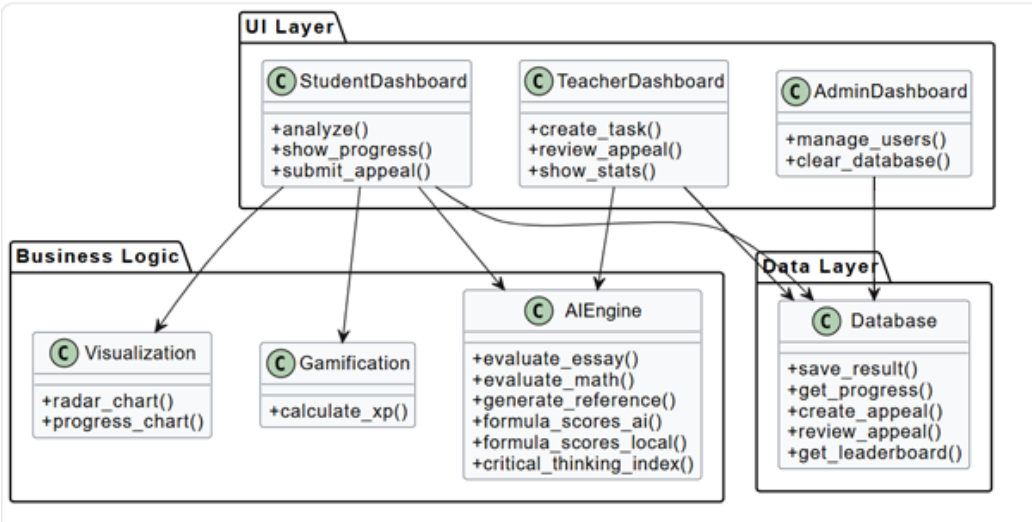

Рисунок 2 – Диаграмма компонент

Ролевая модель и управление доступом

Система имеет три роли пользователей: администратор, преподаватель, студент. Каждая роль имеет свой набор доступных ей функций, что показано на диаграмме компонент (рисунок 2). Студент может выполнять задания (эссе и математическая задача), получать оценку, просматривать прогресс и рейтинг, подавать апелляции и получать рекомендованную литературу для развития критического мышления. Преподаватель может создавать задания сам или с помощью искусственного интеллекта, просматривать успеваемость студентов своего класса, принимать апелляции и выставлять оценки по апелляциям. Администратор может создавать пользователей, изменять пароль учетных данных, удалять пользователей (кроме себя) и выборочно очищать базу данных. Задания привязаны к конкретному классу, когда преподаватель создает задание, оно автоматически привязывается к классу учителя, и обучающийся видит задание только для своего класса. Аналогично апелляции привязываются к классу преподавателя, что исключает возможность рассмотрения работ чужих студентов.

Многовекторная модель оценивания критического мышления

Основой системы является многофакторная модель оценивания, и эта модель опирается на таксономию Блума, модель критического мышления Facione и современные подходы к многофакторному оцениванию [4, 9]. Шесть критериев и их формулы представлены ниже:

1. Анализ. Здесь оценивается количество совпавших аргументов студента по отношению ко всем аргументам эталона, созданным либо учителем, либо искусственным интеллектом:

$$Score\_analysis = (Narg / Allarg) * 10, \quad (1)$$

где Score_analysis – оценка умения анализировать;

Narg – количество совпавших аргументов;

Allarg – все аргументы эталонного ответа.

Для определения семантического сходства используется косинусное сходство между предложениями студента и каждым эталонным аргументом [10]. Если косинусное сходство выше порога 0,45, то считается, что студент раскрыл аргумент.

2. Интерпретация. Оценивает, насколько хорошо студент раскрыл и объяснил тему эссе или насколько обоснованно решение задачи студента в общем.

$$Score\_interpretation = cosine\_similarity(StudAns, Ans) * 10, \quad (2)$$

где Score_interpratation – оценка умения интерпретировать;

StudAns – это полный ответ на задание студента;

Ans – общий эталонный ответ.

Для вычисления косинусного сходства используются нейросетевые эмбеддинги на основе модели all-MiniLM-L6-v2 из библиотеки sentence-transformers в качестве резервного механизма.

Оценка.

$$Score\_evalation = ((Eval\_True + Exist\_def)/2) * 10, \quad (3)$$

где Score_evaluation – оценка умения оценивать;

Eval_True – оценивает с помощью ИИ от 0.0 до 1.0, насколько убедительно звучат аргументы, нет ли неуверенности в высказываниях студента на основе маркерных слов;

Exist_def – тоже оценка от 0 до 1, насколько он обосновал аргумент, ответил на вопрос, почему он так считает. Оценка этого параметра производится с помощью ИИ.

Аргументация.

$$Score\_argumentation = ((Claim\_evidence\_ratio + Diversity + Coherence)/3) * 10, \quad (4)$$

где Score_argumentation – оценка умения аргументировать;

Claim_evidence_ratio – соотношение количества утверждений и подкрепленных доказательств, оценивается от 0.0 до 1.0;

Diversity – разнообразие типов аргументов (факты, примеры, аналогии, статистика, цитаты, логические выводы) оценивается с помощью ИИ от 0.0 до 1.0. Если типов аргументов 3, то Diversity = 3/6 = 0.5;

Coherence – ИИ оценивает от 0.0 до 1.0, насколько плавно один аргумент переходит к другому, есть ли логический мост между аргументами или они стоят изолированно.

Рефлексия. Оценивает способность к самоанализу и выявлению собственных слабых мест в аргументах.

$$Score\_reflection = min(1,0,weak\_points/base\_weak\_points) * 10, \quad (5)$$

где Score_reflection – оценка умения рефлексировать;
weak_points – количество слабых мест, найденных студентом в своей работе;
base_weak_points – ожидаемое учителем или ИИ количество слабых мест у работ студентов в среднем.

Логика. Оценивает отсутствие логических ошибок и внутренних противоречий.

$$Score\_logic = (1 - Logic\_error/5) * 10, \quad (6)$$

где Score_logic – оценка отсутствия логических ошибок и внутренних противоречий;

Logic_error – количество логических ошибок, вычисляемое ИИ.

Также нужна будет итоговая оценка критического мышления (CTI):

$$CTI = 0.2 * Score\_analysis + 0.15 * Score\_argumentation + 0.2 * Score\_evaluation + 0.2 * Score\_logic + 0.15 * Score\_interpretation + 0.1 * Score\_reflection \quad (7)$$

Трехуровневый механизм оценивания

В системе был реализован многоуровневый механизм оценивания:

Уровень 1 – AI(Ollama). Основной внешний движок оценивания. Оценивание ИИ построено на обработке prompt-запроса, включающего такую информацию, как эталонный ответ, ответ студента и формулы критериев; на выводе получается JSON-строка с числовыми оценками по каждому критерию. Система поддерживает выбор из нескольких языковых моделей (DeepSeek, GPT-OSS, Qwen3).

Уровень 2 – sentence-transformers. Резервный локальный механизм, активируемый при недоступности языковой модели из Ollama. Использует модель all-MiniLM-L6-v2 для вычисления косинусного сходства между текстами и локальный анализ маркеров.

Уровень 3 – TF-IDF. Минимальный резервный вариант, для слабых компьютеров. Использует пересечение множеств слов (коэффициент Жаккара) для приблизительной оценки семантического сходства.

Такой подход гарантирует работоспособность системы в любых условиях: от серверного развертывания до автономной работы на компьютерах без интернета.

Эталонный ответ, механизм апелляции и обратная связь

Эталонный ответ содержит ключевые аргументы, идеи, типичные ошибки и логическую цепочку рассуждений. Может быть задан двумя способами: преподавателем или искусственным интеллектом. Эталонный ответ учителя имеет приоритет над эталонным ответом искусственного интеллекта.

Студент может подать апелляцию в случае несогласия с оценкой, указав причину несогласия. Преподаватель просматривает исходную работу студента, оценку ИИ и причину несогласия студента и может изменить оценку ИИ. Система автоматически подсчитывает CTI и обновляет итоговый результат.

Также имеется механизм обратной связи, который ИИ показывает студенту и преподавателю. На этой основе преподаватель либо изменяет оценку, либо оставляет без изменений.

Элементы геймификации

Для увеличения вовлеченности студентов в обучение, в систему интегрированы элементы геймификации: система опыта (XP), уровни и бейджи. Очки опыта начисляются из CTI. Бейджи присваиваются за достижение пороговых значений: “Beginner” (CTI < 4), “Critical Thinker” (CTI ≥ 4), “Strong Reasoner” (CTI ≥ 6), “Master Thinker” (CTI ≥ 8). Обезличенный рейтинг (leaderboard) класса визуализируется в виде столбчатой диаграммы.

Визуализация и диаграмма вариантов использования

Модуль визуализации (Visualization) построен на библиотеке Matplotlib и предоставляет четыре типа графиков: радарная диаграмма оценок по критериям, график прогресса CTI по попыткам, столбчатая диаграмма рейтингов класса. Графики отображаются во встроенных окнах приложения через FigureCanvasTkAgg.

Диаграмма вариантов использования (рисунок 3) отражает полную структуру функциональных возможностей для пользователей.

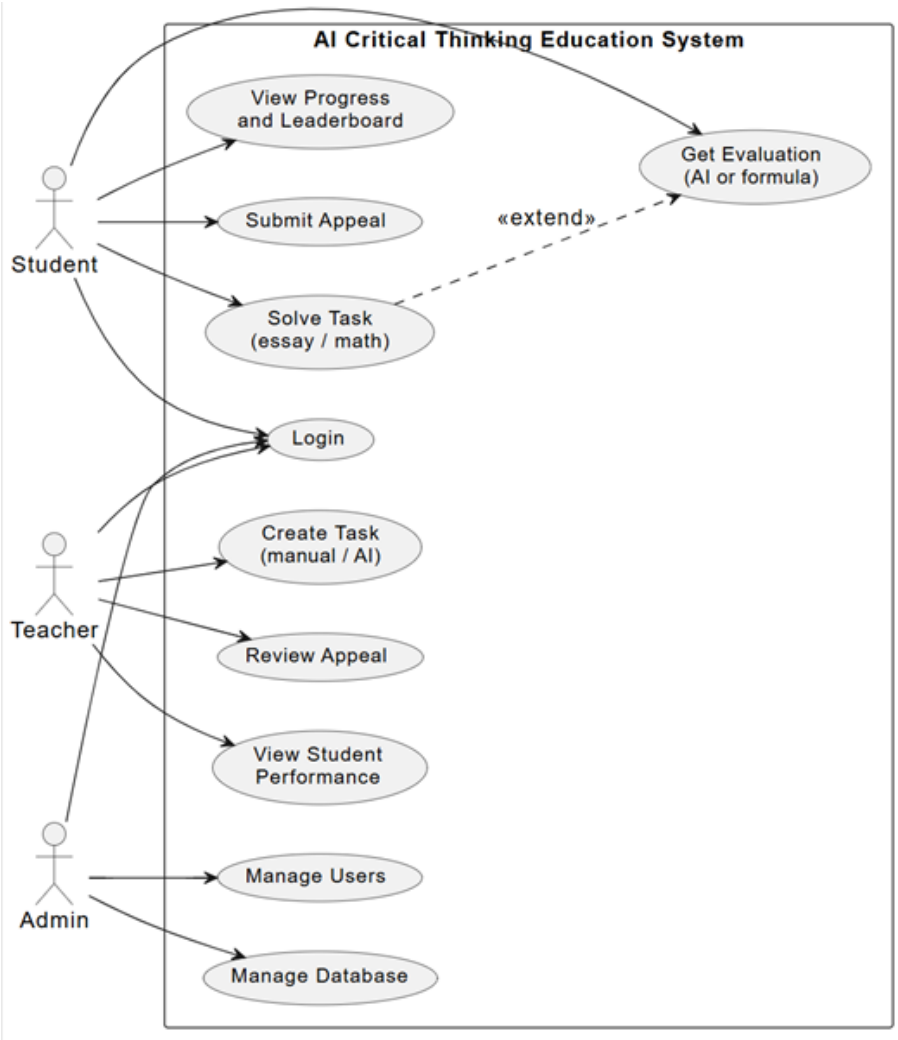

Рисунок 3 – Диаграмма вариантов использования

База данных и ER-диаграмма

Хранение данных осуществляется в реляционной базе данных SQLite (файл education.db), включающей четыре таблицы: users (по ролям и классам), results (результаты оценивания с ис-

ходными и итоговыми баллами), teacher_tasks (задания преподавателей по классам и с эталонным ответом) и appeals (хранит процесс апелляции). Использование SQLite дает автономность системы и отсутствие необходимости в отдельном сервере баз данных.

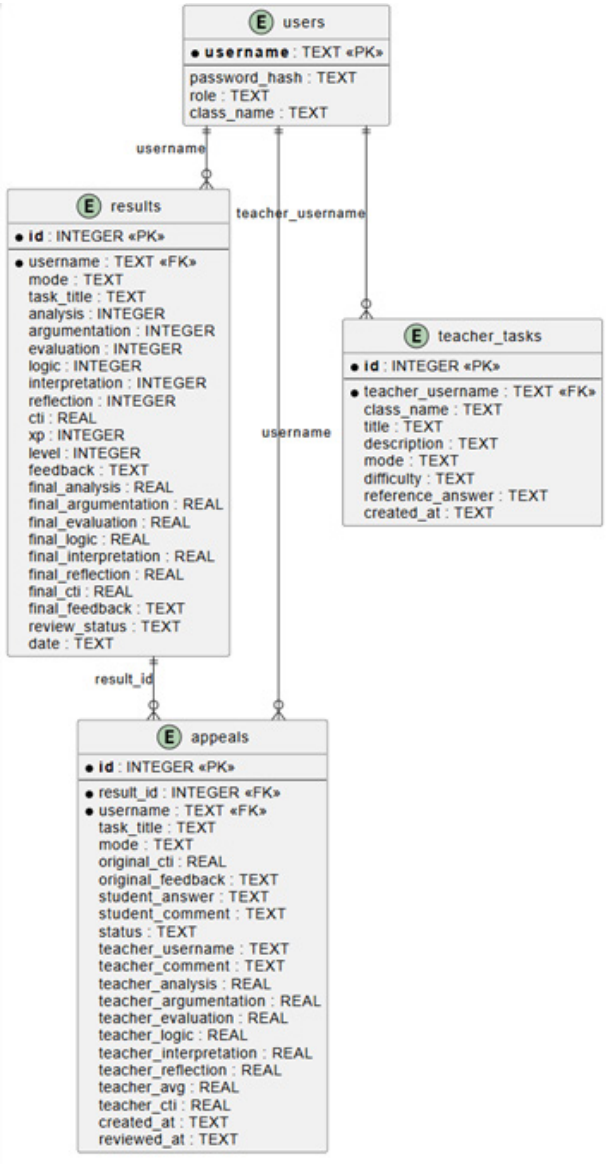

Рисунок 4 – ER-диаграмма

Методика апробации системы и характеристика выборки

Пилотная апробация адаптивной обучающей системы была проведена на выборке из 40 работ (20 эссе и 20 решений задач), выполненных 20 студентами КазНИТУ им. К.И. Сатпаева (Satbayev University). Экспертное оценивание проводилось независимо тремя преподавателями в лице авторов данной работы. Каждый преподаватель оценивал работы студентов по 6 критериям, а именно анализ, интерпретация, оценка, аргументация, рефлексия и логика. Для сопоставления с автоматизированной оценкой рассчитывалась усредненная экспертная оценка трех преподавателей. Такой подход уменьшил субъективность экспертного оценивания и повысил надежность сравнительного анализа. Этапы апробации включали: 1) загрузку заданий и эталонных ответов в систему; 2) выполнение заданий обучающимися; 3) автоматизированное оценивание работ с помощью прямой ИИ-оценки, формульной ИИ-оценки, преобразователей предложений и TF-IDF; 4) независимое экспертное оценивание тремя преподавателями;

5) расчет корреляции Пирсона и среднего абсолютного отклонения между автоматизированными оценками и усредненными экспертными оценками; 6) анализ результатов по каждому критерию критического мышления. Дополнительно фиксировались случаи потенциального несогласия с оценкой, для которых предусмотрен механизм апелляции. Следует отметить, что данная апробация имеет пилотный характер, но она достаточна, чтобы продемонстрировать работоспособность системы. Дальнейшее развитие требует увеличения выборки работ и числа образовательных организаций.

Результаты и обсуждение

Сравнение режимов оценивания

В таблице 1 представлено сравнение четырех режимов оценивания: прямая ИИ-оценка, формульная ИИ-оценка, формульная локальная оценка (sentence-transformers/TF-IDF). Подсчет корреляции и среднего отклонения рассчитывался между оценкой автоматизированной системы и усредненной оценкой преподавателей.

Таблица 1 – Сравнение режимов оценивания (корреляция с усредненной экспертной оценкой трех преподавателей)

Режим оценивания	Корреляция Пирсона (r)	Среднее отклонение от усредненной экспертной оценки
Прямая ИИ-оценка	0,84	1,0 балла
формульная ИИ-оценка	0,79	1,3 балла
Локальная sentence-transformers	0,74	1,5 балла
Локальная TF-IDF	0,38	3,4 балла

Наивысшую корреляцию с усредненной экспертной оценкой трех преподавателей показала прямая ИИ-оценка (r = 0,84), что согласуется с результатами исследований в области автоматизированного оценивания эссе [5, 11]. Формульная ИИ-оценка показала близкий результат (r = 0,79), при этом обеспечивает более высокую прозрачность, поскольку итоговый балл раскладывается на отдельные критерии критического мышления. Локальная модель преобразователей предложений продемонстрировала приемлемое качество, а TF-IDF показал более низкую согласованность, так как этот метод довольно плохо работает со смысловой структурой ответа.

Корреляция по отдельным факторам

В таблице 2 представлена корреляция формульной ИИ-оценки с усредненной экспертной оценкой трех преподавателей по каждому из шести факторов критического мышления.

Таблица 2 – Сравнение формульной ИИ-оценки с усредненной экспертной оценкой трех преподавателей

Критерий	Корреляция Пирсона (r)	Среднее отклонение от усредненной экспертной оценки
Анализ	0,80	1,2 балла
Интерпретация	0,86	0,9 балла
Оценка	0,75	1,4 балла
Аргументация	0,74	1,5 балла
Рефлексия	0,70	1,7 балла
Логика	0,83	1,1 балла
Среднее	0,78	1,3 балла
CTI	0,79	1,3 балла

Наивысшую корреляцию показали логика (r=0,83) и интерпретация (r=0,86). Оба критерия достаточно хорошо поддаются формализации, так как ИИ достаточно хорошо находит ошибки в рассуждениях, а косинусное сходство позволяет сопоставлять ответ студента с эталонным ответом. Наименьшую корреляцию показала Reflection, так как достаточно трудно формализовать этот фактор.

Сравнение методов вычисления по критериям

В таблице 3 показана корреляция с усредненной экспертной оценкой трех преподавателей по отдельным критериям и методам оценивания.

Таблица 3 – Корреляция с экспертной оценкой по критериям и методам

Критерий	Прямая ИИ- оценка	локальный Sentence-transformers	локальный TF-IDF
Анализ	0,86	0,79	0,35
Интерпретация	0,87	0,82	0,5
Оценка	0,81	0,68	0,33
Аргументация	0,77	0,7	0,31
Рефлексия	0,74	0,64	0,26
Логика	0,91	0,79	0,49
Среднее	0,83	0,74	0,37
CTI	0,84	0,74	0,38

Из таблицы 3 можно увидеть, что преобразователи предложений показывают результаты, близкие к прямой ИИ-оценке, где основную роль играет семантическое сходство текстов.

Наибольший разрыв между ними в Evaluation составляет 0,13. Это можно объяснить тем, что внешним эмбеддингам сложно оценить убедительность тех или иных аргументов. Хорошо себя показал TF-IDF в интерпретации, что неудивительно, так как косинусное сходство – одно из применений TF-IDF. В остальных факторах TF-IDF показал не лучшие результаты, так как для этого требуется понимание текстов.

Полученные результаты пилотной апробации подтверждают работоспособность разработанной системы и показывают согласованность автоматизированного оценивания с экспертным оцениванием трех преподавателей. В отличие от Lee et al. [6] и Chu et al. [8], где multi-trait scoring используется преимущественно как метод повышения точности и объяснимости автоматического оценивания эссе, в данной работе multi-trait принцип применяется к развитию критического мышления, а также к даче обратной связи, механизму апелляции и визуализации индивидуального прогресса. Многокомпонентная модель позволяет выделить конкретные аспекты критического мышления, требующие дальнейшего развития навыков студента. Трехуровневый каскадный режим работы позволяет обеспечить работоспособность даже при возникновении сбоев. Формульный подход делает оценивание более прозрачным: преподаватель и обучающийся видят не только итоговый CTI-балл, но и вклад отдельных критериев. Также исследование имеет ограничения. Апробация проводилась на выборке работ 20 студентов и проверке их работ тремя преподавателями. Это позволило доказать осуществимость и работоспособность обучающей платформы для оценивания критического мышления. Но она также не является достаточной для окончательных выводов о ее эффективности во всех образовательных процессах. Также качество формульного оценивания зависит от качества и полноты эталонного ответа. Дальнейшее развитие обучающей платформы предполагает расширение выборки, проведение апробации в разных дисциплинах и группах, а также долговременное исследование влияния обучающей системы на критическое мышление студентов и интеграцию с LMS-платформами.

Заключение

В рамках данной работы разработана адаптивная обучающая система для формирования и оценивания критического мышления у обучающихся с использованием элементов искусственного интеллекта. Система оценивания основана на многофакторной модели, включающей такие параметры, как анализ, интерпретация, оценка, аргументация, рефлексия и логика. Для каждого параметра была разработана своя формула, вычисление которой осуществляется с помощью языковых моделей, а при их недоступности разработан каскадный механизм оценивания.

Основные результаты работы заключаются в том, что реализован каскадный трехуровневый механизм оценивания, обеспечивающий работоспособность системы в различных технических условиях. При этом достигается высокая корреляция между ИИ-оценкой и экспертной оценкой преподавателя. Формульный подход обеспечивает прозрачность и объективность оценивания. Механизм апелляции обеспечивает, с одной стороны, точность оценивания в случае несогласия студента с оценкой, с другой стороны, вкупе с автоматизированной оценкой уменьшает нагрузку на преподавателя. Интеграция элементов геймификации позволит повысить вовлеченность студента в образовательный процесс. Модульная архитектура системы позволяет масштабировать систему и интегрировать ее в образовательные платформы.

Информация о финансировании

В статье представлены результаты исследований, выполненных в рамках программно-целевого финансирования на 2024–2026 гг. по проекту ИРН BR24993072 «Разработка интеллектуального образовательного ресурса на основе искусственного интеллекта для развития критического мышления студентов», финансируемого Комитетом науки Министерства науки и высшего образования Республики Казахстан.

ЛИТЕРАТУРА

1 Batdı, V., Elaldı, Ş., Özçelik, C. et al. Evaluation of the effectiveness of critical thinking training on critical thinking skills and academic achievement by using mixed-meta method. Review of Education, 12, e70001 (2024). https://doi.org/10.1002/rev3.70001

2 Almulla, M.A. Constructivism Learning Theory: A Paradigm for Students' Critical Thinking, Creativity, and Problem Solving to Affect Academic Performance in Higher Education. Sustainability, 15 (17), 12903 (2023). https://doi.org/10.1080/2331186X.2023.2172929

3 Pack, A., Barrett, A., and Escalante, J. Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. Computers and Education: Artificial Intelligence, 6, 100234 (2024). https://doi.org/10.1016/j.caeai.2024.100234

4 Facione, P.A. Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction (The Delphi Report) (Millbrae, CA: California Academic Press, 1990).

5 Uto, M. A Review of Deep-Neural Automated Essay Scoring Models. Behaviormetrika, 48 (2), 459–484 (2021). https://doi.org/10.1007/s41237-021-00142-y

6 Lee, S., Cai, Y., Meng, D., Wang, Z., and Wu, Y. Unleashing Large Language Models' Proficiency in Zero-shot Essay Scoring. arXiv preprint, arXiv:2404.04941 (2024). https://arxiv.org/abs/2404.04941

7 Stahl, M., Biermann, L., Nehring, A., and Wachsmuth, H. Exploring LLM Prompting Strategies for Joint Essay Scoring and Feedback Generation. Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024) (Mexico City: Association for Computational Linguistics, 2024), pp. 283–298. https://aclanthology.org/2024.bea-1.23/

8 Chu, S.Y., Kim, J.W., Wong, B., and Yi, M.Y. Rationale Behind Essay Scores: Enhancing S-LLM's Multi-Trait Essay Scoring with Rationale Generated by LLMs. Findings of the Association for Computational Linguistics: NAACL 2025 (Albuquerque, New Mexico: Association for Computational Linguistics, 2025), pp. 5811–5829. https://doi.org/10.18653/v1/2025.findings-naacl.322

9 Bloom, B.S. Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain (New York: David McKay Company, 1956).

10 Reimers, N., and Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2019). https://doi.org/10.18653/v1/D19-1410

11 Klebanov, B.B., and Madnani, N. Automated Essay Scoring. Synthesis Lectures on Human Language Technologies (Morgan & Claypool Publishers, 2024).

[1]Утебаев К.А.,
магистрант, ORCID ID: 0009-0004-3032-5049,
e-mail: utebayevkuat@gmail.com
[2*]Молдагулова А.Н.,
ф.-м.ғ.к., профессор, ORCID ID: 0000-0002-1596-561X,
*e-mail: a.moldagulova@satbayev.university
[2]Касенхан А.М.,
PhD, профессор, ORCID ID: 0000-0002-6355-9544,
e-mail: a.kassenkhan@satbayev.university

[1]«МИФИ» Ұлттық ядролық зерттеу университетінің Алматы филиалы,
Алматы қ., Қазақстан
[2]Қ.И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті,
Алматы қ., Қазақстан

СТУДЕНТТЕРДІҢ СЫНИ ОЙЛАУ ДАҒДЫЛАРЫН ДАМЫТУҒА АРНАЛҒАН БЕЙІМДЕЛГІШ ОҚЫТУ ЖҮЙЕСІН ӘЗІРЛЕУДЕ ЖАСАНДЫ ИНТЕЛЛЕКТТІ ҚОЛДАНУ

Аңдатпа

Мақалада жасанды интеллект элементтерін пайдалана отырып, білім алушылардың сыни ойлауын қалыптастыруға және бағалауға арналған бейімделгіш оқыту жүйесін әзірлеу және пилоттық апробациялау нәтижелері көрсетілген. Эсселерді автоматтандырылған және үлкен тілдік модельдерге (LLM) негізделген бағалау бағыттарындағы қолданыстағы зерттеулермен салыстырғанда, зерттеудің ғылыми жаңалығы сыни ойлауды көпкомпонентті бағалау, ЖИ негізіндегі бағалаудың каскадты механизмі және апелляция механизмін қолдануымен сипатталады. Бұл студенттің сыни ойлауын дамытудағы әлсіз тұстарын дәлірек анықтауға, сыни ойлауды бағалаудың ашықтығын қамтамасыз етуге, жүйенің істен шығу жағдайларына төзімділігін арттыруға және даулы жағдайларда студентке бағалау нәтижесіне шағымдану мүмкіндігін беруге мүмкіндік береді. Жүйе эсселер мен есептердің шешімдерін сыни ойлаудың алты компоненті бойынша бағалайды: талдау, интерпретация, бағалау, аргументация, логикалық байланыстылық және рефлексия. Әрбір компонент үшін есептеу формулалары ұсынылды; олар тілдік модельдер, сөйлем түрлендіргіштеріне негізделген семантикалық эмбеддингтер және резервтік TF-IDF механизмі арқылы есептеледі. Апробация К.И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университетінің (Satbayev University) 20 студентінен құралған таңдамада жүргізілді. Сараптамалық бағалауға авторлар қатарынан үш оқытушы қатысты. Эксперименттік деректер 40 студенттік жұмысты қамтыды: 20 эссе және 20 есептің шешімі. Әрбір жұмысты үш оқытушы тәуелсіз бағалады, сондай-ақ жұмыстар оқыту жүйесі арқылы бағаланды. Эталондық баға ретінде үш оқытушының орташа бағасы алынды. Автоматтандырылған бағалау мен орташа сараптамалық бағалау арасындағы корреляция r = 0,84 құрады, ал орташа ауытқу он балдық шкала бойынша 1,0 балл болды. Алынған нәтижелер жүйені оқытушыны қолдау құралы ретінде пайдалануға мүмкіндік беретінін көрсетеді, алайда оны үлкенірек таңдама негізінде әрі қарай зерттеу қажет. Ойындандыру элементтері, прогресті визуализациялау және апелляция механизмі білім алушылардың қызығушылығын арттырып, автоматтандырылған және сараптамалық бағалау арасындағы тепе-теңдікті қамтамасыз етеді.

Түйін сөздер: бейімделгіш оқыту жүйесі, сыни ойлау, жасанды интеллект, тілдік модельдер, автоматтандырылған бағалау, көпфакторлы модель, ойындандыру, апелляция механизмі, білім беру платформасы.

[1]Utebayev K.A.,
Master's student, ORCID ID: 0009-0004-3032-5049,
e-mail: utebayevkuat@gmail.com
[2]*Moldagulova A.N.,
Candidate of Physical and Mathematical Sciences, Professor,
ORCID ID: 0000-0002-1596-561X,
*e-mail: a.moldagulova@satbayev.university
[2]Kassenkhan A.M.,
PhD, Professor, ORCID ID: 0000-0002-6355-9544,
e-mail: a.kassenkhan@satbayev.university

[1]Almaty Branch of the National Research Nuclear University MEPhI, Almaty, Kazakhstan
[2]Satbayev University, Almaty, Kazakhstan

USING ARTIFICIAL INTELLIGENCE TO DEVELOP AN ADAPTIVE LEARNING SYSTEM TO DEVELOP CRITICAL THINKING SKILLS IN STUDENTS

Abstract

The article presents the results of the development and pilot testing of an adaptive learning system designed to develop and assess students' critical thinking using elements of artificial intelligence. The scientific novelty of the study, in comparison with existing research on automated essay scoring and LLM-based assessment, lies in the multi-component assessment of critical thinking, the cascade mechanism of AI-based assessment, and the appeal mechanism. These features make it possible to more accurately identify students' weak points in the development of critical thinking, ensure transparent assessment of students' critical thinking, increase the system's resilience to failures, and provide students with the opportunity to challenge the assessment in controversial situations. The system evaluates essays and problem-solving tasks according to six components of critical thinking: analysis, interpretation, evaluation, argumentation, logical coherence, and reflection. For each component, calculation formulas were proposed; these are computed using language models, semantic embeddings based on sentence-transformers, and a backup TF-IDF mechanism. The pilot testing was conducted on a sample of 20 students from Satbayev University. Three instructors, represented by the authors, participated in the expert assessment. The experimental data included 40 student works: 20 essays and 20 problem-solving tasks. Each work was independently assessed by three instructors as well as by the learning system. The reference score was determined as the average score of the three instructors. The correlation between the automated assessment and the averaged expert assessment was r = 0.84, while the mean deviation was 1.0 point on a ten-point scale. The results obtained indicate that the system can be used as a tool to support instructors; however, further research on a larger sample is required. Gamification elements, progress visualization, and the appeal mechanism increase students' engagement and ensure a balance between automated and expert assessment.

Keywords: adaptive learning system, critical thinking, artificial intelligence, language models, automated assessment, multifactor model, gamification, appeal mechanism, educational platform.

Received April 14, 2026; revised June 4, July 22, 2026; accepted August 20, 2026.