МРНТИ 14.01.85 https://doi.org/10.26577/JES872202612

А.А. Әбжәлиева*, Ж.А. Чурбанова, А.Ш. Хайдарова, А.Б. Карибаева, М.З. Абдрахманов

Национальный центр тестирования, Астана, Казахстан
*e-mail: abjalieva.aidana@mail.ru

ГЕНЕРАТИВНЫЙ ИИ ДЛЯ ЕНТ: ОЦЕНКА КАЧЕСТВА КЛОНИРОВАНИЯ ЗАДАНИЙ ПО МАТЕМАТИКЕ

В статье представлены результаты исследования, посвящённого анализу качества и достоверности тестовых заданий по математике с одним правильным ответом, созданных с применением языковых моделей искусственного интеллекта. Основной целью исследования являлась оценка того, насколько такие задания могут соответствовать требованиям национального тестирования и быть использованы в образовательной практике. В рамках исследования было проведено контролируемое клонирование 200 заданий формата ЕНТ на основе предварительно верифицированного банка тестовых материалов. После генерации задания прошли многоэтапную экспертную оценку, включающую проверку математической корректности, определение когнитивного уровня сложности, а также анализ соответствия официальным спецификациям тестирования.

Для генерации заданий были использованы современные языковые модели искусственного интеллекта, в том числе GPT-4, DeepSeek, Qwen версий 2.5 и 3, а также Claude 3. Сравнительный анализ показал, что подавляющее большинство сгенерированных заданий демонстрируют высокий уровень качества. По результатам экспертной проверки 97 % заданий были признаны пригодными для дальнейшего использования. При этом 50,5 % из них не потребовали какой-либо дополнительной доработки и могли быть использованы в исходном виде. Полученные результаты свидетельствуют о низком уровне брака и подтверждают устойчивость применённых алгоритмов генерации тестовых заданий. Таким образом, результаты исследования указывают на значительный потенциал интеграции технологий искусственного интеллекта в процессы обновления, расширения и оптимизации банка заданий ЕНТ, что может повысить эффективность разработки тестовых материалов и ускорить их актуализацию.

Ключевые слова: искусственный интеллект, генеративные языковые модели, автоматизированное клонирование заданий, ЕНТ, педагогические измерения, валидация тестов.

A.A. Abzhaliyeva*, Zh.A. Churbanova, A.Sh. Khaidarova, A.B. Karibayeva, M.Z. Abdirakhmanov

National Testing Center, Astana, Kazakhstan
*e-mail: abjalieva.aidana@mail.ru

Generative AI for UNT: assessing the quality of mathematics item cloning

The paper presents the results of a study devoted to analyzing the quality and reliability of mathematics test items with a single correct answer generated using artificial intelligence language models. The main objective of the research was to evaluate how well such items correspond to the requirements of national testing and whether they can be effectively used in educational practice. Within the framework of the study, a controlled cloning of 200 test items in the format of the Unified National Testing (UNT) was carried out based on a previously verified bank of test materials. After generation, the items underwent a multi-stage expert evaluation process, which included checking mathematical correctness, determining the cognitive level of difficulty, and analyzing compliance with official testing specifications.

Modern artificial intelligence language models were used to generate the tasks, including GPT-4, DeepSeek, Qwen versions 2.5 and 3, and Claude 3. Comparative analysis showed that the vast majority of the generated items demonstrated a high level of quality. According to the results of the expert review, 97% of the tasks were recognized as suitable for further use. Moreover, 50.5% of them did not require any additional revision and could be used in their original form. The obtained results indicate a low defect rate and confirm the stability of the applied algorithms for generating test items. Thus, the findings of the study demonstrate the significant potential of integrating artificial intelligence technologies into the

processes of updating, expanding, and optimizing the UNT item bank, which may improve the efficiency of developing test materials and accelerate their renewal.

Keywords: artificial intelligence, generative language models, automated item cloning, UNT, educational measurement, test validation.

А.А. Әбжәлиева*, Ж.А. Чурбанова , А.Ш. Хайдарова,
А.Б. Карибаева, М.З. Әбдірахманов
Ұлттық тестілеу орталығы, Астана, Қазақстан
*e-mail: abjalieva.aidana@mail.ru

ҰБТ үшін генеративті ЖИ: математикалық тапсырмаларды клондау сапасын бағалау

Мақалада жасанды интеллекттің тілдік модельдерін қолдану арқылы жасалған, бір дұрыс жауабы бар математика пәні бойынша тест тапсырмаларының сапасы мен сенімділігін талдауға арналған зерттеу нәтижелері ұсынылған. Зерттеудің негізгі мақсаты – мұндай тапсырмалардың ұлттық тестілеу талаптарына қаншалықты сәйкес келетінін және оларды білім беру тәжірибесінде қолдану мүмкіндігін бағалау болды. Зерттеу аясында алдын ала верификацияланған тест тапсырмалары банкі негізінде ЕНТ форматына сәйкес 200 тапсырманы бақылаулы түрде клондау жүргізілді. Генерациядан кейін тапсырмалар бірнеше кезеңнен тұратын сараптамалық бағалаудан өтті. Бұл бағалау математикалық дұрыстығын тексеруді, когнитивтік күрделілік деңгейін анықтауды, сондай-ақ тестілеудің ресми спецификацияларына сәйкестігін талдауды қамтыды.

Тапсырмаларды генерациялау үшін жасанды интеллекттің заманауи тілдік модельдері пайдаланылды. Олардың қатарында GPT-4, DeepSeek, Qwen 2.5 және 3 нұсқалары, сондай-ақ Claude 3 модельдері бар. Жүргізілген салыстырмалы талдау нәтижесінде генерацияланған тапсырмалардың басым көпшілігі жоғары сапа деңгейін көрсеткені анықталды. Сараптамалық тексеру қорытындысы бойынша тапсырмалардың 97%-ы әрі қарай қолдануға жарамды деп танылды. Оның ішінде 50,5%-ы ешқандай қосымша түзетуді немесе өңдеуді қажет етпей, бастапқы күйінде пайдалануға болатындығы анықталды. Алынған нәтижелер ақау деңгейінің төмен екенін көрсетіп, тест тапсырмаларын генерациялау барысында қолданылған алгоритмдердің тұрақтылығын дәлелдейді. Осылайша, зерттеу нәтижелері жасанды интеллект технологияларын ЕНТ тапсырмалар банкін жаңарту, кеңейту және оңтайландыру үдерістеріне енгізудің айтарлықтай әлеуеті бар екенін көрсетеді. Бұл өз кезегінде тест материалдарын әзірлеу тиімділігін арттыруға және оларды жаңарту үдерісін жеделдетуге мүмкіндік береді.

Түйін сөздер: жасанды интеллект, генеративті тілдік модельдер, автоматтандырылған тапсырма клондау, ҰБТ, педагогикалық өлшемдер, тест валидтілігі.

Введение

В настоящее время стандартизированные оценки широко применяются для получения надёжных и сопоставимых измерений образовательных результатов и эффективности системы образования. В частности, банки тестовых заданий являются ключевым компонентом таких оценок, поскольку позволяют формировать тесты на основе верифицированных заданий, обеспечивающих валидность и надёжность измерений (OECD, 2023). Однако рост требований к качеству оценивания и необходимость регулярного обновления банков делают традиционные подходы ресурсоёмкими и недостаточно оперативными.

В этой связи внедрение генеративных моделей искусственного интеллекта (ИИ) открывает новые возможности для модернизации разработки тестовых заданий за счёт сокращения временных и экспертных затрат. Актуальность этого подхода подчёркивается и на государственном уровне: К.К. Токаев отметил, что развитие ИИ является ключевым условием глобальной конкурентоспособности и движущей силой трансформации страны (Akorda.kz, 2025). Эта перспектива особенно важна для экзаменов с высокой ставкой, таких как Единое национальное тестирование (далее – ЕНТ), обеспечивающее доступ к высшему образованию и требующее соблюдения принципов валидности, надёжности и справедливости.

Концептуальной основой модернизации является автоматизированная генерация заданий (Automatic Item Generation, AIG), активно применяемая в масштабных оценочных системах и базирующаяся на когнитивных моделях и вычислительных алгоритмах для массового создания заданий (Gierl & Lai, 2016), AIG объединяет экспертные знания и автоматизированные

методы, снижая нагрузку традиционной разработки (Pugh, D. et al., 2022). Центральное место занимает клонирование заданий (Item Cloning) через концепцию «семейства заданий» (Item Family) – набор элементов с общей когнитивной структурой и форматом, но при этом различающихся поверхностными характеристиками (Lathrop & Cheng, 2016).

Теоретической основой клонирования является слабая теория AIG (Weak Theory), согласно которой новые задания создаются на базе верифицированного «родительского» элемента путём изменения поверхностных параметров так, чтобы не влиять на когнитивные процессы решения (Falcão et al., 2022). Эквивалентность заданий формализуется через аддитивные многоуровневые модели структуры заданий (Additive Multilevel Item Structure, AMIS), где клоны вложены в родительские шаблоны, что позволяет минимизировать необходимость индивидуальной калибровки (Lathrop & Cheng, 2016).

Международная практика внедрения ИИ в оценивание подтверждает значимость подобных подходов. Так, в Educational Testing Service (ETS, США) автоматизированная генерация и оценка ответов (системы e-Rater для письменных работ и SpeechRater для устной речи) уже стали неотъемлемой частью многих экзаменов, однако при этом строго соблюдается принцип «human-in-the-loop»: все результаты, полученные с помощью алгоритмов, проходят экспертную верификацию (Deane & Higgins, 2019). Аналогичной позиции придерживаются и британские экзаменационные советы (AQA, JCQ), которые рассматривают контент, созданный ИИ, как вспомогательный инструмент и не допускают его использования без контроля со стороны специалистов, особенно в экзаменах с высокой ставкой (high-stakes). Таким образом, исследование по интеграции ИИ в разработку заданий ЕНТ полностью соответствует мировым стандартам обеспечения надёжности и валидности педагогических измерений.

Современные исследования демонстрируют значительный потенциал больших языковых моделей (Large Language Models, LLM) для автоматической генерации образовательных заданий. В частности, показано, что современные модели способны формировать полный набор вопросов на основе учебного контекста и распределять их по уровням таксономии Блума, что позволяет приближаться по качеству к заданиям, создаваемым преподавателями (Maity et al., 2025). Исследования также подтверждают, что генеративные модели могут создавать структурно корректные экзаменационные вопросы, пригодные для использования в качестве шаблонов, хотя для обеспечения точности содержания и качества дистракторов требуется экспертная валидация (Camarata et al., 2025).

Дополнительные работы показывают, что современные генеративные модели способны достигать высокого уровня согласия с экспертной оценкой при анализе образовательных ответов, демонстрируя тем самым значительный потенциал для использования в системах педагогических измерений (Pecuchova et al., 2025). При этом интеграция LLM с подходами Retrieval-Augmented Generation (RAG) позволяет существенно повысить точность и предметную релевантность автоматизированных систем оценивания за счёт использования внешних источников знаний (Rodríguez et al., 2025).

Национальным центром тестирования инициирован проект по эмпирической оценке возможностей интеграции генеративных моделей ИИ в разработку заданий ЕНТ. Основная цель проекта – проверка способности нейросетей обеспечивать автоматизированное клонирование заданий, соответствующих требованиям валидности, надёжности и содержательного качества. Предметной областью выбрана математика в качестве ключевой дисциплины для абитуриентов STEM-направлений, с фокусом на задания с одним правильным ответом.

Цель статьи – оценить применимость генеративных нейросетей для клонирования тестовых заданий по математике в рамках ЕНТ.

Для достижения этой цели были поставлены следующие задачи:

1. Провести сравнительный анализ эффективности различных крупных языковых моделей (LLM) в процедуре клонирования заданий.

2. Организовать и провести многоуровневую экспертную валидацию сгенерированных заданий в соответствии с установленными требованиями.

Актуальность подобных исследований подтверждается и международной практикой. Так, недавнее исследование качества математических тестов, сгенерированных ChatGPT для восьмого класса по программе Иордании, показало, что 80% заданий соответствуют ожиданиям экспертов, а 85% обладают высоким уровнем содержательной валидности (Aljodeh, 2025). При этом авторы также отмечают необходимость че-

ловеческого контроля для устранения языковых и культурных несоответствий . Аналогичные выводы получены в ходе систематического анализа генерации заданий по STEM-дисциплинам с использованием GPT-3.5 и GPT-4, где применение цепочки рассуждений (chain-of-thought prompting) значительно повысило качество создаваемого контента (Yoon et al., 2025).

3. Проанализировать реакцию экспертов на созданные задания, включая этапы содержательной рецензии, выявления ошибок, оценивания когнитивного уровня и итогового принятия решения о пригодности заданий для включения в экзаменационные материалы.

Материалы и методы исследования

Дизайн исследования и материалы. Проведена эмпирическая экспертиза качества тестовых заданий, созданных методом автоматизированного клонирования с использованием ИИ. Исследование включало 200 заданий по математике формата ЕНТ с одним правильным ответом, сгенерированных контролируемым клонированием. Для генерации использовались крупные языковые модели (LLM): GPT-4, DeepSeek, Qwen 2.5/3 и Claude 3, с сравнительным анализом их эффективности. Исходным материалом служил верифицированный банк заданий ЕНТ 2024 года (139 заданий) с подтверждёнными психометрическими характеристиками.

Процедура автоматизированного клонирования и экспертной валидации. Была внедрена многоуровневая процедура клонирования с участием сотрудника Центра (далее – эксперт-аналитик), отвечающего за математическое направление в цикле Human-in-the-Loop. Процесс клонирования был реализован в разработанном прототипе системы «**AI TestGen NCT**», который представляет собой гибридную архитектуру, сочетающую следующие ключевые компоненты:

1. Векторная база данных эталонных заданий: Содержит верифицированные задания ЕНТ в виде векторных эмбеддингов для семантического поиска релевантных шаблонов клонирования.

2. Модуль RAG (Retrieval-Augmented Generation): Осуществляет извлечение релевантных эталонных заданий и их метаданных (параметры сложности, тематика, психометрические характеристики) для контекстуализации промптов.

3. Ансамбль LLM: Включает несколько языковых моделей (GPT-4, DeepSeek, Qwen, Claude 3), оптимизированных для генерации математических.

Все сгенерированные клоны проходили многоуровневую содержательную проверку (ревизию/экспертизу тестовых заданий) в соответствии с приказом Центра:

- Верификация математической корректности – проверка точности вычислений, корректности математических формул и логической целостности условия.

- Анализ дидактической адекватности – оценка соответствия когнитивного уровня, тематической направленности и отсутствия возможности решения методом подбора.

- Проверка формальных требований – контроль соответствия спецификациям ЕНТ, включая использование стандартных обозначений и терминологии.

- Принятие интегрального решения – классификация заданий на категории «Принято», «Требует доработки» или «Отклонено» на основе комплексной оценки.

Работа системы организована как многоэтапный конвейер. На первом этапе локальная языковая модель (на базе архитектуры GPT) выполняет семантическую деконструкцию эталонного задания, выделяя его инвариантную структуру: проверяемый элемент содержания, логику решения, математические взаимосвязи и типичные ошибки, которые могут служить основой для дистракторов. Этот аналитический контекст формирует строгое техническое задание (промпт) для второго этапа – параллельной изоморфной генерации. На этом этапе два независимых генеративных движка (локальная LLM и облачная модель, например Gemini 2.5 Pro) одновременно создают варианты заданий, что обеспечивает диверсификацию контента и повышает устойчивость системы к сбоям. Генерация проводилась с параметром температуры 0,65, что позволяет поддерживать баланс между воспроизводимостью и вариативностью. После генерации все клоны проходят автоматическую пост-валидацию с использованием библиотек SymPy и NumPy: числовые данные из условия извлекаются, задача решается программно, и полученный ответ сравнивается с ключом, предложенным моделью. Такая проверка позволяет отсеять задания с вычислительными ошибками ещё до передачи экспертам.

Методы анализа данных. Для оценки эффективности процедуры клонирования использовался комплекс количественных и качественных методов:

- Расчет коэффициента валидности как доли заданий, признанных пригодными для использования.

- Сравнительный анализ эффективности различных LLM в процедуре клонирования.

- Качественный анализ экспертных заключений с категоризацией типичных замечаний и выявлением системных проблем.

Данная методология позволила не только оценить потенциал автоматизированного клонирования, но и выявить критические точки, требующие экспертного вмешательства в процессе массовой генерации тестовых заданий.

Результаты и обсуждение

В ходе проекта получены научные и практические результаты, подтверждающие эффективность применения генеративных языковых моделей (LLM) для автоматизации разработки тестовых заданий по математике для ЕНТ. В частности, в работе Ш. Кадырова, Б. Абдрасилова и др. «Оценка LLM на экзамене по математике для поступления в университет в Казахстане» (2025) подробно отражены промежуточные результаты проекта.

Для проверки возможностей проведено тестирование шести моделей – Claude, DeepSeek, Gemini, Llama, Qwen и o1 – на задачах по алгебре, функциям, геометрии, неравенствам и тригонометрии. Важно отметить, что zero-shot режиме DeepSeek, Gemini, Qwen и o1 показали точность 91,2–100%, а Claude и Llama – 43,5–76,5%. Таким образом, гибридный подход повысил точность Claude на 27,4%, Llama – на 39,9%, при многоагентном уточнении Claude достигла 97,8% точности (+58,1% к zero-shot).

Эксперт-аналитик по математике отобрал 139 валидных эталонных заданий из базы в 1000 позиций, исключив задания с графическими элементами или не соответствующие дидактическим и психометрическим критериям (Спецификация теста по математике 2025).

С использованием локального прототипа «AI TestGen NCT» (Отчет о научно-исследовательской работе, 2025) на базе корпуса эталонных заданий с гибридной архитектурой, RAG-механизмами и многоагентным уточнением было сгенерировано 200 новых тестовых заданий, прошедших многоуровневую экспертную валидацию в соответствии с приказом Центра.

Проведённая экспертиза тестовых заданий по предмету «Математика» (язык обучения – русский) установила, что из 200 проверенных заданий:

- 5 заданий были исключены как не соответствующие установленным требованиям;

- 195 заданий рекомендованы для дальнейшего использования, включая задания без замечаний и задания, требующие корректировки.

В ходе анализа тестовых заданий были выявлены три основных аспекта, требующих внимания и доработки.

1. Первый аспект – формулировки, тематическая точность и уровень сложности:

- Обнаружено значительное количество некорректных формулировок и несоответствий тематике.

- В условиях некоторых заданий наблюдается дублирование формулировок, например: «Решите систему уравнений» и «Укажите решение системы уравнений».

- В геометрических задачах встречаются нелогичные формулировки.

Отмечено несоответствие уровня сложности некоторых заданий.

В ряде случаев дистракторы не являются рабочими, что позволяет определить правильный ответ методом исключения.

Кроме того, в ряде уравнений возможно нахождение корней простым подбором ответов, что снижает качество тестовых заданий с позиции тестологии.

2. Второй аспект – оформление, обозначения и стандарты записи:

- Зафиксированы случаи использования англоязычных или международных обозначений, не соответствующих принятым отечественным стандартам.

3. Третий аспект – избыточная сложность и несоответствие алгоритмов решения:

- Некоторые задания характеризуются чрезмерным уровнем сложности. Применяемые в них алгоритмы решения (например, «2 на 3» и «5 на 2») не являются универсальными, что делает задания некорректными.

В отдельных случаях получаются уравнения и системы, решения которых выходят за рамки уровня ЕНТ. Такие задания подлежат удалению.

Полученные результаты согласуются с международными наблюдениями о неоднородности качества математического вывода у LLM и необ-

ходимости применения строгих процедур оценки. Ряд исследований демонстрирует, что модели показывают нестабильность при решении математических задач: формально корректные рассуждения могут сопровождаться логическими неточностями, ошибочными преобразованиями или несоответствием уровня сложности. Так, в работах A. Lewkowycz и др. (2022) показано, что математическое рассуждение крупных языковых моделей остаётся непоследовательным даже при использовании специализированных обучающих методик.

Такая перспектива подтверждает необходимость комплексной экспертизы сгенерированных клонов, поскольку даже при высокой вычислительной точности модели могут допускать методические, формальные или содержательные отклонения, критичные для экзаменов с высокой ставкой.

Ранее проведённое ранее пилотное исследование с участием тех же языковых моделей на заданиях ЕНТ по математике (Kadyrov et al., 2025) показало, что в режиме zero-shot некоторые модели (DeepSeek, Gemini, Qwen, o1) достигают точности 91–100%, в то время как другие (Claude, Llama) демонстрируют существенно более низкие результаты. Это подтверждает, что способность моделей к решению задач коррелирует с их пригодностью для генерации новых корректных заданий. Высокая доля (97%) пригодных к использованию клонов в нашем исследовании косвенно свидетельствует о том, что модели, успешно решающие задачи, способны при правильной настройке промптов и использовании RAG-механизмов создавать валидный контент. Вместе с тем экспертный анализ выявил проблемы, связанные не только с математической корректностью, но и с лингводидактической адаптацией. Как показано в исследованиях по автоматической оценке устной речи, системы ИИ часто нечувствительны к прагматическим и культурно-специфичным аспектам, что проявляется в завышении оценок за формально правильные, но содержательно бедные ответы. Аналогично в нашем случае модель иногда генерировала неестественные для учебной литературы формулировки, использовала кальки с русского языка или устаревшую лексику, что потребовало дополнительного редактирования. Эти наблюдения подчёркивают важность разработки специализированных языковых моделей для казахского языка, подобных KazBERT и KenesBERT (Sakenov et al., 2021; Akhmetov et al., 2023), которые учитывают морфологическую сложность и дискурсивные особенности. Кроме того, с психометрической точки зрения необходимо отметить, что 47% заданий, потребовавших доработки (особенно по параметру сложности), могут при непосредственном включении в тест без предварительной калибровки повлиять на надёжность и валидность измерений. В классической теории тестов и в IRT неверная оценка трудности задания ведёт к смещению итоговых баллов, что недопустимо для экзаменов с высокой ставкой. Поэтому, как и в случае с автоматической оценкой эссе или устной речи (Bernstein et al., 2020; Loukina et al., 2022), необходима тщательная проверка на наличие систематических ошибок (bias) и обеспечение справедливости теста.

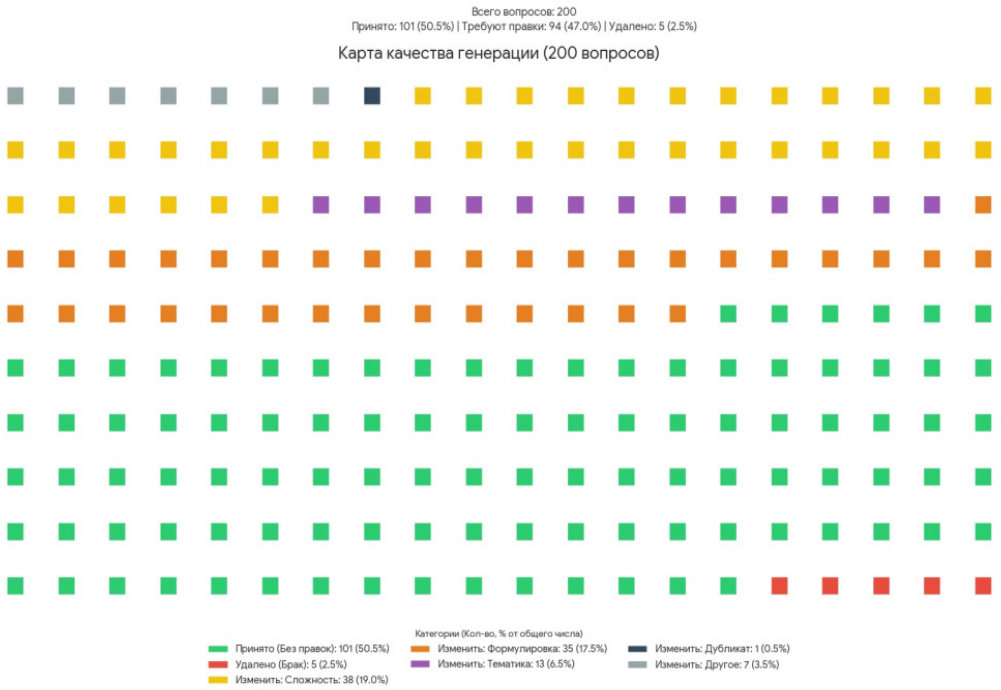
Для систематизации экспертных оценок по каждому вопросу была разработана матрица (*10x20)* качества генерации заданий, визуализирующая распределение 200 автоматически сгенерированных вопросов по категориям экспертных заключений (рисунок 1). Каждая ячейка матрицы соответствует одному вопросу, при этом цветовое кодирование используется для обозначения категорий экспертных заключений по качеству и необходимости доработки (рис.1).

Также матрица качества демонстрирует, что 50,5 % сгенерированных заданий соответствуют требованиям ЕНТ и не требуют дополнительной доработки. Наибольший объем правок требуется в области калибровки уровня сложности (19,9%) и корректировки формулировок (17,5%). Низкий процент дублирования (0,5%) свидетельствует об эффективности алгоритмов вариативности, а минимальное количество бракованных заданий (3,2%) подтверждает принципиальную пригодность технологии для генерации тестовых материалов.

Визуализация позволяет идентифицировать системные проблемы генерации: необходимость улучшения алгоритмов контроля сложности и оптимизации лингвистических компонентов модели. Полученные данные предоставляют четкие ориентиры для дальнейшего совершенствования системы автоматизированной генерации тестовых заданий.

Рисунок 1

Распределение качества автоматической генерации заданий по категориям оценки



Источник / Примечание: Данный рисунок составлен авторами на основе анализа материалов

Кроме того, полученные нами результаты согласуются с выводами масштабного полевого исследования Стэнфордского университета, в котором AI-сгенерированные вопросы по математике, информатике и химии оценивались с использованием модели Раша (IRT) на выборке из почти 1700 студентов колледжей (Isley et al., 2025). Исследование показало, что при итеративном уточнении промптов качество заданий, созданных ИИ, становится сопоставимым с экспертными разработками для стандартизированных экзаменов. Важно отметить, что согласно работе Zhang и Graf (2025), наиболее частыми ошибками LLM при решении математических задач являются процедурные сбои (procedural slips), а не концептуальные непонимания, что открывает возможности для их автоматического выявления и коррекции. Эти данные подкрепляют нашу стратегию двухэтапной валидации (автоматическая проверка + экспертная оценка). Кроме того, исследование, проведённое в Люксембурге с участием более 35 тысяч учащихся, подчеркивает важность учёта культурного контекста при генерации заданий: даже при одинаковом когнитивном уровне сложности могут наблюдаться значимые различия в результатах у детей с разным культурным бэкграундом (Sonnleitner et al., 2025). Это усиливает значимость нашего вывода о необходимости адаптации формулировок к казахстанскому образовательному контексту.

В соответствии с выводами J.Ahn и др. (2024), отсутствие унифицированных стандартов и согласованных методик тестирования приводит к фрагментарности результатов и затрудняет сопоставление математических возможностей различных моделей. Эти выводы имеют особую значимость для национальных процедур оценивания, где требуется строгая привязка заданий к спецификации теста, локальным математическим обозначениям и установленным когнитивным уровням.

В этом контексте полученные результаты подчёркивают, что внедрение комплексной экспертизы – ключевое условие обеспечения качества, надёжности и справедливости тестовых материалов, созданных с применением генеративных моделей.

Результаты исследования демонстрируют высокий потенциал генеративных языковых моделей (LLM) для автоматизированного клонирования тестовых заданий по математике. Анализ

показал, что более половины сгенерированных заданий (50,5%) соответствуют стандарту и не требуют дополнительной доработки, что подтверждает высокий базовый уровень корректности создаваемого контента. Этот показатель сопоставим с результатами исследований в области Automatic Item Generation (AIG), где доля пригодных заданий составляет 40–55%, подтверждая практическое применение подхода (Gierl & Lai, 2016).

Почти половина заданий (47 %) потребовала частичных корректировок, главным образом связанных с уточнением формулировок (17,5 %) и балансировкой уровня сложности (19,0 %). Следует отметить, что указанные изменения носили корректировочный, а не концептуальный характер, что свидетельствует о сохранении когнитивной структуры и дидактической адекватности заданий. Низкий уровень брака (2,5–3,2 %) подтверждает устойчивость алгоритмов генерации и их применение для формирования содержательно валидных заданий, соответствующих требованиям тестологического качества.

Матрица приоритетов исправления заданий показала, что большинство несоответствий сосредоточено в зоне «Текучка / Мелкие правки», что указывает на незначительный характер ошибок. Системные проблемы связаны с калибровкой сложности и лингвистической вариативностью формулировок, что требует совершенствования *prompt engineering* и внедрения механизмов автоматического контроля когнитивного уровня.

Наличие незначительного числа критичных, но легко устранимых ошибок («Срочно / Быстро») подтверждает эффективность применения эксперта-аналитика в цикле Human-in-the-Loop. Таким образом, сочетание человеческого контроля и генеративных возможностей модели обеспечивает высокий уровень содержательной верификации и минимизирует риск внедрения некачественных заданий. Эти результаты согласуются с международными исследованиями, подчёркивающими важность экспертной интеграции на этапе автоматизированной генерации (Falcao et al., 2022).

С практической точки зрения внедрение AIG в работу Национального центра тестирования открывает возможности для создания масштабируемой и экономически эффективной системы производства контента. Как показывают данные хронометража, редактирование ИИ-сгенерированной заготовки занимает 5–7 минут, тогда как создание задания «с нуля» требует 25–30 минут. Это сокращение временных затрат в 4–5 раз позволяет существенно повысить производительность экспертов и оперативность обновления банка заданий ЕНТ, что особенно важно в условиях четырёх ежегодных попыток тестирования и необходимости ротации материалов.

Применение генеративных нейросетей с экспертной валидацией создаёт оптимальную модель для обновления банка тестовых заданий, обеспечивая баланс технологической эффективности и педагогической достоверности, сокращая время и повышая воспроизводимость качества контента.

Заключение

1. Генеративные языковые модели продемонстрировали высокую применимость для автоматизированного клонирования тестовых заданий по математике при условии экспертной валидации.

2. Важно отметить, что более 97 % сгенерированных заданий признаны пригодными для использования, при этом 50,5 % – без доработки, что свидетельствует о высоком уровне автоматической корректности и внутренней логической согласованности.

3. Наиболее частыми типами замечаний стали корректировка сложности и формулировок, которые при этом не влияют на когнитивную структуру заданий.

4. Анализ матрицы приоритетов исправления показал, что преобладают несущественные несоответствия, что, в свою очередь, подтверждает зрелость и устойчивость применённых генеративных алгоритмов.

5. Интеграция эксперта-аналитика в цикл *Human-in-the-Loop* доказала свою эффективность для контроля качества и оптимизации трудоёмкости процедуры клонирования.

6. Таким образом, результаты исследования подтверждают перспективность внедрения генеративных моделей в процесс обновления банка тестовых заданий ЕНТ и демонстрируют возможность масштабного применения данной технологии в национальных оценочных системах.

Ограничения исследования и перспективы дальнейшей работы

Следует отметить, что исследование имеет ряд ограничений. В частности, оно проведено только по предмету «Математика», что огра-

ничивает возможность обобщения результатов на другие дисциплины. Психометрическая проверка с участием реальных испытуемых не проводилась, поэтому дальнейшие этапы должны включать статистическую валидацию параметров заданий (трудность, дискриминативность, надёжность). Кроме того, анализ охватывал ограниченное число языковых моделей, что требует расширения выборки. Перспективным направлением является развитие системы автоматизированной пост-обработки с применением машинного обучения и интеграция динамической обратной связи между экспертной оценкой и генерацией, что повысит точность и адаптивность модели AI TestGen.

Перспективным направлением дальнейших исследований является углублённая работа с дистракторами. Как показывают корейские исследования по автоматической генерации заданий для линейных уравнений, только около 60% сгенерированных неверных ответов отражают реальные типичные ошибки учащихся (Kim & Lee, 2025). Это подтверждает наш тезис о необходимости экспертного контроля именно на этапе разработки отвлекающих ответов. Интеграция психометрических индексов (индекс трудности, дискриминативность) непосредственно в процесс генерации, как это предлагается в работе итальянских исследователей с использованием Google NotebookLM (Fulgione, 2025), позволит в будущем создавать задания с заранее заданными измерительными свойствами

Финансирование

Данное исследование было выполнено при финансовой поддержке Национального центра тестирования Министерства науки и высшего образования Республики Казахстан.

Благодарность

Авторы выражают искреннюю благодарность всем участникам исследования за сотрудничество, активное участие в обсуждениях и ценные предложения.

Заявление о конфликте интересов

Авторы заявляют о полном отсутствии конфликта интересов.

Вклад авторов

А.А. Әбжәлиева – формирование замысла исследования, разработка методических подходов, проведение исследовательской работы, участие в организации и проведении исследования, анализ и интерпретация полученных результатов, подготовка первоначального варианта рукописи, редактирование окончательной версии статьи.

Ж.А. Чурбанова – разработка методологической базы, научное сопровождение, координация работы авторского коллектива, верификация достоверности полученных данных, научная правка материала.

А.Ш. Хайдарова – участие в проведении исследования, обработка и систематизация собранных материалов, визуализация результатов, проверка достоверности и обоснованности выводов, участие в редактировании рукописи.

А.Б. Карибаева – сбор эмпирических данных, участие в аналитической работе, анализ и интерпретация полученной информации, подготовка первоначального варианта статьи.

М.З. Абдрахманов – организационное содействие исследованию, обеспечение необходимыми ресурсами, участие в сборе материалов и редактировании текста статьи.

Литература

Отчет о научно-исследовательской работе «Исследование возможностей генеративных нейросетей для формирования банка валидных и надёжных тестовых заданий для оценивания и анализа в системе образования», 2025 / науч. рук. Н.А. Байжанов. – Астана, 2025.

Спецификация теста по математике. 2025. https://testcenter.kz/upload/iblock/577/05_.pdf

Ahn J., Verma R., Lou R., et al. Large Language Models for Mathematical Reasoning: Progresses and Challenges // Proceedings EACL-SRW-2024. – 2024. DOI: 10.18653/v1/2024.eacl-srw.17. https://aclanthology.org/2024.eacl-srw.17

Akhmetov, A., Nurpeiis, M., Abdimanap, A., et al. KenesBERT: A Conversational Kazakh Language Model. arXiv preprint, 2023. arXiv:2303.09303.

Akorda.kz. Глава государства провел первое заседание Совета по развитию искусственного интеллекта. [Электронный ресурс]. – Официальный сайт Президента Республики Казахстан. Астана, 2025. https://www.akorda.kz/ru/glava-gosudarstva-provel-pervoe-zasedanie-soveta-po-razvitiyu-iskusstvennogo-intellekta-193222.

Aljodeh, M. M. Assessing AI-generated math items: Evidence from eighth-grade triangle congruence // TPM – Testing, Psychometrics, Methodology in Applied Psychology, 2025. – № 32(S2). – P. 251–265. https://tpmap.org/submission/index.php/tpm/article/view/219

Bernstein, J., Van Moere, A., Cheng, J. Validity and fairness of automated speaking assessment // Language Testing. -2020. -№ 37(3). – P. 343–368.

Camarata T., McCoy L., Rosenberg R., Temprine Grellinger K.R., Brettschnieder K., Berman J. LLM-Generated multiple choice practice quizzes for preclinical medical students // Advances in Physiology Education, 2025.- №49(3). P.758-763. doi: 10.1152/advan.00106.2024

Deane, P., Higgins, D. Automated Test Item Generation System and Method. ETS Patent. 2019.

Falcão F., Costa P., Pêgo J. M. Feasibility assurance: a review of automatic item generation in medical assessment. 2022. https://pmc.ncbi.nlm.nih.gov/articles/PMC8886703/

Fulgione, M. Artificial intelligence from Google environment for effective learning assessment // Information, 2025. №16(6). P.462.

Gierl, M. J., Lai, H. Automatic item generation // In S. Lane, M. Raymond, & T. Haladyna (Eds.), Handbook of test development (2nd ed., pp. 410-429). Routledge. 2016.

Isley, C., Gilbert, J., Kassos, E., et al. Assessing the quality of AI-generated exams: A large-scale field study. arXiv preprint. 2025.

Kadyrov S, Abdrasilov B, Sabyrov A, Baizhanov N, Makhmutova A., Kyllonen P.C. Evaluating LLMs on Kazakhstan's mathematics exam for university admission // Frontiers in Artificial Intelligence 2025. №8. P.1642570. doi: 10.3389/frai.2025.1642570

Lathrop Q. N., Cheng Y. Item Cloning Variation and the Impact on the Parameters of Response Models. 2016. https://www.researchgate.net/publication/308878852_Item_Cloning_Variation_and_the_Impact_on_the_Parameters_of_Response_Models

Lewkowycz, A., Andreassen, A., Dohan, D., et al. Solving Quantitative Reasoning Problems with Language Models. 2022. https://arxiv.org/abs/2206.14858

Loukina, A., Madnani, N., Zechner, K. The issues of algorithmic fairness and bias in educational applications // In Proceedings of NAACL-HLT. 2022.

Maity S., Deroy A., Sarkar S. Can large language models meet the challenge of generating educational questions aligned with Bloom's taxonomy? // Computers and Education: Artificial Intelligence. 2025. №8. P.100370. doi: 10.1016/j.caeai.2025.100370

OECD. Untapping the potential of resource banks in the classroom. 2023. Available at: https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/07/untapping-the-potential-of-resource-banks-in-the-classroom_75e4c6fc/f1a19b94-en.pdf.

Pecuchova J., Benko Ľ., Drlik M. Automated grading of open-ended questions in higher education using GenAI models // International Journal of Artificial Intelligence in Education, 2025. № 35. P.3813-3846. doi: 10.1007/s40593-025-00517-2

Sakenov, D., Makhambetov, A., Yessenbayev, Z., et al. KazBERT: Pre-trained Language Model for Kazakh.2021. arXiv preprint, arXiv:2108.11834.

Sonnleitner, P., Bernard, S., Michels, M. A., et al. Establishing cognitive item models for fair and theory-grounded automatic item generation: A large-scale assessment study with image-based math items // Applied Measurement in Education, 2025. P. 1–23. https://orbilu.uni.lu/handle/10993/66488

Yoon, S., Lee, H., Park, J. Automatic item generation in various STEM subjects using large language model prompting // Directory of Open Access Journals. 2025.

Zhang, L., Graf, E. A. Mathematical computation and reasoning errors by large language models. 2025. arXiv preprint.

Zhao P., Zhang H., Yu Q., et al. Retrieval-Augmented Generation for AI-Generated Content: A Survey // Data Science and Engineering. 2025. doi: 10.1007/s41019-025-00335-5

References

Ahn J., Verma R., Lou R., et al. (2024). Large Language Models for Mathematical Reasoning: Progresses and Challenges. Proceedings EACL-SRW-2024. DOI: 10.18653/v1/2024.eacl-srw.17. https://aclanthology.org/2024.eacl-srw.17

Akhmetov, A., Nurpeiis, M., Abdimanap, A., et al. (2023). KenesBERT: A Conversational Kazakh Language Model. arXiv preprint, arXiv:2303.09303.

Akorda.kz (2025). Glava gosudarstva provel pervoe zasedanie Soveta po razvitiju iskusstvennogo intellekta [The President held the first meeting of the Council for the Development of Artificial Intelligence] (2025). [Electronic resource]. – The official website of the President of the Republic of Kazakhstan. https://www.akorda.kz/ru/glava-gosudarstva-provel-pervoe-zasedanie-soveta-po-razvitiyu-iskusstvennogo-intellekta-193222

Aljodeh, M. M. (2025). Assessing AI-generated math items: Evidence from eighth-grade triangle congruence. TPM – Testing, Psychometrics, Methodology in Applied Psychology, 32(S2), 251–265. https://tpmap.org/submission/index.php/tpm/article/view/219

Bernstein, J., Van Moere, A., & Cheng, J. (2020). Validity and fairness of automated speaking assessment. Language Testing, 37(3), 343–368.

Camarata T., McCoy L., Rosenberg R., Temprine Grellinger K.R., Brettschnieder K., Berman J. (2025) LLM-Generated multiple choice practice quizzes for preclinical medical students. Advances in Physiology Education 49(3):758-763. doi: 10.1152/advan.00106.2024

Deane, P., & Higgins, D. (2019). Automated Test Item Generation System and Method. ETS Patent.

Falcão F., Costa P., Pêgo J. M. Feasibility assurance: a review of automatic item generation in medical assessment. (2022). https://pmc.ncbi.nlm.nih.gov/articles/PMC8886703/

Fulgione, M. (2025). Artificial intelligence from Google environment for effective learning assessment. Information, 16(6), 462.

Gierl, M. J., & Lai, H. (2016). Automatic item generation. In S. Lane, M. Raymond, & T. Haladyna (Eds.), Handbook of test development (2nd ed., pp. 410-429). Routledge.

Isley, C., Gilbert, J., Kassos, E., et al. (2025). Assessing the quality of AI-generated exams: A large-scale field study. arXiv preprint.

Kadyrov S, Abdrasilov B, Sabyrov A, Baizhanov N, Makhmutova A and Kyllonen P.C. (2025) Evaluating LLMs on Kazakhstan's mathematics exam for university admission. Frontiers in Artificial Intelligence 8:1642570. doi: 10.3389/frai.2025.1642570

Lathrop Q. N., Cheng Y. Item Cloning Variation and the Impact on the Parameters of Response Models. (2016). https://www.researchgate.net/publication/308878852_Item_Cloning_Variation_and_the_Impact_on_the_Parameters_of_Response_Models

Lewkowycz, A., Andreassen, A., Dohan, D., et al. (2022). Solving Quantitative Reasoning Problems with Language Models. https://arxiv.org/abs/2206.14858

Loukina, A., Madnani, N., & Zechner, K. (2022). The issues of algorithmic fairness and bias in educational applications. In Proceedings of NAACL-HLT.

Maity S., Deroy A., Sarkar S. (2025) Can large language models meet the challenge of generating educational questions aligned with Bloom's taxonomy? Computers and Education: Artificial Intelligence 8:100370. doi: 10.1016/j.caeai.2025.100370

OECD (2023). Untapping the potential of resource banks in the classroom. Available at: https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/07/untapping-the-potential-of-resource-banks-in-the-classroom_75e4c6fc/f1a19b94-en.pdf.

Otchet o nauchno-issledovatel'skoj rabote "Issledovanie vozmozhnostej generativnyh nejrosetej dlja formirovanija banka validnyh i nadjozhnyh testovyh zadanij dlja ocenivanija i analiza v sisteme obrazovanija" [Research report "Exploring the possibilities of generative neural networks to form a bank of valid and reliable test items for assessment and analysis in the education system"] (2025) / scientific adviser N.A. Baizhanov. Astana, 2025.

Pecuchova J., Benko Ľ., Drlik M. (2025) Automated grading of open-ended questions in higher education using GenAI models. International Journal of Artificial Intelligence in Education 35:3813-3846. doi: 10.1007/s40593-025-00517-2

Sakenov, D., Makhambetov, A., Yessenbayev, Z., et al. (2021). KazBERT: Pre-trained Language Model for Kazakh. arXiv preprint, arXiv:2108.11834.

Sonnleitner, P., Bernard, S., Michels, M. A., et al. (2025). Establishing cognitive item models for fair and theory-grounded automatic item generation: A large-scale assessment study with image-based math items. Applied Measurement in Education, 1–23. https://orbilu.uni.lu/handle/10993/66488

Specifikacija testa po matematike [Math Test Specification] (2025). https://testcenter.kz/upload/iblock/577/05_.pdf

Yoon, S., Lee, H., & Park, J. (2025). Automatic item generation in various STEM subjects using large language model prompting. Directory of Open Access Journals.

Zhang, L., & Graf, E. A. (2025). Mathematical computation and reasoning errors by large language models. arXiv preprint.

Zhao P., Zhang H., Yu Q., et al. (2026) Retrieval-Augmented Generation for AI-Generated Content: A Survey. Data Science and Engineering. doi: 10.1007/s41019-025-00335-5

Сведения об авторах:

Әбжәлиева Айдана Асылбекқызы – Магистр, главный эксперт, Национальный центр тестирования (Астана, Республика Казахстан, e-mail: abjalieva.aidana@mail.ru);

Чурбанова Жайран Ахметовна – Магистр, руководитель отдела, Национальный центр тестирования (Астана, Республика Казахстан, e-mail: zhairan2302@gmail.com);

Хайдарова Ақжан Шаймұранқызы – Магистр, главный эксперт-аналитик, Национальный центр тестирования (Астана, Республика Казахстан, e-mail: akzhan.khaidarova@gmail.com);

Карибаева Акмарал Биржановна – магистр наук, ведущий эксперт-разработчик лаборатории искусственного интеллекта, Национальный центр тестирования (Астана, Республика Казахстан, e-mail: akaribayeva@gmail.com);

Абдрахманов Мейрим Зикирьяулы – Бакалавр, ведущий специалист PR, маркетинга и международного сотрудничества, Национальный центр тестирования (Астана, Республика Казахстан, e-mail: mierimabdrahmanov@gmail.com).

Information about authors:

Abzhaliyeva Aidana Assylbekkyzy – Master's degree, Chief expert, National Testing Center (Astana, Republic of Kazakhstan, e-mail: abjalieva.aidana@mail.ru);

Churbanova Zhairan Akhmetovna – Master's degree, Head of department, National Testing Center (Astana, Republic of Kazakhstan e-mail: zhairan2302@gmail.com);

Khaidarova Akzhan Shaimurankyzy – Master's degree, Senior Expert-Analyst, National Testing Center (Astana, Republic of Kazakhstan, e-mail: akzhan.khaidarova@gmail.com);

Karibayeva Akmaral Birzhanovna – Master of Science, Lead expert Developer of the Artificial Intelligence Laboratory at the National Testing Center (Astana, Republic of Kazakhstan, e-mail: akaribayeva@gmail.com) ;

Abdirakhmanov Meirim Zikiriyauly – Bachelor's degree, Leading Specialist of PR, Marketing, and International Cooperation, National Testing Center (Astana, Republic of Kazakhstan, e-mail: mierimabdrahmanov@gmail.com).

Авторлар туралы мәлімет:

Әбжәлиева Айдана Асылбекқызы – Магистр, бас сарапшы, Ұлттық тестілеу орталығы (Астана, Қазақстан Республикасы, e-mail: abjalieva.aidana@mail.ru);

Чурбанова Жайран Ахметовна – Магистр, бөлім басшысы, Ұлттық тестілеу орталығы (Астана, Қазақстан Республикасы, e-mail: zhairan2302@gmail.com);

Хайдарова Ақжан Шаймұранқызы – Магистр, бас сарапшы-талдаушы, Ұлттық тестілеу орталығы (Астана, Қазақстан Республикасы, e-mail: akzhan.khaidarova@gmail.com);

Карибаева Акмарал Биржановна – магистр, Жасанды интеллект зерханасының бас сарапшы-әзірлеушісі, Ұлттық тестілеу орталығы (Астана, Қазақстан Республикасы, e-mail: akaribayeva@gmail.com);

Әбдірахманов Мейрім Зікіриұлы – Бакалавр, PR, маркетинг және халықаралық қатынастар қызметінің жетекші сарапшысы, Ұлттық тестілеу орталығы (Астана, Қазақстан Республикасы, e-mail: mierimabdrahmanov@gmail.com).

Поступило: 13.01.2026
Принята: 25.05.2026