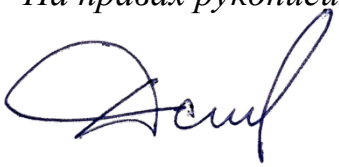
МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования «Московский политехнический университет»

На правах рукописи



Дагаев Александр Евгеньевич

РАЗРАБОТКА АЛГОРИТМОВ ИЗВЛЕЧЕНИЯ И ОБРАБОТКИ ИНФОРМАЦИИ ДЛЯ ГЕНЕРАЦИИ МНОГОФОРМАТНЫХ ТЕСТОВЫХ ЗАДАНИЙ

Специальность 2.3.8. Информатика и информационные процессы

ДИССЕРТАЦИЯ
на соискание учёной степени
кандидата технических наук

Научный руководитель:
доктор технических наук, профессор
Попов Дмитрий Иванович

Москва – 2026 г.

Оглавление

Введение

Актуальность темы исследования

Актуальность исследования обусловлена активным развитием интеллектуальных образовательных систем и цифровых технологий обучения, ориентированных на автоматизацию процессов формирования и диагностики компетенций обучающихся. В условиях реализации компетентностной модели образования возрастает потребность в инструментах, обеспечивающих объективную, масштабируемую и воспроизводимую оценку степени сформированности отдельных компонентов компетенций, включая когнитивные, аналитические и прикладные аспекты.

Одним из ключевых средств такой оценки являются тестовые задания различных форматов, используемые для диагностики усвоения учебного материала и компонентов компетенций. В то же время традиционные подходы к их разработке, основанные преимущественно на ручной подготовке, характеризуются высокой трудоёмкостью, субъективностью и ограниченной масштабируемостью. Это существенно затрудняет их применение в современных образовательных системах, ориентированных на массовое и персонализированное обучение.

Следует подчеркнуть, что разрабатываемые алгоритмы рассматриваются не как замена педагогической деятельности, а как вспомогательный инструментарий цифровой образовательной среды, предназначенный для снижения трудоёмкости рутинных операций преподавателя.

В последние годы всё большее внимание уделяется применению методов искусственного интеллекта и обработки естественного языка для автоматизированной генерации тестовых заданий на основе учебных текстов. Несмотря на достигнутые результаты, существующие решения обычно ориентированы на генерацию отдельных типов заданий или на формальное соответствие исходному тексту без учёта дидактических требований, качества дистракторов и необходимости верификации корректности сформированных

заданий. Также во многих работах отсутствует интеграция этапов извлечения информации, генерации заданий и последующей оценки их качества.

Наиболее сложной является задача автоматизированной генерации дистракторов, то есть неверных вариантов ответа, правдоподобных с точки зрения обучающегося, но логически или содержательно некорректных. Качество дистракторов во многом определяет диагностическую ценность тестового задания. Распространённые методы их формирования основываются на поверхностной семантической близости и не учитывают когнитивные механизмы типичных ошибок обучающихся.

Не менее значимой является проблема верификации сгенерированных тестовых заданий. Отсутствие автоматизированных процедур контроля приводит к появлению заданий с фактическими ошибками, неоднозначными формулировками, отсутствием корректного ответа либо несоответствием заявленному уровню сложности. Это снижает достоверность результатов диагностики и ограничивает практическое применение систем автоматизированной генерации заданий.

В связи с этим актуальной научно-технической задачей является разработка алгоритмов извлечения и обработки информации, обеспечивающих автоматизированную генерацию многоформатных тестовых заданий и дистракторов с последующей интеллектуальной верификацией их качества. Решение данной задачи связано с интеграцией методов обработки естественного языка, алгоритмов генерации контента и формализованных процедур оценки характеристик формируемых заданий, ориентированных на требования компетентностного подхода и практику применения в современных образовательных системах.

Степень разработанности темы.

Проблематика разработки алгоритмов извлечения и обработки информации для автоматизированной генерации тестовых заданий получила значительное развитие в современной научной литературе в области компьютерной лингвистики, искусственного интеллекта и образовательных технологий. Сформирована обширная исследовательская база, охватывающая задачи семантического анализа

учебных текстов, автоматической генерации вопросов различных форматов, формирования дистракторов и последующей верификации результатов генерации.

Базовые подходы к автоматической генерации вопросов (Automatic Question Generation, AQG) были заложены в работах, посвящённых извлечению ключевой информации из текста, синтаксическому и семантическому анализу, а также преобразованию текстового контента в формализованные задания. К данному направлению относятся исследования M. Heilman, N. Smith, R. Mitkov, J. Mostow, S. Kurdi, Y. Zhang, L. Dong, M. Lapata и др., в которых сформированы теоретические основы автоматической генерации вопросов и определены основные классы алгоритмических решений.

Существенный прогресс в развитии AQG связан с внедрением контекстных языковых моделей и трансформерных архитектур. В работах J. Devlin, M. Liu, C. Raffel, A. Vaswani, Y. Tang, Z. Cao показано, что использование контекстно-зависимых представлений текста позволяет повысить семантическую согласованность автоматически формируемых заданий и расширить спектр поддерживаемых форматов.

Современный этап исследований характеризуется активным применением больших языковых моделей, способных выполнять генерацию тестовых заданий в рамках единого генеративного контура. Данные подходы представлены в работах T. Brown, J. Wei, T. Kojima, S. Yao, T. Schick, S. Min и др., где демонстрируется возможность совместной генерации текста вопроса, правильного ответа и набора дистракторов на основе исходного текстового контента. При этом подчёркивается необходимость разработки формализованных процедур контроля корректности и интерпретируемости результатов генерации.

Значительное число исследований посвящено автоматической генерации дистракторов (Automatic Distractor Generation, ADG) как ключевому компоненту заданий с выбором ответа. В работах S. Guo, Z. Jiang, C. Liang, J. Gao, K. Zhou, Y. Sakaguchi рассматриваются методы формирования правдоподобных неверных вариантов ответа с учётом контекста вопроса, семантической структуры текста и типичных ошибок обучающихся.

Отдельное направление исследований связано с автоматизированной верификацией качества сгенерированных тестовых заданий. В более ранних работах применялись преимущественно метрики текстового сходства, однако в современных исследованиях предлагаются многоаспектные подходы, включающие проверку фактической корректности, логической согласованности, релевантности содержания и соответствия уровня сложности. Данные подходы представлены в работах Y. Yuan, W. Wang, D. Kiela, E. Durmus, S. Min и др., включая многоагентные и LLM-ориентированные схемы верификации.

В отечественной научной школе значительный вклад в развитие методов обработки естественного языка и создание языковых моделей для русского языка внесли А. В. Куратов, М. С. Архипов, А. А. Сорокин, М. М. Бурцев, Т. И. Шаврина, А. А. Феногенова, А. А. Панченко, Е. А. Артёмова и др. Разработанные ими русскоязычные языковые модели и корпусы создают технологическую основу для реализации алгоритмов автоматизированной генерации тестовых заданий на русском языке.

Таким образом, при высокой степени разработанности отдельных компонентов автоматизированной генерации тестовых заданий в современной научной литературе отсутствуют комплексные исследования, предлагающие целостный алгоритмический подход, объединяющий указанные компоненты в рамках единого интегрированного контура, ориентированного на многоформатные задания, интеллектуальную верификацию и специфику русскоязычных образовательных текстов.

Цель исследования: разработка и исследование интеллектуальных алгоритмов для извлечения и обработки информации, генерации и верификации многоформатных тестовых материалов из цифрового контента.

Задачи исследования:

1. Провести анализ существующих алгоритмов автоматической генерации тестовых заданий и выявить их ограничения применительно к русскоязычному материалу.

2. Сформировать алгоритм семантического анализа образовательного контента и извлечения из него ключевой информации, релевантной для последующей генерации тестовых заданий.
3. Создать алгоритм генерации многоформатных тестовых заданий, включая задания с выбором одного или нескольких вариантов ответа, задания открытой формы, на установление соответствия и правильной последовательности.
4. Разработать алгоритм интеллектуального формирования дистракторов на основе когнитивного моделирования ошибок обучающихся.
5. Реализовать программный комплекс для генерации контрольно-измерительных материалов и провести его экспериментальную апробацию для подтверждения эффективности и оценки качества генерируемых контрольно-измерительных материалов.

Объект исследования: процессы автоматического извлечения и обработки информации из цифрового образовательного контента для создания тестовых материалов.

Предмет исследования: алгоритмы интеллектуального анализа текстовых данных и автоматической генерации многоформатных тестовых заданий.

Под интеллектуальными алгоритмами в работе понимается совокупность методов, основанных на технологиях искусственного интеллекта, таких как большие языковые модели, мультиагентные системы и когнитивное моделирование, в отличие от жестко детерминированных процедур.

Методологическая база исследования опирается на современные подходы к обработке естественного языка, анализу данных и проектированию алгоритмов. В работе используется сочетание методов компьютерной лингвистики, семантического извлечения информации и методов машинного обучения для разработки и валидации алгоритмов автоматической генерации тестовых заданий. Конкретные подходы включают глубокий семантический, синтаксический и морфологический анализ, а также применение больших языковых моделей для извлечения релевантных данных из образовательных текстов. Экспериментальная

методология включает как качественные, так и количественные методы оценки, а также метрики качества сгенерированных тестовых заданий и сравнительный анализ с существующими алгоритмами.

Теоретическая база исследования опирается на фундаментальные труды в области обработки естественного языка, компьютерной лингвистики и педагогических измерений. Ключевыми теоретическими рамками являются семантическая разметка ролей, теория когнитивной нагрузки в образовательной оценке и принципы психометрического конструирования тестов. Разработанные алгоритмы генерации дистракторов опираются на принципы когнитивного моделирования ошибок, а система верификации – на теорию многоагентных систем. Исследование также учитывает современные достижения в области образовательных технологий, основанных на искусственном интеллекте, интегрируя результаты отечественной и зарубежной научной литературы. Важный теоретический фундамент для семантических и психометрических компонентов разработанных алгоритмов обеспечивают труды таких исследователей, как R. Mitkov, A. E. Goldberg, J. Sweller, F. M. Lord и L. M. Rudner.

Эмпирическая база исследования включает обширные наборы данных цифрового образовательного контента на русском языке: учебники, научные публикации и предметно-ориентированные материалы по различным дисциплинам. Среди эмпирических методов исследования – сравнительный анализ работы предложенных алгоритмов с их базовыми решениями, оценка качества результатов как с помощью автоматизированных метрик, так и на основе экспертной оценки, а также экспериментальная проверка адаптивности алгоритмов к различным образовательным контекстам. Интеграция указанных эмпирических данных с методологическими и теоретическими рамками обеспечивает надёжность и практическую применимость разработанных алгоритмов в условиях реальных образовательных практик.

Научная новизна исследования заключается в следующем:

1. Предложен алгоритм абстрактивного извлечения информации, отличающийся от существующих алгоритмов, основанных на

статистических или общих семантических оценках качества тем, что вводит параметрическую интегральную метрику с доменно-специфичными весовыми коэффициентами, что позволяет повысить релевантность извлекаемой информации в среднем на 13,2 % относительно лучшей из рассмотренных базовых моделей (О1).

2. Разработан алгоритм автоматической генерации многоформатных тестовых заданий, отличающийся от существующих подходов на основе больших языковых моделей тем, что интегрирует модуль мультиагентной верификации, позволяющий снизить долю дефектных заданий на 84,5 %.
3. Предложен алгоритм генерации дистракторов, отличающийся от алгоритмов, основанных на семантической близости к правильному ответу тем, что моделирует процесс построения корректной цепочки рассуждений с последующим внесением когнитивных ошибок для генерации неверных ответов, что позволяет повысить качество дистракторов на 22,7 % относительно базовых алгоритмов.

Теоретическая значимость исследования состоит в развитии теоретических основ компьютерной лингвистики применительно к образовательным технологиям, расширении научных представлений о методах автоматического извлечения и структурирования знаний, формулировании новых принципов создания интеллектуальных образовательных систем.

Результаты исследования вносят вклад в теорию педагогических измерений через разработку новых подходов к автоматизации процессов создания тестовых материалов, в теорию искусственного интеллекта путём создания специализированных алгоритмов обработки образовательного контента.

Практическая значимость работы определяется возможностью использования разработанных алгоритмов в системах электронного обучения, платформах массовых открытых онлайн-курсов и корпоративных системах обучения.

Практические результаты исследования могут быть применены для:

- автоматизации создания тестовых материалов в образовательных учреждениях
- повышения эффективности процессов разработки электронных образовательных ресурсов
- создания адаптивных систем контроля знаний
- разработки инструментов поддержки преподавательской деятельности

Положения, выносимые на защиту.

1. Алгоритм абстрактивного извлечения информации из образовательного контента. Применение в алгоритме интеллектуального отбора аннотаций на основе откалиброванной доменно-настроенной интегральной метрики обеспечивает формирование релевантной базы для генерации тестов с количественным преимуществом в интегральной оценке качества извлекаемой информации на 13,2 % по сравнению с лучшей из рассмотренных современных больших языковых моделей (О1).
2. Алгоритм генерации многоформатных тестовых заданий на основе LLM, интегрированный с мультиагентной системой верификации. Использование ансамбля узкоспециализированных агентов позволяет проводить фильтрацию логически и фактически некорректных формулировок, обеспечивая снижение количества дефектных контрольно-измерительных материалов на 84,5 %.
3. Алгоритм интеллектуального формирования дистракторов, основанный на когнитивном моделировании ошибок. Явное построение цепочки рассуждений с имитацией типичных заблуждений обучающихся обеспечивает повышение правдоподобности и семантического разнообразия неверных вариантов ответа, что дает интегральный прирост качества дистракторов на 22,7 % по сравнению с базовыми алгоритмическими подходами.

Соответствие научно-квалификационной работы Паспорту научной специальности.

Диссертационное исследование соответствует паспорту научной специальности 2.3.8 «Информатика и информационные процессы» по пп. 4 и 5.

Пункт 4. Разработка методов и технологий цифровой обработки аудиовизуальной информации с целью обнаружения закономерностей в данных, включая обработку текстовых и иных изображений, видео контента. Разработка методов и моделей распознавания, понимания и синтеза речи, принципов и методов извлечения требуемой информации из текстов.

Пункт 5. Лингвистическое обеспечение информационных систем и процессов. Методы и средства проектирования словарей данных, словарей индексирования и поиска информации, тезаурусов и иных лексических комплексов. Методы семантического, синтаксического и прагматического анализа текстовой информации для представления в базах данных и организации интерфейсов информационных систем с пользователями.

Разработанные в диссертации алгоритмы по своей функциональной роли являются прямой реализацией «методов и технологий» цифровой обработки текстовой информации и извлечения из нее требуемых сведений, что соответствует пункту 4 паспорта специальности.

В то же время, будучи реализованными в виде программного комплекса, данные алгоритмы выступают в качестве «средств» лингвистического обеспечения информационных систем, направленных на семантический и прагматический анализ текста. Это устанавливает соответствие работы пункту 5 паспорта специальности.

Апробация и реализация результатов исследования.

Результаты исследования апробированы на 4 международных научных конференциях:

- III Международная научно-практическая конференция «Компьютерные приложения для управления и устойчивого развития производства и промышленности»
- XVI Международная научно-практическая конференция «Управление устойчивым развитием России в условиях цифровой трансформации»

- XII Международная научная конференция «Математическое и компьютерное моделирование»
- IX Международная научно-практическая конференция студентов, аспирантов и молодых учёных «Фундаментальные и прикладные исследования молодых учёных»

Основные результаты исследования представлены в 9 работах, включая 4 статьи в журналах, входящих в перечень ВАК (из них 3 (К2) по направлению 2.3.8), 5 публикаций в изданиях РИНЦ, а также 11 зарегистрированных в Роспатенте программ для ЭВМ.

Личный вклад автора. Все основные результаты, выносимые на защиту, получены автором лично. Автором выполнены постановка научной задачи, анализ состояния исследований по теме диссертации, разработка алгоритмов абстрактивного извлечения информации, генерации многоформатных тестовых заданий, интеллектуального формирования дистракторов, программная реализация предложенных решений, организация и проведение экспериментальных исследований, обработка, анализ и интерпретация полученных результатов.

Объем и структура диссертации.

Диссертационная работа состоит из введения, четырёх глав, заключения, списка литературы и приложений. Общий объём работы составляет 146 страниц машинописного текста, содержит 30 таблиц; список литературы включает 136 наименований.

Во введении обоснована актуальность исследования, сформулированы его цель и задачи, определены объект и предмет исследования, описаны методологическая, теоретическая и эмпирическая базы, представлены научная новизна, теоретическая и практическая значимость, положения, выносимые на защиту, а также сведения об апробации результатов.

В первой главе проведён анализ современных подходов к автоматической генерации тестовых заданий и дистракторов, отражающий эволюцию данной области от шаблонных методов к нейросетевым архитектурам и большим языковым моделям. Проанализированы гибридные стратегии с использованием

внешних источников знаний, подходы к генерации дистракторов, а также модели предобученных языковых представлений и метрики оценки качества текстов. Выявлены основные ограничения существующих решений применительно к русскоязычному образовательному контенту.

Во второй главе представлены результаты разработки и исследования алгоритма абстрактивного извлечения информации, предназначенного для формирования текстов, используемых при последующей генерации контрольно-измерительных материалов. Разработан и формализован алгоритм, включающий классификацию предметной области, генерацию множества аннотаций с использованием языковой модели и отбор по интегральной метрике с доменно-зависимыми весовыми коэффициентами. Определены весовые профили для предметных областей. Приведены результаты экспериментальной проверки на корпусе русскоязычных текстов, подтверждающие улучшение качества аннотаций по совокупности используемых метрик по сравнению с базовыми решениями.

В третьей главе изложены результаты разработки алгоритма генерации многоформатных тестовых заданий с модулем мультиагентной верификации. Описана архитектура системы, включающая пять типов заданий и семь специализированных агентов верификации на основе принципа многоагентного оценивания. Представлены результаты экспериментальной оценки на корпусе из 1000 заданий, показавшие снижение доли дефектных заданий на 84,5%.

В четвёртой главе изложены результаты исследования, направленного на разработку алгоритма автоматической генерации дистракторов на основе когнитивного моделирования ошибок. Описана трёхэтапная процедура генерации. Представлены результаты экспериментов с участием пяти независимых экспертов на двух русскоязычных наборах данных – RuOpenBookQA и RuWorldTree. Результаты демонстрируют превосходство предложенного подхода над базовыми алгоритмами по показателям правдоподобия, семантической релевантности и разнообразия. Описан программный комплекс из 11 программ для ЭВМ.

В заключении сформулированы основные выводы исследования, подтверждающие достижение поставленной цели и решение всех задач, намечены

перспективы дальнейших исследований в области автоматизации создания образовательного контента.

Глава 1. Анализ существующих подходов к автоматизированной генерации тестовых заданий

1.1. Теоретические основы генерации тестовых заданий

Автоматизированная генерация тестовых заданий представляет собой самостоятельное и активно развивающееся научное направление, формирующееся на пересечении методов обработки естественного языка, теории педагогических измерений и современных образовательных технологий. В научной литературе для обозначения данного направления используются две устоявшиеся аббревиатуры. Под AQG (Automatic Question Generation) понимается генерация вопросов на основе естественно-языковых текстов. Под AIG (Automatic Item Generation) рассматривается формирование законченных оценочных единиц, включающих условие задания, набор дистракторов и правильный ответ [51, 34].

Принципиальное разграничение понятий AQG и AIG носит не только терминологический, но и методологический характер. Подход AQG ориентирован на генерацию вопросов с опорой на заданный текстовый фрагмент. В рамках AIG, согласно [68], осуществляется формирование полноценного задания контроля знаний, включающее выбор ключевой концепции, формулирование условия задания, синтез вариантов ответа и последующую проверку их дидактической корректности. В связи с этим AIG может рассматриваться как задача более высокого уровня сложности по сравнению с генерацией вопросительной формулировки.

Научная значимость рассматриваемого направления определяется совокупностью взаимосвязанных факторов. Во-первых, масштабирование образовательных платформ в условиях постоянного увеличения числа обучающихся определяет необходимость автоматизированного пополнения банков заданий при обеспечении требуемого уровня качества [108]. Во-вторых, развитие адаптивных систем тестирования требует динамического формирования новых вариантов заданий, что позволяет предотвратить утечку содержания и минимизировать эффект запоминания правильных ответов [103]. В-третьих,

индивидуализация обучения предполагает возможность конструирования диагностических тестов, адаптируемых к уровню подготовки и характерным затруднениям обучающихся, что также определяет необходимость автоматизации процессов генерации и обновления оценочных материалов [103, 118].

Автоматизированная генерация тестовых заданий представляет собой многоэтапный алгоритмический процесс, а не отдельную процедуру текстовой генерации. Указанный процесс включает последовательность взаимосвязанных этапов: извлечение релевантного содержания из исходного материала; выбор ключевой концепции и формулирование правильного ответа; синтез текста вопроса или иного условия задания; генерацию дистракторов; верификацию качества сформированного задания. Нарушение или пропуск любого из перечисленных этапов приводит к снижению дидактической ценности результата [87]. Данное положение имеет методологическое значение при проведении последующего анализа ограничений существующих подходов.

Наряду с AQG и AIG в научной литературе выделяются генерация заданий с множественным выбором как специализированная задача построения MCQ, а также генерация дистракторов (DG) как самостоятельная подзадача формирования правдоподобных неправильных вариантов ответа [23]. Указанные задачи объединяются общим контекстом и относятся не к области генерации текста как таковой, а к области формирования оценочных единиц, к которым предъявляются специфические дидактические и психометрические требования.

Качество автоматически формируемого тестового задания должно оцениваться по критериям, выходящим за пределы грамматической корректности и связности текста. В образовательном контексте выделяется совокупность ключевых измерений качества, непосредственно определяющих диагностическую ценность задания [58].

Дидактическая корректность предполагает соответствие тестового задания заявленной учебной цели и соответствующему таксономическому уровню. Задание, направленное на воспроизведение фактической информации, не тождественно заданию, ориентированному на применение концепции или анализ

причинно-следственных связей. Несоответствие уровня когнитивной нагрузки рассматривается как один из типичных дефектов автоматически формируемых заданий, что отмечается в современных обзорных исследованиях [57].

Однозначность формулировки предполагает, что условие задания не допускает множественных интерпретаций, а правильный ответ является единственным среди предложенных вариантов. Наличие нескольких правильных ответов рассматривается как распространённый дефект, систематически выявляемый при экспертной проверке заданий [25]. В автоматически формируемых заданиях данный дефект обусловлен недостаточной точностью моделирования семантических различий, в результате чего система способна сформировать несколько вариантов ответа, каждый из которых может оказаться частично корректным с точки зрения предметного содержания.

Диагностическая ценность дистракторов определяется степенью их правдоподобия и избирательности по отношению к уровню подготовки обучающихся. Дистрактор должен быть ориентирован на привлечение внимания тех обучающихся, которые не освоили проверяемую концепцию, и не должен вводить в заблуждение обучающихся, обладающих соответствующими знаниями [20]. Указанное требование определяет необходимость детальной проработки механизмов формирования дистракторов, представляющих собой самостоятельную исследовательскую задачу в рамках AQG и AIG.

Фактическая и логическая корректность предполагают, что задание должно содержать достоверные сведения, а правильный ответ должен быть подтверждённым надёжными источниками. Указанное требование приобретает критическое значение при использовании больших языковых моделей, способных генерировать убедительно сформулированный, но фактически некорректный текст [63].

Соответствие учебному материалу предполагает тематическую соотнесённость задания с конкретным разделом курса и требуемым уровнем детализации, установленным образовательной программой. Автоматически формируемые задания могут отклоняться от заданного содержательного контекста,

а также характеризоваться избыточным упрощением либо чрезмерной сложностью по отношению к проверяемому материалу [66].

Совокупность перечисленных требований определяет сложный характер оценки качества в задачах AQG и AIG, которая не может быть сведена к применению стандартных метрик генерации текста. По этой причине в ряде современных исследований оценка качества формируемых заданий рассматривается как один из наиболее сложных и наименее формализованных компонентов систем автоматизированной генерации заданий [108, 54].

Многообразие форматов тестовых заданий, применяемых в современной образовательной практике, определяет спектр задач автоматизированного формирования заданий. Наиболее распространённым является формат задания с множественным выбором (multiple-choice question, MCQ), включающий условие задания в виде вопроса либо незавершённого утверждения, один правильный ответ и несколько неправильных вариантов ответа (дистракторов) [94]. Указанный формат выступает наиболее распространённым объектом исследований в области AQG и DG.

Задания с кратким свободным ответом (short answer) предполагают самостоятельное формирование ответа обучающегося и требуют использования иной архитектуры построения заданий, при которой основное внимание уделяется формулированию открытого вопроса, допускающего проверку корректности ответа без использования фиксированного набора вариантов. Задания на заполнение пропусков (fill-in-the-blank, FITB) являются близкими к указанному формату, однако по своей структуре приближаются к заданиям с множественным выбором. В данном случае система определяет ключевое слово или фрагмент, подлежащий изъятию из исходного текста, а задача обучающегося заключается в восстановлении пропущенного элемента [32].

Задания на установление соответствия направлены на проверку способности обучающихся устанавливать связи между понятиями, датами, определениями и примерами. Формат упорядочивания используется для проверки знаний о последовательности процессов, этапов или событий. Указанные форматы

представлены в исследованиях по автоматизированному формированию заданий существенно реже по сравнению с заданиями с множественным выбором, несмотря на их широкое применение в образовательной практике [87, 97].

Автоматизированный процесс формирования тестовых заданий, независимо от целевого формата, включает последовательность функциональных этапов: извлечение содержания, предполагающее идентификацию ключевых концепций, фактов, терминов и связей в исходном материале; выбор фокуса, заключающийся в определении проверяемого компонента содержания; синтез условия задания, предусматривающий формулирование условия в вопросительной или утвердительной форме; формирование вариантов ответа, включающее построение правильного ответа и дистракторов; верификацию, направленную на контроль качества сформированного задания [118].

Характерной особенностью большинства существующих систем является охват лишь части перечисленных этапов и поддержка лишь ограниченного набора форматов тестовых заданий. В частности, системы, ориентированные исключительно на формирование заданий с множественным выбором на основе англоязычных текстов, не могут быть непосредственно перенесены в условия многоформатного формирования заданий на основе русскоязычных учебных материалов. Указанное несоответствие рассматривается как одно из ключевых ограничений предметной области.

1.2. Эволюция подходов к автоматизированной генерации тестовых заданий

На ранних этапах развития автоматизированного формирования вопросов и тестовых заданий преобладали шаблонные и правило-ориентированные подходы. В их основе лежит преобразование входного предложения в вопросительную форму с использованием фиксированных синтаксических шаблонов и лингвистических правил. Указанные подходы получили широкое распространение в 1990-х и 2000-х годах и заложили терминологическую основу соответствующего направления. Вместе с тем, в более поздних исследованиях они сохраняют

актуальность в качестве базового решения для сравнения при оценке современных методов [53].

Типичный шаблонный подход предполагает выполнение следующих этапов: разбор синтаксической структуры исходного предложения; выбор фокусного элемента, представленного именной или глагольной группой; применение трансформационного шаблона для получения вопросительной конструкции. При корректной разработке шаблонов данный подход обеспечивает грамматическую правильность формируемых вопросов [84].

Онтологические и правило-ориентированные системы, использующие предметные онтологии, обеспечивают формирование заданий, привязанных к конкретной предметной области, такой как биология, медицина, география и другие [71]. Преимуществом данного подхода является высокая предсказуемость и интерпретируемость, поскольку разработчик системы располагает полной информацией о правилах, в соответствии с которыми формируется каждое задание. Вместе с тем, ограничения данного подхода являются существенными: жестко заданные шаблоны обеспечивают ограниченное лексическое разнообразие, а их разработка и поддержка требуют значительных экспертных усилий при переходе к каждой новой предметной области.

Масштабируемость шаблонных систем существенно ограничена, поскольку расширение охвата предметных областей или форматов заданий требует ручной разработки новых шаблонов. При переходе к морфологически богатым языкам, к которым относится русский язык, трудоемкость поддержки шаблонов многократно возрастает вследствие необходимости учета падежных форм, согласования и вариативности порядка слов. Вместе с тем шаблонные системы первого поколения определили ключевые этапы процесса формирования заданий и сформулировали базовые требования к их качеству, сохраняющие актуальность в настоящее время.

Развитие корпусной лингвистики и методов статистической обработки текстов обусловило переход к подходам, основанным на использовании параллельных и аннотированных корпусов вида «контекст – вопрос». Ключевой идеей данного перехода является рассмотрение генерации вопроса как задачи

последовательного преобразования: на вход подается текстовый фрагмент и, при необходимости, размеченный ответ, а на выходе формируется вопрос, соответствующий заданному контексту и ответу.

Использование статистических языковых моделей и методов машинного перевода позволило представить задачу автоматизированного формирования вопросов как задачу последовательного преобразования, что открыло возможность обучения на размеченных данных вместо ручного задания правил [41]. Ранние исследования в данном направлении опирались на параллельные корпуса, в которых вопросы и соответствующие им контексты рассматривались как пары «исходная последовательность – целевая последовательность» в терминах машинного перевода.

Преимуществом статистического подхода является способность выявлять закономерности, присутствующие в больших корпусах, и формировать вопросы с более высоким уровнем лексического разнообразия по сравнению с шаблонными методами. Ограничения данного подхода обусловлены зависимостью качества модели от объема и обучающих данных: при дефиците размеченных пар «контекст – вопрос» для конкретного предметного домена или языка модели демонстрируют существенное снижение качества. Кроме того, статистическая парадигма не обеспечивает формирование вопросов с достаточной смысловой глубиной: модели склонны к поверхностному перефразированию без полноценного учета содержания.

Существенное изменение методологии автоматизированного формирования вопросов связано с внедрением нейросетевых архитектур типа encoder–decoder. Исследования с использованием рекуррентных нейронных сетей, в частности работы Du, Shao и Cardie [41], продемонстрировали возможность обучения моделей формированию осмысленных вопросов на основе входного текста без явного задания трансформационных правил. Дальнейшее развитие получили модели, в которых повышается управляемость процесса формирования вопросов за счет разделения операций копирования и генерации слов, а также за счет выделения признаков, определяющих фокус вопроса [76]. Введение механизмов

внимания позволило моделям устанавливать соответствие между отдельными элементами формируемого вопроса и релевантными фрагментами исходного текста [135].

Модели генерации вопросов с учетом заданного ответа предполагают включение в процесс формирования заданий этапа явного маркирования целевого ответа во входном тексте, что позволяет формировать вопросы, семантически соотнесенные с конкретным фрагментом ответа [110]. Это обеспечивает повышение релевантности и связности формируемых вопросов по сравнению с моделями, не учитывающими информацию о целевом ответе. Дальнейшее развитие получили также модели, ориентированные на формирование вопросов в расширенном контекстном и диалоговом окружении [39].

Важным преимуществом нейросетевых архитектур типа encoder–decoder является их способность обучаться на данных без явного задания лингвистических правил, что способствует расширению охвата языковых конструкций. Вместе с тем при применении к низкоресурсным языкам и специализированным предметным областям обучение таких моделей сталкивается с выраженным дефицитом размеченных данных. Кроме того, обработка дальних зависимостей в длинных текстах в архитектурах LSTM и GRU осуществляется менее эффективно по сравнению с трансформерными моделями, разработанными позднее.

Дополнительным ограничением данного класса моделей является отсутствие встроенных механизмов оценки качества: сформированные вопросы принимаются без применения автоматизированных процедур верификации фактической корректности и диагностической ценности. Задача формирования дистракторов в большинстве систем указанного периода либо отсутствует, либо решается с использованием простых эвристических методов.

Введение архитектуры трансформера, основанной на механизме самовнимания [120], ознаменовало переход к качественно новому уровню языкового моделирования. Предварительно обученные на больших корпусах модели, такие как BERT [36], T5 [96], BART [72], продемонстрировали способность использовать языковые и фактические знания, сформированные на этапе

предобучения, при решении прикладных задач, включая автоматизированное формирование вопросов посредством дообучения на относительно небольших размеченных наборах данных.

Модель T5, рассматривающая любую задачу обработки естественного языка как задачу преобразования «текст – текст» [96], получила широкое распространение в качестве базовой архитектуры для автоматизированного формирования вопросов. Аналогичный подход применяется и в прикладных исследованиях, посвященных генерации вопросов на основе T5 [56]. Дообучение модели на структуре «контекст, ответ, вопрос» позволяет формировать вопросы, характеризующиеся связностью, грамматической правильностью и тематической релевантностью. На основе T5 была реализована одна из первых комплексных систем сквозного формирования заданий с множественным выбором, охватывающая этапы от генерации вопроса до формирования дистракторов [100]. Модель BART [72], реализующая архитектуру шумоподавляющего автокодировщика, также широко применяется для задач генерации вопросов, в частности, в системах генерации вопросов и ответов, объединяющих соответствующие этапы в едином процессе.

Предварительно обученные языковые модели не только повышают качество формируемых вопросов, но и снижают зависимость от объемных доменно-специфичных обучающих данных. Предобучение на масштабных корпусах обеспечивает перенос языковых и фактических знаний, которые могут быть адаптированы к новому предметному домену при ограниченном последующем обучении [96, 56, 45]. Вместе с тем, как отмечается в ряде исследований [63, 61], предобученные модели наследуют смещения, присущие обучающим корпусам, и способны формировать задания, корректные с синтаксической точки зрения, но содержащие фактические неточности или логические противоречия.

Следует отметить, что переход к предварительно обученным трансформерным моделям не устранил проблему верификации качества: большинство систем по-прежнему формируют задания без использования автоматизированных механизмов контроля фактической и дидактической

корректности, что приводит к необходимости возложения данной функции на преподавателя.

Появление больших языковых моделей, таких как GPT-4 [19], LLaMA 2 [116], а также инструкционно настроенных моделей, включая InstructGPT [89], существенно расширило возможности автоматизированного формирования тестовых заданий. Способность таких моделей следовать инструкциям, сформулированным на естественном языке, без специализированной настройки на конкретную задачу либо при использовании ограниченного числа примеров позволила упростить процесс формирования заданий. В частности, вместо отдельного дообучения для каждого формата задания становится возможным управлять результатом через формализацию запроса [28].

Для задач автоматизированного формирования вопросов и тестовых заданий применяются различные стратегии формализации запросов: прямая постановка задачи, включение нескольких демонстрационных примеров, а также поэтапная декомпозиция задачи с раздельным заданием требований к содержанию вопроса, правильному ответу и дистракторам [62, 65]. Практические исследования показывают, что структурированные запросы с четко заданными требованиями к формату, уровню сложности и качеству дистракторов обеспечивают более высокое качество формируемых заданий по сравнению с минимально специфицированными запросами [62, 65].

Принципиальные ограничения подходов, основанных на использовании больших языковых моделей, подробно отражены в научной литературе. Одним из ключевых ограничений является феномен галлюцинаций, заключающийся в формировании фактически некорректной информации, представленной с высокой степенью уверенности [61]. В контексте образовательного содержания это означает, что модель может формировать ответы, противоречащие содержанию исходного учебного материала, либо создавать задания, в которых правильный ответ оказывается фактически некорректным.

Кроме того, большие языковые модели демонстрируют нестабильность качества генерации в зависимости от формулировки запроса, а их способность

формировать задания, соответствующие высоким таксономическим уровням (анализ, оценка, синтез), существенно уступает способности формировать вопросы, направленные на воспроизведение и понимание [80]. Поверхностная обработка предметного содержания, особенно при работе с узкоспециализированными предметными областями, приводит к формированию технически корректных, но дидактически недостаточно проработанных заданий.

Большие языковые модели при отсутствии внешнего контроля не обладают встроенными механизмами надежной самопроверки выходных данных: сформированные задания не проходят автоматизированную проверку фактической и дидактической корректности, если соответствующий этап специально не включен в архитектуру системы. Современные исследования подчеркивают, что отсутствие встроенных процедур контроля качества остается одним из главных нерешенных вопросов применения больших языковых моделей в образовательной оценке [54, 48].

Отдельное направление составляют гибридные подходы, в которых сочетаются генеративные возможности языковых моделей с контролируемыми шаблонными механизмами формирования заданий. Такие решения позволяют повысить структурную корректность формируемых заданий при сохранении вариативности формулировок. Комбинирование генеративных моделей с шаблонными конструкциями способствует снижению числа синтаксических и логических дефектов в заданиях, а также повышению их дидактической пригодности [66].

Использование генеративных моделей показывает, что высокая языковая связность результата не гарантирует его фактической корректности и тематического соответствия исходному материалу. Это способствовало развитию гибридных подходов, в рамках которых генеративная модель дополняется внешними источниками верифицированного знания и специализированными процедурами контроля результатов.

Метод Retrieval-Augmented Generation (RAG) [73] представляет собой подход, при котором генеративная модель перед формированием ответа или

вопроса осуществляет извлечение релевантных фрагментов из внешнего корпуса и включает их в контекст генерации. Для автоматизированной генерации тестовых заданий данный подход обеспечивает опору формируемого задания на конкретный учебный материал, а не исключительно на параметрические знания модели [49]. Расширения RAG, основанные на использовании графовых структур (GraphRAG) [93], применяют представление знаний в виде графа для повышения точности и контекстной согласованности поиска. Это важно при работе со сложными многоуровневыми концептуальными связями, характерными для учебных дисциплин.

Онтологии и графы знаний используются для структурирования предметного содержания и обеспечения корректности формируемых заданий [90]. В исследованиях граф сущностей предметной области применяется в качестве источника кандидатов для формирования правильного ответа и дистракторов. Узлы графа, семантически близкие к правильному ответу, используются как основа для генерации правдоподобных дистракторов [131]. Такой подход позволяет снизить риск тематического рассогласования между формулировкой задания, правильным ответом и дистракторами.

Мультимодальные данные, включая изображения, схемы, формулы и диаграммы, постепенно интегрируются в системы автоматизированной генерации тестовых заданий для поддержки форматов, требующих визуального анализа [133]. Образовательный контент, особенно в естественно-научных, математических и инженерных дисциплинах, в значительной степени представлен в мультимодальной форме, что создает дополнительные сложности для систем генерации, ориентированных преимущественно на обработку текста.

В совокупности гибридные стратегии демонстрируют, что эффективная система генерации тестовых заданий представляет собой не отдельную языковую модель, а архитектурную композицию модулей извлечения содержания, генерации, обогащения знаниями и верификации. Такое представление об архитектуре формирует методологическую основу для последующей постановки научной задачи настоящего исследования.

1.3. Современные подходы к генерации дистракторов

Дистрактор (отвлекающий вариант ответа) является структурно необходимым элементом задания с множественным выбором, поскольку именно качество неверных вариантов ответа в значительной степени определяет различительную способность и диагностическую ценность задания [58, 20]. Очевидно неправдоподобные варианты ответа снижают диагностическую ценность задания и фактически сводят его к исключению заведомо неподходящих альтернатив.

Требования к качеству дистрактора носят двойственный характер: он должен одновременно быть тематически и семантически близким к правильному ответу и при этом оставаться логически и содержательно некорректным по отношению к поставленному вопросу [20, 101]. Указанное сочетание требований определяет высокую сложность задачи генерации дистракторов в составе автоматизированного построения тестовых заданий.

С точки зрения педагогических измерений наибольшую ценность представляют дистракторы, отражающие типичные ошибки и частичные заблуждения обучающихся, поскольку такие варианты ответа позволяют не только дифференцировать уровни освоения материала, но и выявлять характер пробелов в знаниях [27]. В этой связи генерация дистракторов должна рассматриваться не как вспомогательная, а как самостоятельная и методически значимая задача в составе общего контура автоматизированной генерации тестовых заданий.

Отсутствие систематического учета типичных когнитивных ошибок обучающихся является одним из ключевых ограничений, выявляемых при итоговом анализе степени разработанности рассматриваемой темы.

Первое поколение методов автоматизированной генерации дистракторов основывалось на использовании лексических ресурсов и простых эвристических правил. Типичный подход предполагает применение тезаурусов (например, WordNet) для поиска слов, близких к правильному ответу по семантическому расстоянию, включая синонимы, элементы одной гиперонимической группы и семантически смежные концепты [99].

Корпусные методы опираются на статистические меры близости слов (в частности, частотную совместную встречаемость и косинусное сходство векторных представлений, полученных с использованием моделей word2vec) для идентификации кандидатов в дистракторы, тематически близких к правильному ответу [111]. Морфологические преобразования применяются при формировании дистракторов для заданий, направленных на проверку лексических и грамматических навыков, и включают изменение словообразовательных элементов, формы слова или замену на графически сходные лексемы.

Достоинством лексических методов является относительная простота реализации и высокая интерпретируемость получаемых результатов. Вместе с тем их ограничения носят принципиальный характер: лексическая близость не гарантирует семантической правдоподобности в контексте конкретного задания. Дистрактор может быть близким к правильному ответу по данным тезауруса, но оставаться некорректным или абсурдным в контексте поставленного вопроса. Также наблюдается обратная ситуация: в узкоспециализированных предметных областях тезаурусные методы формируют дистракторы, которые оказываются семантически тривиальными и легко распознаваемыми обучающимися [38].

Ключевой нерешенной проблемой методов первого поколения является поверхностная правдоподобность дистракторов: формируемые варианты ответа демонстрируют сходство с правильным ответом преимущественно на уровне лексических признаков и не отражают содержательных ошибок и типичных заблуждений обучающихся.

Ограничения лексических методов обусловили переход к контекстно-зависимым подходам генерации дистракторов. Ключевая идея данных подходов заключается в том, что качество дистрактора должно оцениваться не абстрактно, а относительно конкретного вопроса и его контекста: дистрактор, приемлемый в одном задании, может оказаться некорректным в другом [119].

Модели ранжирования кандидатов в дистракторы [74] предполагают генерацию широкого множества кандидатов с последующим их упорядочиванием на основе специализированных признаков, включая семантическую близость к

правильному ответу, правдоподобность в контексте и грамматическую согласованность с формулировкой задания. Данный подход позволяет существенно повысить качество итогового набора дистракторов по сравнению с методами, основанными на использовании тезаурусов.

Перспективное направление представляют подходы, использующие данные о реальных ошибках обучающихся при генерации дистракторов, поскольку такие данные позволяют формировать варианты ответа, отражающие фактически наблюдаемые заблуждения, а не только формальную семантическую близость [27]. В рамках данного подхода дистракторы формируются не как лексически близкие варианты, а как конкретные ошибочные ответы, фиксируемые в образовательной практике, что обеспечивает их высокую дидактическую правдоподобность.

Вместе с тем вероятностные подходы, опирающиеся на общеязыковые корпуса, нередко приводят к формированию дистракторов, которые оказываются корректными ответами на поставленный вопрос, особенно в условиях семантической многозначности или перефразирования. Указанный дефект, при котором дистрактор совпадает с правильным ответом по смыслу, относится к числу типичных ошибок автоматически сформированных заданий.

С появлением предварительно обученных языковых моделей генерация дистракторов была переосмыслена как задача преобразования «текст → текст». Существенный вклад в развитие данного направления внесли исследования, посвящённые генерации дистракторов с использованием внешних знаний и контекста для заданий с заполнением пропусков и заданий на понимание прочитанного текста, в которых продемонстрированы преимущества нейросетевых и гибридных моделей по сравнению с эвристическими методами [99, 69].

Совместная генерация вопроса и дистракторов представляет собой подход, при котором одна модель одновременно формирует условие задания, правильный ответ и дистракторы, что позволяет учитывать их взаимную согласованность. Системы на основе архитектур T5, BART и других моделей класса «текст → текст», а также подходы, учитывающие показатели дискриминативности задания и

свойства набора дистракторов, получили развитие в работах 2023–2024 гг. [100, 101, 115, 121].

Применение больших языковых моделей для генерации дистракторов на основе запросов расширяет возможности формирования разнообразных вариантов без специализированного дообучения. В рамках данного направления используются стратегии, предполагающие первоначальное формирование расширенного множества кандидатов с последующим их ранжированием и фильтрацией [26]. Такой подход способствует увеличению разнообразия дистракторов, однако требует наличия надёжных механизмов их последующей оценки.

Многоэтапные подходы к формированию заданий с множественным выбором [79], основанные на использовании цепочек рассуждений [123], предполагают последовательное выполнение ряда шагов, включая выявление типичных концептуальных ошибок по теме, формирование дистракторов, отражающих указанные ошибки, и последующее исключение вариантов, совпадающих с правильным ответом. Данный подход концептуально близок к методам, ориентированным на учет типичных ошибок обучающихся, однако в значительной степени опирается на внутренние знания модели, а не на эмпирические данные образовательной практики.

Несмотря на достигнутый прогресс, методы генерации дистракторов, основанные на использовании больших языковых моделей, не устраняют проблему принципиально: формируемые варианты могут оказаться корректными ответами, быть тривиальными либо семантически избыточными. Согласно современным обзорным исследованиям, задача формирования диагностически ценных дистракторов, отражающих реальные когнитивные ошибки обучающихся, остается нерешённой [20, 59].

Понимание необходимости сопровождения генерации дистракторов процедурами их оценки и отбора обусловило развитие многоэтапных конвейеров, включающих ранжирование и фильтрацию. В рамках таких конвейеров на первом этапе формируется расширенное множество кандидатов, на втором этапе

применяются многомерные критерии оценки, на третьем этапе осуществляется отбор итогового набора дистракторов.

Критерии ранжирования кандидатов включают: (1) семантическую близость к правильному ответу; (2) дискриминативность, понимаемую как вероятность выбора дистрактора неподготовленным обучающимся; (3) независимость, выражающуюся в минимальном семантическом перекрытии между дистракторами в пределах одного задания; (4) синтаксическую согласованность с формулировкой задания [74, 77].

Оценочный инструментарий для генерации дистракторов включает как автоматические метрики (NDCG, F1@k, P@k), так и специализированные метрики, разработанные с учетом специфики данной задачи. Метрика DISTO [50], предложенная в качестве обучаемой функции оценки сгенерированных дистракторов, демонстрирует более высокую корреляцию с экспертными оценками по сравнению со стандартными метриками BLEU и ROUGE, не ориентированными на оценку качества дистракторов. Сходное направление исследований связано с объединением генерации дистракторов и их специализированной оценки в рамках единого алгоритмического контура [29]. Указанное наблюдение имеет принципиальное значение, поскольку универсальные метрики качества генерации текста недостаточно отражают специфические требования, предъявляемые к дистракторам.

Переход от оценки на основе семантической близости к моделированию типичных заблуждений обучающихся рассматривается в научной литературе как перспективное, но в недостаточной степени реализованное направление [81]. Основная трудность связана с отсутствием масштабных корпусов, содержащих реальные ошибки обучающихся, для большинства предметных областей и языков. Для русскоязычного образовательного контекста данная проблема является особенно актуальной.

1.4. Подходы к оценке и верификации качества автоматически сформированных заданий

Оценка качества автоматически сформированных тестовых заданий традиционно основывается на двух взаимодополняющих подходах: использовании автоматических метрик и экспертной оценке. Автоматические метрики, первоначально разработанные для задач машинного перевода, такие как BLEU [91] и ROUGE [75], широко применялись и в задачах автоматизированной генерации тестовых заданий для количественного сопоставления сгенерированных вопросов с эталонными. Метрики BLEURT [106] и BERTScore [132], основанные на использовании семантических представлений вместо лексического перекрытия, обеспечивают более высокую корреляцию с экспертными оценками качества текста.

Вместе с тем, принципиальное ограничение указанных метрик применительно к образовательным заданиям заключается в их нечувствительности к педагогически значимым характеристикам. Высокие значения метрик BLEU или ROUGE, отражающие лексическое совпадение с эталонным вопросом, не позволяют сделать вывод о фактической корректности, диагностической ценности или соответствии вопроса требуемому таксономическому уровню [88].

Специализированные метрики, разработанные для задач автоматизированной генерации тестовых заданий, направлены на преодоление данного ограничения. Метрика RQUGE [85], на оценке разрешимости сгенерированного вопроса, позволяет определить, может ли быть получен корректный ответ на основе заданного контекста без прямого сопоставления с эталонной формулировкой. Метрика PMAN [122] предназначена для оценки соответствия сгенерированного вопроса предложенному правильному ответу, что является критически важным для проверки корректности ключа задания. Более детализированные многоаспектные схемы оценки представлены в современных оценочных наборах, в частности в QGEval, где качество вопроса анализируется по ряду содержательных и языковых критериев [48].

Экспертная оценка сохраняет статус основного метода верификации качества образовательных заданий, поскольку только эксперт в предметной области способен надёжно оценить их фактическую корректность, дидактическую ценность и соответствие образовательной программе. Вместе с тем данный подход обладает существенным ограничением, связанным с отсутствием масштабируемости: проведение экспертной оценки требует значительных временных и финансовых затрат, что затрудняет её применение в условиях массового использования систем автоматизированной генерации заданий на образовательных платформах [130]. Эмпирические исследования качества заданий с множественным выбором в медицинском образовании показывают, что даже формально корректные вопросы нередко не соответствуют требованиям к заданиям высокого когнитивного уровня, качеству дистракторов и психометрическим характеристикам [40, 127, 42, 70].

Дефекты автоматически сформированных заданий систематизированы в современных обзорных исследованиях. Наиболее распространённые из них могут быть классифицированы по нескольким категориям.

Фактологические ошибки представляют собой один из наиболее существенных дефектов генерации на основе больших языковых моделей: модель формирует задания, содержащие ложные предпосылки или некорректные правильные ответы. Данный тип ошибок широко отражён в научной литературе [63, 61] и обусловлен использованием параметрических знаний модели без их верификации по внешним источникам.

Логические противоречия возникают в случаях, когда условие задания и правильный ответ оказываются несовместимыми, когда дистрактор может быть отнесён к числу корректных ответов, либо когда дистракторы пересекаются между собой по смыслу. Указанный класс дефектов является более сложным для автоматического выявления, поскольку требует анализа семантических отношений между компонентами задания.

Нарушение принципа однозначности проявляется в ситуациях, при которых задание допускает несколько правильных ответов. Подобные дефекты часто

возникают при генерации вопросов на основе содержательно насыщенных текстов. Современные обзорные и прикладные исследования показывают, что автоматически сформированные задания нередко характеризуются неоднозначностью формулировок и зависимостью ответа от интерпретационного контекста [55].

Дидактические дефекты, включая несоответствие таксономическому уровню, отсутствие диагностической нагрузки, тривиальность либо избыточную сложность, практически не поддаются автоматическому выявлению без привлечения педагогической экспертизы или специализированных оценочных модулей. Их систематическое обнаружение требует разработки задач-ориентированных критериев оценки, выходящих за рамки стандартных метрик генерации текста [44].

Недостаточность существующих метрик, а также высокая трудоёмкость экспертной оценки обусловили развитие автоматизированных механизмов верификации. В данном направлении можно выделить несколько классов решений, каждый из которых обладает специфическими преимуществами и ограничениями.

Правило-ориентированная верификация основана на применении фиксированного набора формализованных правил, направленных на выявление типовых дефектов: дублирования дистракторов, нарушения синтаксического параллелизма между вариантами ответа, наличия очевидно некорректных вариантов, а также несоответствия грамматических характеристик (числа, рода) между ключом и дистракторами [86]. Данный подход хорошо масштабируется, однако принципиально ограничен в обнаружении семантических и содержательных дефектов.

Верификация на основе извлечения внешнего контекста предполагает использование поиска по базе знаний или текстовому корпусу для проверки фактологической корректности задания. В случае, если утверждение, соответствующее правильному ответу, противоречит данным из верифицированных источников, задание помечается как потенциально

некорректное. Эффективность данного подхода напрямую зависит от полноты, актуальности и доменной релевантности используемой базы знаний.

Использование больших языковых моделей в роли оценщика [136] представляет собой подход, при котором LLM применяются для автоматической оценки качества текстовых результатов. В контексте AQG данный подход используется для анализа сгенерированных заданий по совокупности критериев, включая корректность ключа, правдоподобность дистракторов и ясность формулировки. Исследования, в которых GPT-4 применяется в качестве оценщика [136], демонстрируют высокую степень согласованности с экспертными оценками. Вместе с тем данный подход характеризуется рядом ограничений, включая позиционное смещение (position bias), предпочтение более развёрнутых ответов и склонность к самоподтверждению (self-enhancement bias).

Многоагентные подходы к верификации предполагают распределение функций оценки между несколькими специализированными агентами, каждый из которых отвечает за отдельный аспект качества задания: фактическую корректность, дидактическую обоснованность, логическую согласованность компонентов и т.д. [125]. Подход ChatEval [31], основанный на организации взаимодействия между несколькими LLM-агентами в форме аргументированного диалога, демонстрирует преимущество по сравнению с одноагентной оценкой в ряде задач. Применительно к образовательному контенту разработка полноценных многоагентных конвейеров верификации представляет собой актуальное направление исследований; при этом решения, адаптированные к специфике автоматической генерации тестовых заданий (AIG), в существующей литературе представлены ограниченно.

Применительно к русскоязычным системам инфраструктура использования больших языковых моделей в роли оценщика развита существенно слабее по сравнению с англоязычными исследованиями. Наличие универсальных русскоязычных оценочных наборов, таких как MERA [47] и TAPE [114], имеет важное значение для общей диагностики моделей, однако специализированные

инструменты верификации образовательных заданий на русском языке в настоящее время остаются недостаточно развитыми.

1.5. Особенности применения существующих подходов к русскоязычному образовательному контенту

Русский язык относится к флективному типу языков и характеризуется развитой морфологической системой, высокой степенью изменяемости словоформ и относительно свободным порядком слов. Указанные свойства формируют дополнительные сложности при автоматизированной генерации тестовых заданий, нехарактерные для систем, ориентированных на аналитический английский язык.

В русском языке существительные, прилагательные, числительные и причастия изменяются по падежу, роду и числу, а глаголы — по лицу, числу, времени, виду и залогу. Для именной группы в структуре тестового задания это означает необходимость строгого морфологического согласования всех компонентов: вопросительного слова, ключевого понятия и зависимых определений [113]. Автоматически сформированные задания с нарушениями падежного или родо-числового согласования воспринимаются носителями языка как грубые ошибки, что существенно снижает доверие к системе.

Проблема морфологического согласования приобретает особую значимость при генерации дистракторов: варианты ответа, семантически правдоподобные, могут становиться грамматически некорректными при их включении в контекст условия вследствие несоответствия падежной формы. Поскольку большинство существующих систем генерации дистракторов разработано для английского языка, характеризующегося сравнительно низкой степенью флективности, их прямое применение к русскоязычным заданиям без специализированной морфологической постобработки является некорректным [64].

Относительно свободный порядок слов в русском языке формирует дополнительную вариативность при преобразовании декларативных предложений в вопросительные: возможны несколько грамматически корректных формулировок, различающихся информационной структурой высказывания. Для

систем автоматической генерации вопросов это означает необходимость явного управления информационной структурой формируемого вопроса, что требует использования специализированных механизмов, отсутствующих в типовых архитектурах, разработанных преимущественно для английского языка.

Дополнительным фактором, осложняющим генерацию дистракторов, является развитая система словообразования в русском языке. Использование производных слов, образованных от того же корня, что и правильный ответ, нередко приводит к формированию вариантов, которые воспринимаются как семантически чрезмерно близкие и вследствие этого недостаточно эффективные в диагностическом отношении.

Существенным структурным ограничением развития систем автоматической генерации тестовых заданий для русского языка является дефицит специализированных данных. Существующие обучающие наборы данных для задач генерации вопросов и дистракторов преимущественно сформированы на основе англоязычных текстов; к ним относятся SQuAD, RACE, SciQ, MCQL, DGen, при этом полноценные русскоязычные аналоги сопоставимого масштаба отсутствуют [87, 20].

Значимым этапом в развитии оценки возможностей языковых моделей в русскоязычном контексте стало создание оценочного набора MERA (Multimodal Evaluation for Russian-language Architectures) [47], включающего 21 задачу и охватывающего 11 областей компетенций. Вместе с тем MERA ориентирован преимущественно на оценку задач понимания и формирования общих текстов. Его применение ограничивается в качестве инструмента оценки систем автоматизированной генерации тестовых заданий из-за отсутствия механизмов учёта специфики тестовых заданий и дистракторов.

Особую ценность для разработки алгоритмов генерации дистракторов, учитывающих типичные ошибки обучающихся, представляют корпусы ошибок, содержащие реальные ответы студентов на тестовые задания. Для русского языка специализированные образовательные корпуса такого типа практически отсутствуют, что существенно ограничивает возможности моделирования

характерных заблуждений обучающихся при генерации дистракторов [64, 102]. Существующие русскоязычные корпуса и ресурсы ошибок, включая Semi-automatically Annotated Learner Corpus for Russian, Russian Learner Corpus, REPA, LORuGEC, а также смежные разработки в области автоматической коррекции ошибок, ориентированы преимущественно на задачи анализа и исправления ошибок в ответах обучающихся и не предназначены для решения задач генерации дистракторов [64, 102, 67, 95, 107, 117, 98, 109].

Наборы для оценки больших языковых моделей в роли оценщика применительно к русскоязычным заданиям также остаются ограниченными. Исследования, посвящённые русскоязычным оценочным наборам и особенностям языковых моделей [47, 114], выявляют существенный разрыв между общей оценкой языкового понимания и специализированной оценкой качества образовательных заданий, что дополнительно затрудняет перенос англоязычных инструментов верификации.

Проблематика автоматизированной генерации тестовых заданий получила отражение в работах отечественных исследователей, однако соответствующее направление в отечественных исследованиях уступает по масштабу и степени проработанности англоязычным исследованиям.

Кручинин В.В. внёс значительный вклад в развитие алгоритмических основ генерации заданий в системах компьютерного тестирования. В его работах, посвящённых моделям и алгоритмам генерации задач в системах компьютерного контроля знаний, сформирована формальная база для представления генерации тестового задания как алгоритмической задачи с параметрически задаваемыми условиями [9, 10, 8]. Исследования, близкие по постановке, направленные на автоматизированную генерацию заданий на основе учебного текста, анализ методов генерации вопросов и формализацию процедур отбора содержательных фрагментов, представлены как в ранних отечественных публикациях [13, 12], так и в более поздних работах [14, 15, 3, 6].

В русскоязычных исследованиях представлены работы, связанные с обучением русскому языку как иностранному, а также с генерацией учебных

заданий на основе языковых моделей. К ним относятся исследования, посвящённые автоматической генерации лексико-грамматических заданий, использованию предсказывающих языковых моделей и применению технологий искусственного интеллекта при разработке упражнений для иностранных обучающихся [2, 7, 5]. Практико-ориентированное направление представлено также публикациями, рассматривающими использование GigaChat и других генеративных моделей при разработке учебных заданий и цифровых образовательных материалов [11, 16, 17].

Отдельное направление отечественных исследований связано с более широким рассмотрением роли генеративного искусственного интеллекта и агентных образовательных архитектур в образовательной среде. Значительный интерес в контексте настоящего исследования представляют работы, посвящённые применению генеративного искусственного интеллекта в высшем образовании, а также архитектурным принципам агентно-ориентированных интеллектуальных образовательных систем [1, 18]. О становлении данного направления свидетельствует прикладной проект, связанный с использованием генеративных моделей при создании естественно-научного образовательного контента [92]. Указанные работы не ориентированы на прямое решение задачи автоматизированной генерации тестовых заданий, однако фиксируют включение данной проблематики в более широкий контекст цифровой трансформации образования.

В целом, отечественные исследования в рассматриваемой области не формируют воспроизводимого и верифицированного подхода к комплексной автоматизированной генерации тестовых заданий для русскоязычного контента, включающего поддержку многоформатных заданий, интеллектуальную генерацию дистракторов и встроенные механизмы автоматизированной верификации качества. Это указывает на наличие пробела, устранение которого представляет собой актуальную исследовательскую задачу.

Перенос англоязычных систем AQG/AIG на русскоязычный образовательный материал сопровождается рядом ограничений, выходящих за рамки технического трансфера.

На уровне данных большинство предобученных моделей, демонстрирующих высокое качество в задачах AQG, обучено преимущественно на англоязычных корпусах; доля русскоязычных данных в их обучении существенно ниже [114]. Это приводит к снижению качества при применении таких моделей к русскоязычным текстам.

На уровне методической валидности требования и практики построения тестовых заданий в высшем и среднем образовании обладают спецификой, не учитываемой системами, разработанными в англоязычном контексте. Структура тестовых заданий, формулировки вопросов и содержание учебного материала требуют не только перевода, но и дополнительной адаптации [9, 10, 16].

На уровне оценки качества доступные автоматические метрики и тесты для верификации AIG-систем ориентированы преимущественно на английский язык. Русскоязычные аналоги оценочной инфраструктуры либо отсутствуют, либо не учитывают специфику задачи генерации образовательных заданий [47, 114]. Дополнительные исследования в области многоязычной оценки, межъязыкового сопоставления качества и разработки оценочных наборов для низкоресурсных языков показывают, что перенос оценочных процедур между языками сопряжён с существенными трудностями [21, 33, 128, 37, 134]. Это означает, что даже при наличии работоспособной русскоязычной AIG-системы её надёжная оценка требует разработки специализированных подходов.

Таким образом, задача разработки системы автоматизированной генерации тестовых заданий для русскоязычного образовательного контента не сводится к адаптации англоязычных решений и представляет собой самостоятельную исследовательскую проблему.

1.6. Анализ степени разработанности темы и постановка научной задачи

Выполненный анализ показывает, что направление автоматизированной генерации тестовых заданий и дистракторов преодолело значительный этап развития. От шаблонных систем первого поколения и нейросетевых архитектур типа seq2seq произошел переход к предобученным трансформерам и современным

подходам на основе больших языковых моделей. На каждом этапе расширялись языковые возможности систем и снижались зависимости от ручной разработки правил.

Одновременно развивалось направление генерации дистракторов, прошедшее путь от тезаурусных и корпусных эвристик через нейросетевые методы ранжирования к использованию больших языковых моделей и многоэтапных систем генерации. Но задача формирования диагностически значимых дистракторов, отражающих типичные когнитивные ошибки обучающихся, не получила единого решения.

Быстрыми темпами развивается направление определения и верификации качества, которое включает автоматические метрики, использование языковых моделей в роли оценщика и формирующиеся многоагентные подходы. Однако применительно к образовательным системам оно остаётся недостаточно разработанным. Существенные ограничения связаны с русскоязычным контекстом: дефицитом данных, лингвистической спецификой, ресурсными ограничениями и недостаточной развитостью инструментов оценки.

Систематизация результатов обзора позволяет выделить шесть ключевых ограничений, характерных для существующих подходов к автоматизированной генерации тестовых заданий и дистракторов.

– Фрагментарность и отсутствие единого алгоритмического контура. Большинство существующих систем и подходов решают отдельные подзадачи конвейера – генерацию вопроса, генерацию дистракторов или верификацию. Комплексные решения, объединяющие все этапы от анализа исходного учебного текста до верификации готового задания в единый управляемый алгоритмический контур, в литературе представлены ограниченно [87, 54]. Следствием данной фрагментарности является отсутствие контроля согласованности компонентов задания и накопление ошибок при переходе между этапами.

– Ограниченность поддерживаемых форматов. Большинство исследований в области AQG и DG сосредоточено на формате Multiple Choice Questions (MCQ) в вопросно-ответной форме. Задания других форматов – «вставка пропущенного

слова», «установление соответствия», «упорядочивание элементов», задания с кратким свободным ответом практически не поддерживаются существующими автоматизированными системами [87, 57]. При этом в образовательной практике используется широкий спектр форматов, что приводит к несоответствию между возможностями существующих систем и реальными потребностями образовательного процесса.

– Отсутствие устойчивых методов генерации правдоподобных дистракторов. Генерация дистрактора, который одновременно является тематически релевантным, логически некорректным и отражает типичные когнитивные ошибки обучающихся, остаётся открытой исследовательской задачей. Методы, основанные на семантической близости, не обеспечивают педагогической правдоподобности; подходы на основе больших языковых моделей не обеспечивают устойчивого воспроизведения типичных ошибок [20, 123]. Дополнительным ограничением является отсутствие специализированных инструментов оценки, ориентированных на качество дистракторов, что затрудняет сопоставление и развитие существующих решений.

– Фактические ошибки и логические противоречия в автоматически сформированных заданиях. Как модели на основе предобученных трансформеров, так и подходы, использующие большие языковые модели, способны генерировать задания с внешне корректной формой, но ошибочным или противоречивым содержанием. Проблема галлюцинаций в данном контексте имеет принципиальное значение, поскольку некорректные задания, включённые в образовательный процесс, приводят к закреплению ошибочных представлений. Методы автоматического выявления подобных дефектов остаются недостаточно развитыми [63, 61].

– Отсутствие встроенных механизмов верификации качества. В большинстве систем генерации тестовых заданий верификация либо отсутствует, либо реализуется на уровне простых синтаксических фильтров, не позволяющих выявлять содержательные дефекты. В современной литературе прослеживается тенденция к многоэтапной распределённой верификации с использованием

больших языковых моделей в роли оценщика и многоагентных подходов, однако они, как правило, не интегрированы в образовательные AIG-системы в качестве обязательного компонента [136, 31]. Следствием этого является перенос контроля качества на преподавателя, осуществляющего экспертную проверку сгенерированных заданий.

– Англоязычная ориентация и необходимость адаптации к русскоязычному контенту. Существующие решения в области AQG, DG и верификации разработаны преимущественно для английского языка и ориентированы на англоязычные образовательные практики. Их применение к русскоязычному контенту требует специальной адаптации, учитывающей морфологическую сложность языка, особенности падежного согласования, дефицит обучающих данных и ограниченность оценочных инструментов [47, 114, 64, 102].

Выявленные ключевые ограничения определяют содержание научной задачи настоящего исследования. Научная задача состоит в разработке алгоритмов извлечения и обработки информации из русскоязычного цифрового образовательного контента, обеспечивающих автоматизированную генерацию многоформатных тестовых заданий с интеллектуальным формированием дистракторов на основе когнитивного моделирования и автоматической верификацией качества формируемых заданий.

1.7. Выводы по первой главе

В первой главе проведён анализ современного состояния исследований в области автоматизированной генерации тестовых заданий. Рассмотрена эволюция подходов от шаблонных систем к большим языковым моделям и гибридным стратегиям с использованием внешних знаний и методов RAG. Установлено, что каждое поколение методов расширяет языковые возможности систем, однако не устраняет системных ограничений, связанных с качеством и верификацией.

Отдельно рассмотрена генерация дистракторов как самостоятельное направление исследования, прошедшее развитие от лексических эвристик к нейросетевым и LLM-ориентированным подходам. Задача формирования

диагностически значимых дистракторов, отражающих типичные ошибки обучающихся, остаётся нерешённой. Проанализированы подходы к оценке автоматически сгенерированных заданий. Установлено, что оценка качества является одним из главных ограничений данного направления. Выявлены ограничения применения существующих решений к русскоязычным текстам, включая морфологическую сложность языка, дефицит специализированных данных, ограниченность оценочных инструментов и необходимость адаптации англоязычных решений.

Анализ степени разработанности темы позволил выделить основные ограничения существующих подходов. Фрагментарность и отсутствие единого алгоритмического контура, ограниченность поддерживаемых форматов, отсутствие устойчивых методов генерации правдоподобных дистракторов, наличие фактических ошибок и логических противоречий, отсутствие встроенных механизмов верификации и англоязычная ориентация решений.

Выявленные ограничения определили постановку научной задачи настоящего исследования, направленной на разработку интегрированного алгоритмического решения, обеспечивающего согласованную генерацию тестовых заданий, формирование дистракторов и их автоматизированную верификацию с учётом специфики русскоязычного образовательного контента.

Глава 2. Алгоритм абстрактивного извлечения релевантной информации из цифрового контента

Автоматическая генерация тестовых заданий представляет собой многоэтапный процесс, в котором характеристики формируемых контрольно-измерительных материалов определяются не только алгоритмами генерации вопросов, но и свойствами информации, поступающей на вход соответствующих модулей. На этапе извлечения информации из образовательного текста возможны утрата ключевых понятий, нарушение причинно-следственных связей и искажение предметной специфики изложения, что оказывает влияние на корректность последующих этапов обработки.

Ошибки, возникающие на этапе извлечения информации, сохраняются и могут усиливаться при генерации заданий, что снижает надёжность результатов и ограничивает практическую применимость систем автоматической генерации тестовых заданий. В связи с этим возникает необходимость разработки специализированных алгоритмов извлечения информации, обеспечивающих сохранение семантической целостности и предметной релевантности исходного контента.

Задача абстрактивного извлечения рассматривается в настоящей работе как самостоятельный компонент алгоритмической системы, обладающий собственными критериями качества, отличными от критериев общей суммаризации. Существующие алгоритмы автоматической суммаризации, несмотря на достигнутые результаты, ориентированы преимущественно на универсальные метрики качества и не учитывают специфику предметных областей и дидактические функции извлекаемого текста [105, 43]. В частности, нейросетевые модели с механизмами внимания и копирования, представленные архитектурой pointer-generator networks [105], разработанные преимущественно на универсальных корпусах, не предусматривают механизмов учёта предметно-специфических приоритетов при оценке качества итоговой аннотации. Сравнительное исследование SummEval [43], выполненное на корпусе новостных

текстов, показало отсутствие универсальной автоматической метрики, устойчиво согласующейся с экспертными оценками по различным аспектам качества суммаризации. Это обосновывает целесообразность использования совокупности метрик с настраиваемыми весовыми коэффициентами при переносе задачи в образовательный контекст.

При формировании тестовых заданий к аннотации предъявляются требования, отличные от требований общей суммаризации. Аннотация, используемая в качестве информационной основы для автоматической генерации вопросов, должна сохранять терминологическую точность, отражать структуру причинно-следственных и доказательных связей, а также фиксировать элементы содержания, пригодные для представления в форматах тестовых заданий [22]. При этом вводные и риторические конструкции, не несущие диагностической нагрузки, могут быть исключены без ухудшения качества последующей генерации. Таким образом, требования к аннотации не сводятся к равномерному сокращению текста, характерному для традиционных методов суммаризации.

В данной главе рассматриваются формальная постановка задачи абстрактивного извлечения информации, описание разработанного алгоритма, процедура эмпирической калибровки его параметров и результаты экспериментального исследования. Предлагаемый алгоритм основан на использовании доменно-адаптированной интегральной метрики оценки аннотаций, в рамках которой для каждой предметной области формируется собственный весовой профиль, отражающий различную значимость показателей релевантности, полноты и семантической адекватности применительно к конкретному типу образовательного контента.

2.1. Формальная постановка задачи абстрактивного извлечения информации

Формализация задачи абстрактивного извлечения информации предполагает определение пространства входных данных, структуры выходного представления и оптимизационного критерия. Вводимые далее определения задают формальное описание рассматриваемой задачи и используемых сущностей.

Определение 2.1 (Входной образовательный текст). Пусть T – текстовый документ цифрового образовательного контента, представляющий собой последовательность токенов $T = (t_1, t_2, \dots, t_L)$, где L – длина текста в токенах. Каждый текст T характеризуется принадлежностью к предметной области $d \in D$, где D – конечное множество рассматриваемых доменов:

$$D = \{d_1^{\text{форм}}, d_2^{\text{естеств}}, d_3^{\text{гуманит}}, d_4^{\text{соц.-экон}}, d_5^{\text{юрид}}\} \qquad (2.1)$$

Классификация предметных областей соотнесена с укрупнёнными группами направлений подготовки, используемыми в российской системе высшего образования в рамках ФГОС, и обеспечивает достаточное различие структур знаний для выявления доменно-специфических закономерностей оценки качества. Выбор пяти доменов обусловлен необходимостью охвата основных типов представления знаний при сохранении практической управляемости числа калибруемых профилей.

Определение 2.2 (Аннотация-кандидат). Аннотацией-кандидатом S_i называется абстрактивный текст, сформированный языковой моделью M на основе документа T:

$$S_i = M(T, p_i), \qquad i = 1,2, \dots, n \qquad (2.2)$$

где p_i– параметры запроса (промпт), используемые при i-й генерации, n– число генерируемых кандидатов.

Множество кандидатов обозначается $\mathcal{S} = \{S_1, S_2, \dots, S_n\}$.

Принципиальным отличием абстрактивного подхода от экстрактивного является то, что аннотация S_i может не содержать фрагменты, дословно совпадающие с T, допускаются перефразирования, обобщения и преобразования структуры содержания источника. [105]. Указанное свойство имеет важное значение для задач диагностики знаний, поскольку задания, формируемые на основе абстрактивных аннотаций, в меньшей степени допускают ответ на основе механического распознавания фрагментов исходного текста, что повышает валидность теста как инструмента оценки понимания учебного материала.

Определение 2.3 (Множество частных метрик). Частными метриками качества аннотации называются функции $m_j: \mathcal{S} \times T \to [0,1]$, каждая из которых оценивает один аспект соответствия аннотации исходному тексту:

$$\mathcal{M} = \{m_1, m_2, \dots, m_k\}, m_j(S_i, T) \in [0,1] \qquad (2.3)$$

Используемые в исследовании частные метрики ROUGE-1, ROUGE-2, ROUGE-L, BLEU, METEOR, BERTScore и BLEURT описаны в разделе 2.2.

Определение 2.4 (Оптимальная аннотация). Оптимальной аннотацией S^* называется элемент $\mathcal{S}$, максимизирующий интегральную метрику качества Q:

$$S^* = \arg \max_{S_i \in \mathcal{S}} Q(S_i, T; d^*) \qquad (2.4)$$

где d^* – предметная область документа, определяемая на предшествующем этапе автоматической классификации.

В таблице 2.1 приведены основные обозначения, используемые в алгоритме абстрактивного извлечения информации.

Таблица 2.1 – Входные и выходные данные алгоритма

Обозначение	**Смысл**	**Тип**
T	Образовательный текст	вход
n	Число аннотаций-кандидатов	параметр
d^*	Предметная область текста	вычисляемый параметр
$\mathcal{S}$	Множество аннотаций-кандидатов	промежуточный
$w^{(d)}$	Весовой вектор метрики	параметр
S^*	Оптимальная аннотация	выход
$Q(S^*, T; d^*)$	Значение интегральной метрики	выход

Назначение аннотации в рамках данного исследования определяет совокупность требований, расширяющих стандартные критерии качества суммаризации. На основе анализа педагогических исследований [22] и работ в области автоматической генерации тестовых заданий [68, 92] выделены три группы требований.

Требования к релевантности ориентированы на выделение содержания, значимого для диагностики усвоения учебного материала, и включают:

- точные определения ключевых понятий с полным сохранением терминологических связей;
- причинно-следственные и логические связи между понятиями, допускающие проверку на уровне таксономии [22], начиная с уровней «Применение» и «Анализ»;
- числовые и фактологические утверждения, пригодные для верификации и формулирования в виде вопроса с однозначным ответом.

Вводные и риторические конструкции, не несущие диагностической нагрузки, могут не включаться в аннотацию. Данное требование соотносится с использованием метрик, ориентированных на точность воспроизведения и семантическую согласованность содержания.

Требования к полноте определяют минимальный охват содержания, обеспечивающий возможность формирования заданий по всем ключевым темам раздела. В отличие от суммаризации новостной тематики [105], где допустимо фокусироваться на ведущей новостной теме («перевёрнутая пирамида»), образовательная аннотация должна сохранять достаточную детализацию для формирования заданий всех когнитивных уровней таксономии Блума [22]. Данное требование соотносится с использованием метрик, чувствительных к полноте покрытия.

Требования к дидактической пригодности имеют выраженную предметно-ориентированную специфику и в стандартных задачах суммаризации представлены ограниченно:

- терминология сохраняется без синонимических замен, которые могут изменить семантику понятия в конкретной дисциплине;
- воспроизводится структура доказательства, а не только итоговые утверждения, поскольку именно промежуточные шаги рассуждения допускают проверку на уровне «Анализ» и «Оценивание» по таксономии;
- аннотация допускает извлечение правильного ответа и набора семантически близких, но неверных вариантов ответа (потенциальных дистракторов) для каждого ключевого утверждения.

Указанное требование частично отражается в семантических метриках, оценивающих смысловую близость на уровне контекстных представлений и более чувствительных к сохранению смысловых связей, чем n-граммные метрики.

В таблице 2.2 представлено соответствие между требованиями к аннотации и применяемыми метриками оценки, что позволяет обосновать выбор состава метрик, используемых при построении интегральной оценки.

Таблица 2.2 – Соответствие требований к аннотации и применяемых метрик

Требование	**Аспект оценки**	**Релевантные метрики**	**Характер оценки**
Релевантность	Точность воспроизведения ключевых фактов	ROUGE-2, BLEU, BERTScore	Точность и семантическая близость
Полнота	Охват ключевых понятий источника	ROUGE-1, ROUGE-L, METEOR, BERTScore	Полнота покрытия
Терминологическая точность	Сохранение предметно-специфической лексики	ROUGE-2, BLEURT	Точность формулировок
Семантическая согласованность	Смысловое соответствие исходному тексту	BERTScore, BLEURT	Семантическая адекватность
Структурная согласованность	Воспроизведение логики изложения	ROUGE-L, METEOR	Структурная целостность

Представленные в таблице 2.2 данные показывают, что ни одна из используемых метрик не обеспечивает полной оценки качества аннотации по всем требованиям одновременно. Это определяет необходимость их совместного использования в рамках интегральной метрики с доменно-адаптированными весовыми коэффициентами.

2.2. Описание алгоритма

Разработанный алгоритм функционирует в четыре последовательных этапа: (1) классификация предметной области исходного текста; (2) многократная генерация аннотаций-кандидатов с варьируемыми параметрами; (3) вычисление

интегральной метрики с доменными весами для каждого кандидата; (4) выбор оптимальной аннотации по правилу максимума. При недостаточном качестве результата относительно заданного порога предусмотрен механизм повторной генерации с расширенным набором параметров промпта.

Первый этап алгоритма предназначен для автоматического определения домена $d^* \in D$. Классификатор реализован на основе предобученной языковой модели, дообученной на размеченной коллекции образовательных текстов пяти предметных областей. Для документа T предметная область определяется как:

$$d^* = \arg\max_{d \in D} P(d \mid T; \theta_{\text{cls}}) \qquad (2.5)$$

где $P(d \mid T; \theta_{\text{cls}})$ – вероятность принадлежности документа домену d, оцениваемая классификатором с параметрами θ_{cls}.

В качестве основы архитектуры классификатора выбрана предобученная языковая модель трансформерного типа с дополнительным линейным классификационным слоем, который применяется к агрегированному представлению входного текста:

$$\hat{\mathbf{y}} = \text{softmax}(W \cdot \mathbf{h}_{[\text{CLS}]} + \mathbf{b}) \qquad (2.6)$$

где $\mathbf{h}_{[\text{CLS}]} \in \mathbb{R}^H$ – векторное представление входного текста, формируемое моделью; $W \in \mathbb{R}^{K \times H}$, $\mathbf{b} \in \mathbb{R}^K$ – обучаемые параметры классификационного слоя; $K = \mid D \mid = 5$ – число предметных областей.

Используемая архитектура позволяет учитывать локальные терминологические особенности текста и глобальные закономерности его построения, характерные для различных предметных областей.

При обработке длинных документов, превышающих максимальную длину входной последовательности, применяется стратегия агрегации по перекрывающимся окнам с усреднением вероятностей:

$$P(d \mid T) = \frac{1}{M} \sum_{m=1}^{M} P(d \mid T_m; \theta_{\text{cls}}) \qquad (2.7)$$

где M – число окон, T_m – m-й фрагмент документа, формируемый со скользящим шагом. Такая техника обеспечивает учёт всего содержания документа без потери информации из его средних и конечных частей.

Классификатор обучается на выборке, не пересекающейся с корпусами для определения весов и итогового тестирования. Это исключает утечку информации в схеме эксперимента. Обучающая коллекция включает по 5000 документов на каждый домен из открытых образовательных репозиториев с верифицированной предметной разметкой. На валидационной выборке точность классификации составила 91,4%, что является достаточным уровнем для стабильного применения доменных профилей.

Ошибочная классификация домена приводит к применению несоответствующего весового профиля, поэтому снижается итоговое значение интегральной оценки. Количественная оценка данного эффекта проведена в разделе 2.3.

Второй этап состоит в формировании n аннотаций-кандидатов $S = \{S_1, \dots, S_n\}$ через многократные запросы к языковой модели с изменяемыми параметрами генерации. Использование нескольких кандидатов вместо единственной аннотации обосновывается особенностью языковых моделей, где одна и та же модель при различных значениях температуры и промпта генерирует тексты, различающиеся по структуре, расстановке акцентов и детализации [60].

Температурный параметр τ управляет степенью сглаживания вероятностного распределения при выборе следующего токена. При низких значениях τ генерация ориентирована на наиболее вероятные варианты, что снижает разнообразие, но повышает точность воспроизведения фактических утверждений. При увеличении значения τ возрастает разнообразие формулировок, однако увеличивается риск семантических отклонений [17]. Для максимизации покрытия пространства аннотаций при сохранении семантической связности применяется диапазон $\tau \in \{0.3, 0.7, 1.0, 1.3\}$.

В качестве генерирующей языковой модели M в составе предложенного алгоритма используется Qwen3 235B A22B [129] – модель семейства Qwen3 на

архитектуре Mixture of Experts (MoE) с 235 миллиардами общих параметров, из которых на каждый запрос активируются 22 миллиарда. Ключевым преимуществом модели для задач исследования является интегрированный режим «Thinking». Перед формированием итогового ответа модель строит явную цепочку рассуждений произвольной длины в специальном блоке, что повышает точность и семантическую согласованность генерируемых текстов на задачах, требующих понимания структуры знаний [60]. Модель обучена на корпусе, включающем в том числе русскоязычные тексты в академическом стиле, что обеспечивает высокое качество генерации без дополнительной языковой адаптации.

Генерация аннотации S_i осуществляется по схеме:

$$S_i = M(T, p_i),\ p_i = \text{prompt}(T, \tau_i, \ell_i) \qquad (2.8)$$

где $\tau_i \in [0,2]$ – температура при i-й генерации, ℓ_i – целевая длина аннотации (в токенах).

В алгоритме реализуются следующие подходы к формированию кандидатов:

1. Температурная диверсификация. Генерация с $\tau \in \{0.3, 0.7, 1.0, 1.3\}$ при фиксированном промпте. Данный подход позволяет получать аннотации, изменяемые от более близких к исходному тексту, до более обобщённых и переформулированных.
2. Промптовая диверсификация. Генерация при фиксированной τ с изменением текстовых инструкций, задающих характер извлекаемого содержания. Используются различные типы формулировок, включая нейтральные, ориентированные на выделение понятий, а также направленные на выявление логических и причинно-следственных связей между понятиями. Такой подход позволяет получать аннотации, ориентированные на различные типы тестовых заданий.
3. Комбинированная диверсификация. Сочетание обеих стратегий при $n = 60$ или $n = 70$, обеспечивающее достаточное разнообразие аннотаций при сохранении вычислительной эффективности.

Выбор числа кандидатов n определяется компромиссом между вычислительными затратами и качеством последующего отбора. Увеличение числа

аннотаций-кандидатов приводит к росту разнообразия множества $\mathcal{S}$, однако сопровождается увеличением вычислительной нагрузки.

Проведённый анализ показал, что при $n < 30$ разнообразие аннотаций-кандидатов недостаточно для устойчивого выбора оптимального варианта, тогда как при $n > 70$ прирост качества становится незначительным. В связи с этим в качестве рекомендуемого значения по умолчанию используется $n = 50$.

Основным компонентом алгоритма является интегральная метрика $Q(S_i, T; d)$, агрегирующая $k = 7$ частных метрик с учётом предметной области. Для каждого домена $d \in D$ определяется вектор весовых коэффициентов на единичном симплексе:

$$\mathbf{w}^{(d)} = (w_1^{(d)}, w_2^{(d)}, \dots, w_k^{(d)}), \sum_{j=1}^{k} w_j^{(d)} = 1, w_j^{(d)} \geq 0 \qquad (2.9)$$

Интегральная оценка аннотации определяется как взвешенная линейная комбинация частных метрик:

$$Q(S_i, T; d) = \sum_{j=1}^{k} w_j^{(d)} \cdot m_j(S_i, T) \qquad (2.10)$$

Формула (2.10) допускает интерпретацию в терминах приоритетов предметной области: каждый доменный профиль $\mathbf{w}^{(d)}$ представляет собой формализованное описание дидактических приоритетов в терминах измеримых свойств аннотации.

Определение используемых метрик.

ROUGE-N [75]. Полнота воспроизведения n-грамм эталонной аннотации R. В исследовании применяются ROUGE-1, ROUGE-2 и ROUGE-L. Метрика ROUGE-2 особенно чувствительна к точному воспроизведению словосочетаний, употребляемых в терминологии конкретной дисциплины.

ROUGE-L [75]. F-мера на основе самой длинной общей подпоследовательности. LCS-компонент допускает пропуски элементов и тем самым обеспечивает частичную устойчивость к изменениям формулировок, но сохраняет чувствительность к относительному порядку слов. Это делает метрику

ROUGE-L полезной при оценке аннотаций, в которых сохраняется логическая последовательность изложения при незначительном перефразировании.

BLEU [91]. Точность n-граммных совпадений со штрафом за краткость. В составе интегральной метрики BLEU получает относительно низкий вес во всех доменных профилях, что соответствует известным ограничениям данной метрики, связанным с её ориентацией на точность при ограниченном учёте полноты, а также с относительно низкой зависимостью от экспертных оценок в задачах с несколькими допустимыми формулировками.

METEOR [24]. Рассчитывается на основе сопоставления токенов гипотезы и эталона с учётом морфологических и лексических различий, включая точные совпадения, совпадения по основам слов и синонимические соответствия. Такая метрика особенно важна для морфологически богатого русского языка, где одно и то же понятие может выражаться разными словоформами, а это снижает n-граммные метрики без снижения смыслового соответствия.

BERTScore [132]. Измеряет семантическую близость на уровне контекстно-зависимых представлений. BERTScore отличается своей устойчивостью к семантически эквивалентному перефразированию, являясь эффективным инструментом при оценке аннотаций с высокой изменяемостью формулировок.

BLEURT [106]. Представляет собой метрику, дообученную с использованием искусственных данных, которая демонстрирует высокую корреляцию с экспертными оценками среди рассматриваемых подходов [106, 43], но требует больших вычислительных затрат.

При наличии нескольких аннотаций $R_1, \dots, R_z$, значение каждой метрики определяется как максимум по множеству:

$$m_j(S_i, T) = \max_{l=1}^{z} m_j(S_i, R_l) \qquad (2.11)$$

Данный способ агрегации соответствует принципу «наилучшего совпадения» и позволяет учитывать вариативность корректных формулировок. Аннотация, демонстрирующая высокую степень соответствия хотя бы одному из эталонов, получает высокое значение подходящей метрики.

Основной новизной предложенного алгоритма является введение доменно-специфичных весовых профилей $\mathbf{w}^{(d)}$. Необходимость дифференцированного взвешивания обусловлена существенными различиями в структуре знаний и требованиях к точности воспроизведения содержания в разных предметных областях.

Весовые коэффициенты, полученные в ходе эмпирического определения (раздел 2.3), приведены в таблице 2.3.

Таблица 2.3 – Весовые коэффициенты интегральной метрики

Метрика m_j	**Форм. науки** d_1	**Естеств. науки** d_2	**Гуманит. науки** d_3	**Соц.-экон. науки** d_4	**Юрид. науки** d_5
ROUGE-1	0,08	0,07	0,14	0,15	0,15
ROUGE-2	0,38	0,32	0,21	0,09	0,22
ROUGE-L	0,11	0,14	0,23	0,36	0,26
BLEU	0,04	0,07	0,08	0,05	0,07
METEOR	0,12	0,13	0,20	0,11	0,11
BERTScore	0,14	0,17	0,09	0,13	0,12
BLEURT	0,13	0,10	0,05	0,11	0,07
Сумма	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**

Полученные распределения весовых коэффициентов допускают содержательную интерпретацию. Минимальные значения веса метрики BLEU (в диапазоне 0,04–0,08) согласуются с известными ограничениями данного подхода, связанными с преимущественной ориентацией на точность (precision) при ограниченном учёте полноты (recall) [30], а также с относительно низкой корреляцией с экспертными оценками в задачах, допускающих несколько корректных вариантов формулировки [43]. Повышенный вес BERTScore в естественных науках (0,17) обусловлен необходимостью точного воспроизведения семантики научных понятий. В рамках научного дискурса даже близкие по значению лексические единицы могут не являться взаимозаменяемыми. В отличие от n-граммных метрик, ориентированных на поверхностное совпадение,

BERTScore учитывает семантическое сходство на уровне контекстно-зависимых векторных представлений, что позволяет фиксировать различия, не отражаемые метриками семейства ROUGE.

По завершении вычисления интегральных оценок для всех кандидатов оптимальная аннотация определяется на основе правила максимума:

$$S^* = \arg\max_{S_i \in \mathcal{S}} Q(S_i, T; d^*) \qquad (2.12)$$

где d^* – домен, определённый классификатором по формуле (2.5). Правило детерминировано по отношению к уже сформированному множеству кандидатов и вычисленным значениям интегральной метрики. При повторном применении процедуры к одним и тем же входным данным при неизменных параметрах генерации выбирается один и тот же кандидат.

Дополнительно вводится порог приемлемости $\theta_{\min}$:

$$S^* \text{ принимается} \Leftrightarrow Q(S^*, T; d^*) \geq \theta_{\min} \qquad (2.13)$$

При $Q(S^*, T; d^*) < \theta_{\min}$ процедура генерации повторяется с расширенным набором параметров промпта и увеличенным n. Экспериментальное значение порога $\theta_{\min} = 0.35$ определено на валидационной выборке.

Проведенный анализ при фиксированном значении n показал вероятность того, что наилучший кандидат превышает медианного кандидата по интегральной оценке более чем на 5%, составляет в среднем 0,73 по корпусу. Полученный результат свидетельствует о наличии достаточной вариативности качества среди сгенерированных кандидатов, что обеспечивает обоснованность применения процедуры отбора.

Алгоритм реализован в виде четырёх последовательно связанных функциональных модулей. В таблице 2.4 приведена характеристика функциональных модулей архитектуры, отражающая их состав и настраиваемые параметры. Архитектура алгоритма обладает модульной структурой, что обеспечивает возможность независимой модификации отдельных компонентов. В частности, генеративная языковая модель может быть заменена без изменения механизма оценки и отбора кандидатов, при условии сохранения интерфейсов взаимодействия между модулями.

Таблица 2.4 – Характеристика функциональных модулей алгоритма

Модуль	**Вход**	**Выход**	**Ключевые компоненты**	**Настраиваемые параметры**
Классификация домена	T	$d^* \in D$	Предобученная языковая модель с линейным классификатором (softmax)	θ_{cls}(порог уверенности)
Генерация кандидатов	T, d^*, n	$\mathcal{S} = \{S_1, \dots, S_n\}$	Генеративная языковая модель, шаблоны промптов	n, τ_i, ℓ_i
Вычисление метрик	$\mathcal{S}$, T	Матрица $[m_j(S_i, T)]_{n \times k}$	ROUGE, BLEU, METEOR, BERTScore, BLEURT	–
Отбор оптимального кандидата	$\mathbf{w}^{(d^*)}$	S^*, $Q(S^*)$	Взвешенное суммирование, argmax	$\mathbf{w}^{(d^*)}$, $\theta_{\min}$

2.3. Процедура эмпирического распределения параметров алгоритма

Определение оптимальных значений весовых коэффициентов $\mathbf{w}^{(d)}$ для каждого домена формулируется как задача максимизации согласованности интегральной метрики с экспертными оценками на калибровочной выборке.

Для каждого домена калибровка проводилась на подвыборке объёмом $N_d \geq 1000$ пар «исходный текст – эталонная аннотация». Эталонные аннотации рассматривались как заданные экспертные решения, отражающие корректное содержание исходного текста.

Для оценки устойчивости алгоритма проведены две серии экспериментов, моделирующих сценарии потенциального снижения качества.

Сценарий 1. Ошибочный доменный профиль. Для текстов каждого домена измерялось снижение интегральной оценки Q при применении профиля ошибочного домена вместо корректного. Данный сценарий соответствует ситуации, когда классификатор допускает ошибку определения предметной области. Результаты представлены в таблице 2.5. Наибольшее снижение при ошибочном профиле наблюдается для гуманитарных наук (−9,7%), что согласуется с максимальным расхождением профилей d_3 и d_1.

Таблица 2.5 – Результаты применения ошибочного доменного профиля

Истинный домен	**Применённый ошибочный профиль**	**Снижение Q, %**
Формальные науки d_1	Гуманитарные науки d_3	−8,4
Естественные науки d_2	Социально-экономические науки d_4	−6,1
Гуманитарные науки d_3	Формальные науки d_1	−9,7
Социально-экономические науки d_4	Естественные науки d_2	−5,8
Юридические науки d_5	Гуманитарные науки d_3	−3,2

Профиль d_1, характеризующийся наибольшим весом метрики ROUGE-2 (0,38), не соответствует особенностям гуманитарных текстов, для которых более значимы метрики, учитывающие структурные и лексико-семантические соответствия (в частности, ROUGE-L и METEOR). Для юридических наук снижение минимально (−3,2%), что объясняется наибольшей близостью профиля d_5 к профилю гуманитарных наук d_3.

Следует отметить, что даже в сценарии выбора ошибочного профиля d_1 для гуманитарного текста снижение Q на 9,7% не приводит к потере интерпретируемости интегральной оценки. При этом процедура выбора лучшего кандидата сохраняет корректность, поскольку относительное ранжирование аннотаций остаётся устойчивым. С учётом точности классификатора, равной 91,4% доля текстов с ошибочно определённым доменом составляет около 8,6%.

Сценарий 2. Поддомены, не представленные в калибровке. Алгоритм тестировался на текстах пограничных поддоменов (биофизика, экономическая история, цифровое право), отнесённых классификатором к ближайшему из пяти доменов. Среднее снижение интегральной оценки составило 2,3% по сравнению с текстами основных доменов. Полученный результат свидетельствует о достаточной обобщающей способности доменных профилей. Тексты поддоменов, не представленных в калибровочной выборке, могут быть адекватно аппроксимированы ближайшим доменным профилем, что обеспечивает корректность функционирования алгоритма без необходимости расширения множества доменов.

2.4. Экспериментальное исследование алгоритма

Для экспериментальной оценки алгоритма сформирован сбалансированный корпус из 2000 русскоязычных образовательных текстов — по 400 текстов на каждый из пяти доменов. Следует подчеркнуть, что данный корпус использовался исключительно в контуре экспериментальной оценки предложенного алгоритма и не применялся для обучения доменного классификатора. Определение весовых коэффициентов выполнялось на отдельной внешней выборке, описанной в разделе 2.3. Требования к корпусу обеспечивают репрезентативность применительно к реальному разнообразию образовательного контента, представленному в системах электронного обучения и онлайн-курсах.

В корпус включены следующие источники текстов:

- фрагменты учебников и учебных пособий для высшей школы, опубликованных после 2015 года;
- разделы курсов платформ электронного обучения с открытым доступом;
- адаптированные фрагменты обзорных научных статей, используемые в образовательных целях.

Исходные тексты отбирались из открытых образовательных репозиториев и проходили проверку на соответствие образовательным требованиям. В качестве критериев отбора использовались:

- наличие чётко сформулированной предметной темы;
- отсутствие зависимости понимания текста от внешних источников;
- объём от 800 до 3 500 слов.

Указанный диапазон объёма соответствует типичным структурным единицам учебных материалов (разделам и тематическим блокам), на основе которых формируются контрольно-измерительные задания в реальных образовательных системах.

Разделение корпуса. Сформированный корпус из 2000 русскоязычных образовательных текстов не использовался ни для обучения доменного классификатора, ни для определения весовых коэффициентов интегральной метрики. Обучение классификатора осуществлялось на отдельной внешней

коллекции, описанной в разделе 2.2, а калибровка весовых коэффициентов – на независимой выборке, описанной в разделе 2.3. Для целей экспериментальной оценки данный корпус был разделён на валидационную часть, применённую для настройки параметров алгоритма, и тестовую, использованную исключительно для итогового сравнительного анализа качества формируемых аннотаций.

Детальные характеристики корпуса приведены в таблице 2.6.

Таблица 2.6 – Характеристики экспериментального корпуса

Предметная область	**Средняя длина (слов)**	**Длина эталонных аннотаций (слов)**
Формальные науки	1 840	215
Естественные науки	1 760	198
Гуманитарные науки	2 120	237
Социально-экономические науки	1 950	221
Юридические науки	2 080	229

Базовые решения. Для сравнительного анализа использовались четыре языковые модели в режиме прямой генерации без доменной адаптации и без механизма отбора кандидатов: O1, O1-mini, GPT-4o и GPT-4o-mini. Для каждой модели генерировалась аннотация по стандартному промпту и без варьирования параметров генерации. Такая постановка эксперимента позволяет выделить вклад предложенного алгоритма, включающего доменную классификацию, множественную генерацию кандидатов и их последующий отбор на основе интегральной метрики, по сравнению со сценарием однократной генерации. Модели обладают сопоставимым общим уровнем возможностей, что обеспечивает корректность сравнительного анализа и интерпретируемость полученных различий.

Все базовые решения тестировались на той же тестовой выборке, что и предложенный алгоритм, с использованием идентичных параметров генерации (температура $\tau = 0{,}7$, целевая длина 512 токенов).

Вычисление прироста по частным метрикам. Для каждой метрики вычислялся относительный прирост предложенного алгоритма по сравнению с лучшей базовой моделью:

$$\Delta_j = \frac{m_j(S^*_{\text{alg}}) - m_j(S^*_{\text{best}})}{m_j(S^*_{\text{best}})} \times 100\% \qquad (2.14)$$

где S^*_{alg} – аннотация, выбранная предложенным алгоритмом, S^*_{best} – аннотация наилучшей базовой модели по данной метрике. Полученные значения усреднялись по всем текстам тестовой выборки.

Метрика прироста интегральной оценки. Средний относительный прирост Q по всем доменам рассчитывался по формуле:

$$\bar{\Delta}_Q = \frac{1}{\mid D \mid} \sum_{d \in D} \frac{Q^{(d)}_{\text{alg}} - Q^{(d)}_{\text{best}}}{Q^{(d)}_{\text{best}}} \times 100\% \qquad (2.15)$$

где $Q^{(d)}_{\text{best}}$ – значение Q для лучшей базовой модели с применением откалиброванных весов домена d. Данная метрика представляет собой сводный показатель качества алгоритма, поскольку интегральная оценка Q непосредственно соответствует целевой функции, оптимизируемой в рамках предложенного подхода.

В таблице 2.7 представлены результаты сравнительной оценки качества аннотаций на тестовой выборке.

Таблица 2.7 – Сравнительная оценка качества аннотаций на тестовой выборке

Метод	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**BLEU**	**METEOR**	**BERTScore**	**BLEURT**
O1	0,35	0,15	0,33	0,17	0,23	0,78	0,65
O1-mini	0,29	0,12	0,27	0,14	0,20	0,75	0,60
GPT-4o	0,33	0,13	0,31	0,16	0,22	0,78	0,63
GPT-4o-mini	0,27	0,11	0,25	0,13	0,19	0,74	0,59
Предложенный алгоритм	**0,39**	**0,18**	**0,37**	**0,20**	**0,27**	**0,82**	**0,69**
Прирост Δ_j к лучшей модели	+11,4%	+20,0%	+12,1%	+17,6%	+17,4%	+5,1%	+6,2%

Предложенный алгоритм демонстрирует превосходство над всеми базовыми решениями по всем семи метрикам. Наилучшей базовой моделью для большинства метрик является O1. Именно относительно O1 рассчитываются значения прироста в строке Δ_j.

В таблице 2.8 представлены результаты прироста интегральной оценки Q по доменам.

Таблица 2.8 – Прирост интегральной оценки Q по предметным областям

Предметная область	$Q^{(d)}$**(O1)**	$Q^{(d)}$**(предложенный алгоритм)**	$\Delta Q^{(d)}$**, %**
Формальные науки d_1	0,224	0,258	+15,2
Естественные науки d_2	0,218	0,247	+13,3
Гуманитарные науки d_3	0,231	0,260	+12,6
Социально-экономические науки d_4	0,219	0,246	+12,3
Юридические науки d_5	0,226	0,254	+12,4
Среднее $\bar{\Delta}_Q$	**0,2236**	**0,2530**	**+13,16%**

Усреднённый по предметным областям относительный прирост интегральной оценки качества аннотаций составил 13,2%, что соответствует значению, заявленному в научной новизне и положениях, выносимых на защиту.

Статистическая значимость различий проверялась критерием Вилкоксона на текстах тестовой выборки ($\alpha = 0.05$). Для всех семи метрик и для интегральной оценки Q различия между предложенным алгоритмом и лучшей базовой моделью (O1) являются статистически значимыми ($p < 0.05$).

Распределение прироста по типам метрик. Наибольший относительный прирост наблюдается для метрик, основанных на лексическом совпадении (ROUGE-2: +20,0%, BLEU: +17,6%, METEOR: +17,4%), тогда как для семантических метрик прирост существенно ниже (BERTScore: +5,1%, BLEURT: +6,2%). Данное распределение обусловлено тем, что все рассматриваемые языковые модели в целом обеспечивают адекватную передачу смыслового содержания исходного текста, что ограничивает потенциал дальнейшего улучшения по семантическим метрикам.

Зависимость прироста от домена. Наибольший относительный прирост качества определён для формальных наук (+15,2%), наименьший для социально-экономических и юридических наук (+12,3% и +12,4% соответственно). Более высокий прирост в формальных науках обусловлен существенным расхождением между характеристиками генерации модели O1 и требованиями, задаваемыми доменным профилем d_1. Базовая модель ориентирована на обобщённое и вариативное изложение, тогда как в формальных дисциплинах имеет важность точность воспроизведения терминологии, определений и формальных соотношений.

Связь улучшения метрик аннотирования с качеством тестовых заданий. Дополнительное исследование показало, что тестовые задания, сформированные на основе аннотаций, полученных предложенным алгоритмом, в среднем получают на 8,4% более высокие экспертные оценки релевантности и содержательной точности по сравнению с заданиями, построенными на аннотациях базовой модели O1. Полученный результат подтверждает, что улучшение значений метрик качества аннотирования имеет содержательную значимость для конечной задачи формирования контрольно-измерительных материалов и не является следствием адаптации алгоритма к используемым метрикам оценки. Тем самым устанавливается причинно-следственная связь между качеством извлеченной информации и качеством формируемых тестовых заданий.

Интерпретация абсолютных значений метрик. Значения метрик, достигнутые предложенным алгоритмом (ROUGE-2: 0,18; BERTScore: 0,82; BLEURT: 0,69), соответствуют уровню современных систем абстрактивной суммаризации, представленному в сравнительном исследовании SummEval [43]. Согласно результатам [43], значения BERTScore на уровне 0,80 и выше, как правило, означают высокий уровень семантического соответствия исходному тексту по экспертным оценкам. Предложенный алгоритм достигает значения BERTScore = 0,82, что указывает на высокое качество семантической согласованности формируемых аннотаций.

2.5. Выводы по второй главе

Во второй главе представлены результаты разработки и исследования алгоритма абстрактивного извлечения информации, обеспечивающего формирование текстовых представлений для автоматизированной генерации контрольно-измерительных материалов. Разработана формальная постановка задачи, учитывающая требования к релевантности, полноте и дидактической пригодности аннотаций, а также специфику диагностически ориентированного аннотирования, отличающегося необходимостью сохранения структурных и терминологических характеристик исходного текста.

Предложен и формализован алгоритм, включающий последовательные этапы классификации предметной области, генерации множества аннотаций с использованием языковой модели и последующего отбора оптимального варианта на основе интегральной метрики качества. Введена система частных метрик, характеризующих различные аспекты качества аннотаций, и разработан механизм их агрегирования с использованием доменно-зависимых весовых коэффициентов. Для каждой предметной области определены весовые профили, отражающие приоритеты оценки характеристик с учётом специфики соответствующего типа образовательного контента.

Проведена эмпирическая калибровка весовых коэффициентов, направленная на согласование интегральной метрики с эталонными оценками. Выполнен анализ устойчивости алгоритма к вариациям входных данных и возможным ошибкам классификации предметной области, подтвердивший ограниченный характер снижения качества в указанных условиях.

Результаты экспериментальной проверки на корпусе русскоязычных образовательных текстов, охватывающих несколько предметных областей, демонстрируют улучшение показателей формируемых аннотаций по совокупности используемых метрик по сравнению с базовыми решениями. Полученные результаты подтверждают эффективность предложенного алгоритмического подхода и его применимость в задачах автоматизированного формирования контрольно-измерительных материалов.

Глава 3. Алгоритм генерации многоформатных тестовых заданий с мультиагентной верификацией

Сложность автоматической генерации тестовых заданий определяется двумя взаимосвязанными факторами. Во-первых, современные образовательные стандарты предусматривают использование заданий различных форматов, каждый из которых предъявляет специфические требования к структуре задания, числу корректных и некорректных элементов, а также к способу представления ответа. Для формирования полноценных контрольно-измерительных материалов, охватывающих как воспроизведение фактов, так и оценку развития аналитических компетенций, требуется поддержка не менее пяти типов заданий [22, 68]. Во-вторых, большие языковые модели не обеспечивают гарантированной фактической корректности, дидактической пригодности и соответствия заявленному уровню сложности каждого отдельного задания [19]. Существенным ограничением современных больших языковых моделей (LLM) является проблема галлюцинаций, заключающаяся в генерации правдоподобно сформулированных, но фактически некорректных утверждений [19, 83]. Дополнительно следует учитывать и другие типы дефектов, не связанные напрямую с фактической ложностью, включая недостаточную ясность формулировок, нарушения структуры вариантов ответа и несоответствие заданий требуемому уровню сложности.

Предлагаемое решение основано на разделении процессов генерации и контроля качества. Первый модуль алгоритма формирует тестовое задание на основе большой языковой модели, тогда как второй выполняет его многоаспектную верификацию. Выделение процедуры оценки в отдельный этап представляется целесообразным, поскольку совмещение генерации и оценки в рамках единого запроса приводит к конкуренции инструкций в модели и снижает надёжность выявления дефектов [83, 31]. В данной работе этот этап реализован в виде двухфазовой мультиагентной системы. На первой фазе специализированные ИИ-агенты независимо проверяют различные аспекты качества задания, на второй фазе при необходимости уточняют свои решения через регламентированный обмен

сообщениями по заранее заданному графу коммуникации. Такая архитектура позволяет повысить корректность итогового массива заданий при сохранении генеративной вариативности базовой модели.

3.1. Формальная постановка задачи генерации многоформатных тестовых заданий

Задача генерации многоформатных тестовых заданий формализуется как отображение тройственного входного представления (C, τ, σ) в пространство структурированных заданий $(q, V(q))$.

Определение 3.1 (Входной контент). Входным контентом C называется текстовое представление образовательного материала, служащее информационной базой для генерации задания. В контексте разрабатываемого конвейера $C \equiv S^*$, где S^* – оптимальная аннотация, полученная по алгоритму главы 2. Допускается также прямое использование фрагментов исходного текста при обходе этапа аннотирования.

Определение 3.2 (Формат задания). Формат задания τ принадлежит конечному множеству:

$$\tau \in \mathcal{T} = \{\text{single}, \text{multiple}, \text{open}, \text{match}, \text{sequence}\} \qquad (3.1)$$

где «single» – задание с выбором одного верного ответа; «multiple» – с выбором нескольких верных ответов; «open» – открытая форма с развёрнутым ответом; «match» – установление соответствия; «sequence» – установление правильной последовательности.

Определение 3.3 (Уровень сложности). Уровень сложности σ задаётся дискретным параметром:

$$\sigma \in \{1, 2, 3\} \qquad (3.2)$$

где $\sigma = 1$ – базовый уровень (уровни таксономии Блума L_1: запоминание и L_2: понимание); $\sigma = 2$ – средний уровень (L_3: применение и L_4: анализ); $\sigma = 3$– продвинутый уровень (L_5: оценивание и L_6: синтез/создание). Обозначим через $\mathcal{L} = \{L_1, L_2, L_3, L_4, L_5, L_6\}$ полное множество уровней таксономии.

Определение 3.4 (Тестовое задание). Тестовым заданием формата τ называется пара $(q, V(q))$, где q – текст формулировки задания, $V(q)$– структура ответа, зависящая от формата:

$$V(q) = \begin{cases} (A^*, D), & \tau \in \{\text{single, multiple}\}, \\ (a_{\text{ref}}, \mathbf{K}), & \tau = \text{open}, \\ \{(l_i, r_i)\}_{i=1}^{p}, & \tau = \text{match}, \\ (e_1, \dots, e_s), & \tau = \text{sequence}, \end{cases} \qquad (3.3)$$

где A^* – множество правильных ответов, $D = \{d_1, \dots, d_m\}$ – множество дистракторов, при этом $A^* \cap D = \emptyset$; для формата «single» выполняется $| A^* | = 1$, для формата «multiple» – $| A^* | \geq 2$; a_{ref} – эталонный ответ, $\mathbf{K} = (k_1, \dots, k_n)$ – вектор критериев оценивания (l_i, r_i) – i-я пара соответствия; $(e_1, \dots, e_s)$ – элементы правильной последовательности.

Определение 3.5 (Задача генерации). Задача генерации многоформатных тестовых заданий формулируется как вычисление:

$$(q, V(q)) = G(C, \tau, \sigma) \qquad (3.4)$$

где G– генератор, реализованный на базе языковой модели $\mathcal{M}$.

В таблице 3.1 систематизированы характеристики поддерживаемых форматов.

Таблица 3.1 – Характеристика поддерживаемых форматов тестовых заданий

Формат τ	**Структура $V(q)$**	**Число вариантов**	**Пример применения**
single	1 верный + 3–4 дистрактора	4–5	Выбор корректного определения термина
multiple	2–4 верных + 2–3 неверных	4–7	Выбор характеристик явления
open	Эталонный ответ и критерии оценивания	–	Объяснение явления или процесса
match	Два столбца для сопоставления	4–6 пар	Соотнесение понятий и определений
sequence	Неупорядоченные элементы	4–6 шагов	Установление этапов процесса

Многообразие форматов обеспечивает покрытие всех уровней таксономии Блума [22]. Данное соответствие имеет важное значение для составления

контрольно-измерительных материалов, обеспечивающих диагностику сформированности компетенций на различных когнитивных уровнях.

На основе анализа нормативных документов в области педагогических измерений и психометрической литературы сформулированы требования к корректности тестового задания. Совокупность требований определяет систему критериев, каждый из которых реализуется через соответствующего верификационного агента.

Требование 1 (Фактическая корректность). Все содержательные утверждения формулировки задания q и верного ответа A^* (либо a_{ref}) должны соответствовать фактическому содержанию контента C. Как показано в работе Min et al. [83], длинные ответы языковых моделей могут содержать совокупность как подтверждаемых, так и неподтверждаемых утверждений, что предполагает необходимость их декомпозиции и последующей верификации относительно исходного источника.

Требование 2 (Тематическая релевантность). Задание должно проверять знание содержания, непосредственно связанного с C; использование знаний, не представленных в C, приводит к снижению корректности оценки результатов.

Требование 3 (Валидность структуры ответа). Структура $V(q)$ должна задавать корректное решение, однозначно определяемое для соответствующего формата задания. Один правильный ответ для «single», множество не менее двух правильных ответов для «multiple», эталонный ответ и критерии оценивания для «open», единственное корректное соответствие или порядок для «match» и «sequence». Отсутствие однозначно верного ответа делает задание непригодным для диагностики вне зависимости от качества формулировки самого задания.

Требование 4 (Ясность и однозначность формулировки). Формулировка задания q должна допускать единственную интерпретацию; использование двусмысленных конструкций, неопределённых указательных выражений и недостаточно чётко заданных ограничений недопустимо.

Требование 5 (Соответствие уровня сложности). Когнитивные операции, необходимые для ответа, должны соответствовать параметру σ по таксономии [22].

Нарушение характерно при $\sigma = 3$, когда модель формирует вопросы на прямое воспроизведение факта, маскируя их под аналитические.

Требование 6 (Отсутствие предвзятости). Задание не должно содержать формулировок, создающих преимущества или ограничения для отдельных групп обучающихся по признакам, не связанным с проверяемой компетенцией.

Требование 7 (Самостоятельность содержания). Формулировка q не должна представлять собой прямое воспроизведение фрагмента C. Задание, допускающее получение ответа посредством распознавания цитируемого фрагмента, не обеспечивает диагностику компетенций и противоречит принципу валидности теста [68].

Требования 1–7 отражают различные аспекты качества тестового задания и поэтому проверяются раздельно специализированными агентами. При этом отдельные аспекты качества содержательно связаны между собой, что учитывается на этапе избирательной межагентной коммуникации.

3.2. Описание алгоритма

Алгоритм объединяет два функциональных модуля: генерацию задания языковой моделью и мультиагентную верификацию с двухфазовым протоколом взаимодействия. Если задание не проходит верификацию, инициируется повторная генерация с учётом выявленных дефектов.

На первом этапе языковая модель $\mathcal{M}$ принимает на вход набор параметров (C, τ, σ) и формирует задание $(q, V(q))$ в соответствии со спецификацией форматов. В качестве $\mathcal{M}$ используется Qwen3 235B A22B [129] – модель MoE-архитектуры с 235 млрд. суммарных и 22 млрд. активных параметров, представленная в конфигурации, ориентированной на задачи многошагового рассуждения. Режим Thinking в данном случае критически важен для генерации заданий уровня $\sigma = 3$, а также форматов «match» и «sequence», требующих структурированного вывода.

Запрос к модели формируется в виде последовательного объединения четырёх блоков:

$$p = [\mathrm{SYS}(\tau, \sigma) \parallel \mathrm{CONTENT}(C) \parallel \mathrm{FORMAT\_SPEC}(\tau) \parallel \mathrm{FEEDBACK}(\phi)] \quad (3.5)$$

где $\mathrm{SYS}(\tau, \sigma)$ – системная инструкция с указанием формата, уровня сложности и роли генератора; $\mathrm{CONTENT}(C)$ – блок входного контента; $\mathrm{FORMAT\_SPEC}(\tau)$ – схема ожидаемой. структуры вывода; $\mathrm{FEEDBACK}(\phi)$ – текстовое описание выявленных несоответствий на предыдущей итерации. При первичной генерации $\phi = \emptyset$, в связи с чем соответствующий блок в запрос не включается.

Схема вывода применяется совместно с механизмом ограниченного декодирования [124], который сужает пространство допустимых токенов на каждом шаге генерации путём маскирования некорректных токенов с последующей нормализацией распределения вероятностей. Применение данного механизма обеспечивает структурную корректность вывода на уровне механизма декодирования.

Схемы структуры выходных данных для всех форматов тестовых заданий задаются в виде формализованных описаний, определяющих состав полей и их семантическую интерпретацию. Такая формализация обеспечивает единообразие представления результатов генерации и возможность их последующей автоматизированной обработки.

Для формата «single» структура задания включает текст формулировки задания, корректный ответ, набор дистракторов, уровень сложности и тематическую принадлежность. При этом дистракторы представляют собой множество неверных, но содержательно релевантных вариантов ответа, предназначенных для проверки понимания материала.

Для формата «match» структура задания содержит текст инструкции, два множества элементов для установления соответствия, а также множество корректных пар. Соответствия задаются посредством индексов элементов, что обеспечивает однозначность представления решения. Дополнительно указывается уровень сложности задания.

Для формата «open» структура задания включает текст формулировки, эталонный ответ и набор критериев оценивания, определяющих требования к развернутому ответу. Критерии формализуют процедуру оценки и обеспечивают

воспроизводимость результатов проверки. Также задаётся уровень сложности задания.

Верификационный модуль разрабатываемого алгоритма реализуется в виде координируемой мультиагентной системы, в которой специализированные ИИ-агенты выполняют локальную экспертную проверку различных аспектов качества тестового задания, а детерминированный координатор управляет протоколом их взаимодействия. Данная архитектура принципиально отличается как от одноагентной схемы, где все критерии качества объединяются в одном запросе, так и от ансамбля независимых агентов без межагентной коммуникации. В предлагаемой системе агенты обладают различной предметной специализацией, используют разные подмножества входной информации и разные инструментальные контуры; взаимодействие между ними допускается только в тех случаях, когда между проверяемыми критериями существует содержательно обоснованная зависимость.

Необходимость применения мультиагентной организации обусловлена тем, что требования к качеству тестового задания частично независимы, но не полностью изолированы друг от друга. Фактическая ошибка в помеченном правильном ответе закономерно влияет на валидность общей структуры решения; неоднозначная формулировка способна изменить вывод о существовании единственного правильного ответа; почти дословное копирование исходного текста способно искусственно снижать фактическую когнитивную сложность задания. Следовательно, эффективная система контроля должна сочетать два свойства: независимость первичных суждений по каждому критерию и регламентированное взаимодействие в тех случаях, когда между критериями возникает содержательная зависимость.

Функциональное разделение между агентами и координатором имеет принципиальное значение. Агенты осуществляют содержательную экспертную оценку задания; координатор не принимает собственных предметных решений и не подменяет собой ни один из агентов. Его функции ограничены запуском фаз взаимодействия, фиксацией результатов, маршрутизацией по графу

коммуникации, однократным запуском процедуры пересмотра решений на второй фазе и применением детерминированного правила допуска.

Определение 3.6 (ИИ-агент верификации). ИИ-агентом верификации A_j называется кортеж $\langle \Psi_j, I_j, \mathcal{T}_j, \mathcal{M}, O_j \rangle$, где:

- Ψ_j – системный промпт, определяющий роль, предметную экспертизу и критерий проверки агента;
- $I_j \subseteq \{q, V(q), C, \sigma\}$ – подмножество входных данных из разделяемой среды, доступных агенту;
- $\mathcal{T}_j$ – множество инструментов (вычислительных функций), которые агент вызывает в процессе рассуждения;
- $\mathcal{M}$ – языковая модель, реализующая способность агента к рассуждению;
- O_j – функция вывода, формирующая результат вида $O_j(I_j) = \langle s_j, r_j, F_j \rangle$, где $s_j \in \{0,1\}$ – бинарное решение, $r_j \in \mathcal{R}$ – языковое обоснование, $F_j \subseteq \mathcal{F}$ – множество выявленных проблем.

Принципиальное отличие агента от детерминированной процедуры заключается в том, что решение s_j формируется языковой моделью $\mathcal{M}$ на основе рассуждения, в рамках которого результаты инструментов $\mathcal{T}_j$ выступают одним из источников информации, но не являются единственным основанием для принятия решения. В отличие от процедур, основанных на жёстко заданных правилах, агент допускает отклонение от числовых сигналов метрик в тех случаях, когда содержательный анализ задачи требует иной интерпретации полученных значений.

Определение 3.7 (Мультиагентная система верификации). Мультиагентной системой верификации называется кортеж $\mathcal{V} = \langle \mathcal{A}, \mathcal{E}, \mathcal{B}, \mathcal{G}_{\text{com}}, \mathcal{R}, Coord \rangle$, где:

- $\mathcal{A} = \{A_1, A_2, \dots, A_7\}$ – конечное множество специализированных ИИ-агентов;
- $\mathcal{E} = \{(q, V(q)), C, \sigma\}$ – разделяемая среда (верифицируемое задание и его контекст);
- $\mathcal{B}$ – доска объявлений (blackboard), в которую агенты публикуют результаты первой фазы и из которой читают результаты на второй фазе;

- $\mathcal{G}_{\text{com}} = (\mathcal{A}, \mathcal{E}_{\text{com}})$ – граф коммуникации, в котором ориентированные рёбра задают разрешённые каналы межагентного взаимодействия на Фазе 2;
- $\mathcal{R}: \{0,1\}^7 \to \{0,1\}$ – детерминированная функция агрегации окончательных решений агентов.
- *Coord* – управляющий координатор.

Система верификации включает $N_A = 7$ ИИ-агентов, каждый из которых проверяет одно из требований раздела 3.1. Выбор специализированной архитектуры обусловлен тем, что объединение нескольких критериев в одном промпте приводит к конкуренции критериев за ресурсы внимания модели и снижает выявляемость дефектов [31]. Полная характеристика агентов приведена в таблице 3.2.

Таблица 3.2 – Состав и характеристика агентов системы верификации

№	Агент	Требование	Входные данные I_j	Инструменты $\mathcal{T}_j$
1	FactChecker	Фактическая корректность	$q, V(q), C$	LLM-декомпозиция, поиск опорных фрагментов, NLI-entailment
2	ContextGrounding	Тематическая релевантность	q, C	BERTScore, TF-IDF cosine
3	AnswerValidity	Валидность структуры ответа	$q, V(q)$	LLM-рассуждение, независимое решение задания
4	Clarity	Ясность формулировки	q	LLM-лингвистический анализ
5	DifficultyMatch	Соответствие уровня сложности σ	$q, V(q), \sigma$	Классификатор таксономии, вспомогательный набор поиска
6	BiasDetection	Отсутствие предвзятости	$q, V(q)$	LLM-аудит, набор потенциальных выражений
7	Originality	Самостоятельность содержания	q, C	ROUGE-L, n-gram overlap

Агент A_1 – «FactChecker».

Роль и цель. Агент «FactChecker» является экспертом по верификации фактических утверждений. Агент проверяет, все ли содержательные утверждения задания поддерживаются исходным контентом C, а также выявляет «галлюцинации» модели.

Инструменты.

– LLM-декомпозиция – декомпозиция на отдельные утверждения из задания;

– NLI-классификатор $\mathcal{M}_{\text{NLI}}$ – специализированная модель, обученная на задачах Natural Language Inference, отдельная от генерирующей модели $\mathcal{M}$.

Процесс рассуждения:

На первом шаге агент применяет LLM-декомпозицию для выделения множества атомарных фактических утверждений из формулировки задания q и верного ответа A^*:

$$\mathcal{U}(q,V(q)) = \{u_1, u_2, \dots, u_k\} \qquad (3.6)$$

Атомарным считается такое утверждение, которое допускает самостоятельную проверку без обращения к соседним предложениям.

На втором шаге для каждого $u_i \in \mathcal{U}$ агент запрашивает инструмент NLI, вычисляющий вероятность логического следствия:

$$\text{ent}(u_i, C) = \mathcal{M}_{\text{NLI}}(C, u_i) \in [0,1] \qquad (3.7)$$

где $\text{ent}(u_i, C) \approx 1$ означает, что контент C логически влечёт утверждение u_i (то есть $C \vDash u_i$), а $\text{ent}(u_i, C) \approx 0$ – противоречие.

На третьем шаге агент проводит рассуждение о каждом утверждении, анализируя его значение $\text{ent}(u_i, C)$; конкретный фрагмент C, в котором утверждение может быть найдено или опровергнуто; возможные перефразирования, логические следования, допустимые обобщения. Агент не сводит решение к сравнению с фиксированным порогом, а интерпретирует числовой сигнал в контексте конкретного утверждения и содержательного представления.

Формирование решения. Агент принимает положительное решение, если все атомарные утверждения поддерживаются контентом либо являются допустимыми смысловыми обобщениями. Если хотя бы одно утверждение признано

неподдерживаемым или противоречащим источнику, агент формирует отрицательное решение и фиксирует его в $\mathcal{F}_1$:

$$s_1 = \begin{cases} 1, & \text{если все } u_i \text{ поддерживаются } C, \\ 0, & \text{если существует противоречащее } u_i \end{cases} \quad (3.8)$$

Агент A_2 – «ContextGrounding».

Роль и цель. Агент «ContextGrounding» проверяет, обеспечивается ли решение задания содержанием контента C, и определяет, достаточно ли информации, представленной в источнике, для ответа без привлечения внешнего знания. В отличие от проверки фактической корректности, выполняемой агентом «FactChecker», данный агент оценивает не истинность отдельных утверждений, а степень содержательной опоры задания на исходный контент и тематическую релевантность проверяемого материала.

Инструменты.

– BERTScore – семантическое сходство между q и C, вычисляемое на уровне векторных представлений токенов;

– TF-IDF cosine – лексическое сходство на основе взвешенного частотного представления.

Процесс рассуждения:

На первом шаге агент запрашивает оба инструмента:

$$b = \text{BERTScore}(q, C), c = \text{cosine\_tfidf}(q, C) \quad (3.9)$$

где $b, c \in [0,1]$;

На втором шаге агент использует полученные значения как ориентиры. Низкие значения обоих инструментов указывают на вероятную недостаточность контекстной опоры или тематическую нерелевантность, однако сами по себе не являются достаточным основанием для отрицательного решения.

На третьем шаге агент содержательно отвечает на вопрос: может ли обучающийся корректно решить задание, опираясь только на сведения, представленные в C, или для ответа требуется внешнее знание. На второй фазе агент учитывает сообщения от «FactChecker». Если в задании выявлены

неподтверждаемые утверждения, агент переоценивает, не выходит ли задание за пределы содержания исходного контента.

Формирование решения. Принимается $s_2 = 1$, если агент устанавливает, что для ответа достаточно информации из C и проверяемое содержание тематически соотнесено с источником; $s_2 = 0$ – если задание требует знаний, отсутствующих в C, либо выходит за пределы его содержательной области. В множество F_2 агент записывает конкретные понятия, утверждения или фрагменты решения, для которых не обнаружена достаточная опора в источнике.

Агент A_3 – «AnswerValidity».

Роль и цель. Агент «AnswerValidity» отвечает за валидность структуры ответа. Его задача заключается в установлении, задаёт ли $V(q)$ однозначно корректный ответ в соответствии с выбранным форматом задания.

Инструменты:

– Независимое решение задания – агент самостоятельно формирует ответ, не ориентируясь на заранее размеченный правильный вариант.

– LLM-рассуждение – интерпретация результатов независимого решения и анализ допустимых вариантов ответа.

Процесс рассуждения:

Для форматов «single» и «multiple» агент самостоятельно отвечает на вопрос q, после чего сопоставляет собственный вывод с A^*. Далее он анализирует дистракторы, проверяя, не существует ли среди них вариантов, которые при допустимой интерпретации вопроса также могут быть признаны верными.

Для формата «open» агент проверяет, охватывает ли a_{ref} смысл вопроса и согласованы ли критерии K.

Для форматов «match» и «sequence» агент оценивает однозначность каждой пары соответствия или каждой позиции последовательности. При наличии альтернативных и содержательно допустимых решений фиксируется дефект структуры ответа.

На второй фазе агент учитывает сообщения от FactChecker и Clarity. Если FactChecker указывает на фактическую ошибку в помеченном правильном ответе,

агент пересматривает сам факт существования корректного решения в $V(q)$. Если Clarity выявляет неоднозначность формулировки, агент переоценивает, остаётся ли ответ однозначным при альтернативных интерпретациях вопроса.

Формирование решения. $s_3 = 1$, если однозначно корректный ответ установлен; $s_3 = 0$ – при выявлении неоднозначности или структурного несоответствия.

Агент A_4 – «Clarity».

Роль и цель. Агент «Clarity» выполняет лингвистическую экспертизу задания и проверяет, допускает ли формулировка задания единственную интерпретацию.

Инструменты: LLM-лингвистический анализ – синтаксическая, семантическая и прагматическая интерпретация формулировки.

Процесс рассуждения: Агент анализирует задание q по следующим аспектам:

(а) Неоднозначные референции. Местоимения или указательные слова без явного антецедента («оно», «это» без предшествующего существительного).

(б) Двойные отрицания. Конструкции «не…не…», «кроме исключений», усложняющие понимание.

(в) Скрытая конъюнкция требований. Вопросы вида «Какое из утверждений верно И применимо в данном контексте?», предполагающие одновременное выполнение двух условий.

(г) Культурные и языковые допущения. Термины, понятные лишь части аудитории вследствие культурной или языковой специфики, не связанной с проверяемой компетенцией.

(д) Неопределённые кванторы: «как правило», «обычно», «часто» без критерия размывают границу правильного ответа.

Выявленные фрагменты не интерпретируются как механические маркеры брака. Агент оценивает, действительно ли они создают неоднозначность, препятствующую диагностической интерпретации задания.

Формирование решения. $s_4 = 1$, если не обнаружено ни одного значимого источника неоднозначности; $s_4 = 0$ – при наличии хотя бы одной неясности,

способной повлиять на интерпретацию ответа. В $\mathcal{F}_4$ агент цитирует проблемный фрагмент и поясняет тип неоднозначности.

Агент A_5 – «DifficultyMatch».

Роль и цель. Агент «DifficultyMatch» оценивает соответствие задания заявленному уровню сложности σ и определяет, какую когнитивную операцию фактически должен выполнить обучающийся при ответе.

Инструменты:

– Классификатор, необходимый для отнесения задания к одному из уровней таксономии.

– Набор когнитивных глаголов, представляющий собой детерминированный словарь маркеров уровней таксономии:

– Определение функции $\mathrm{map}(\sigma)$**.** Отображение уровней таксономии на параметр сложности задаётся явно:

$$\mathrm{map}(\sigma) = \begin{cases} \{L_1, L_2\}, & \sigma = 1, \\ \{L_3, L_4\}, & \sigma = 2, \\ \{L_5, L_6\}, & \sigma = 3, \end{cases} \qquad (3.10)$$

где L_1: запоминание, L_2: понимание, L_3: применение, L_4: анализ, L_5: оценивание, L_6: синтез/создание [22].

Процесс рассуждения:

На первом шаге агент использует набор когнитивных глаголов и формирует предварительную гипотезу о требуемом уровне сложности. На втором шаге агент запрашивает классификатор:

$$\hat{L} = \Phi_{\mathrm{Bloom}}(q, V(q)) \in \mathcal{L} \qquad (3.11)$$

На третьем шаге агент сопоставляет сигналы, отвечая на вопрос, какая когнитивная операция требуется для решения задания: воспроизведение, применение, анализ, оценивание или создание. При расхождении лексического сигнала и классификатора приоритет отдаётся содержательному анализу задачи.

На второй фазе агент учитывает сообщения от FactChecker, AnswerValidity и Originality. Фактическая ошибка в правильном ответе, структурная неоднозначность решения и почти дословное воспроизведение контента делают

оценку реального когнитивного уровня менее надёжной и требуют пересмотра вывода. В частности, при обнаружении почти буквального копирования агент переоценивает, не снижается ли фактическая сложность задания до уровня механического распознавания фрагмента исходного текста.

Формирование решения. Агент принимает

$$s_5 = 1 \Leftrightarrow \hat{L} \in \mathrm{map}(\sigma). \qquad (3.12)$$

то есть $s_5 = 1$, если предсказанный когнитивный уровень $\hat{L}$ входит в множество уровней, соответствующих параметру σ; иначе $s_5 = 0$. В обоснование r_5 агент указывает, какая когнитивная операция требуется и почему она соответствует (или не соответствует) уровню σ.

Агент A_6 – «BiasDetection».

Роль и цель. Агент проверяет, не содержит ли задание формулировок, ставящих в невыгодное положение группы обучающихся по признакам, не связанным с проверяемой компетенцией.

Инструменты:

– LLM-аудит формулировок (анализ потенциально дискриминирующих эффектов);

– Сканирование по набору потенциально релевантных выражений.

Процесс рассуждения: Агент анализирует задание по следующим измерениям:

(а) Гендерный и культурный нейтралитет. Использование нейтральных примеров; отсутствие неявных предположений о поле, этнической принадлежности, культурном контексте.

(б) Социально-экономические допущения. Предполагает ли задание наличие знаний, доступных только части обучающихся по имущественному признаку.

(в) Языковая сложность. Не превышает ли сложность формулировки проверяемый предметный уровень (то есть задание оценивает владение языком, а не предметом).

(г) Предвзятость в дистракторах.

Формирование решения. Агент принимает $s_6 = 1$, если ни один из перечисленных дефектов не выявлен, и $s_6 = 0$ – при наличии хотя бы одного признака предвзятости или нерелевантной языковой недоступности. В $\mathcal{F}_6$ агент описывает конкретный предвзятый элемент и его потенциальный дискриминирующий эффект.

Агент A_7 – «Originality».

Роль и цель. Агент проверяет, переработано ли содержание задания контента в новую форму, а не воспроизводит его дословно.

Инструменты:

– ROUGE-L – метрика наибольшей общей подпоследовательности между q и C;

– N-gram overlap – доля совпадающих биграмм между q и C.

Процесс рассуждения:

На первом шаге агент вычисляет обе метрики:

$$\rho = \text{ROUGE-L}(q, C),\ \ \omega_2 = \text{ngram\_overlap}(q, C, n = 2) \qquad (3.13)$$

где $\rho, \omega_2 \in [0,1]$; значения, близкие к 1, указывают на высокую лексическую близость.

На втором шаге агент анализирует совпадающие последовательности, различая неизбежные совпадения терминологии и случаи фактического копирования исходного фрагмента.

На третьем шаге агент отвечает на вопрос, заключающийся в том, допускает ли задание ответ без понимания содержания, путём механического поиска соответствующего места в исходном тексте. Если да, то задание нарушает требование самостоятельности.

Формирование решения. Агент принимает $s_7 = 1$, если задание формулируется через переработку внутреннего содержания и подразумевает понимание; $s_7 = 0$ – при обнаружении копирования. В $\mathcal{F}_7$ агент указывает совпадающие фрагменты.

Взаимодействие агентов организовано по схеме доски объявлений, дополненной явным графом содержательных зависимостей между критериями

качества. Такая организация позволяет сочетать независимость первичной оценки с ограниченным взаимодействием по содержательно обоснованным каналам.

Требование независимости. До завершения первой фазы агенты не имеют доступа к решениям друг друга: $\forall j \neq j' : A_j \perp A_{j'}$ в фазе 1. Указанное требование исключает распространение ошибок.

Фаза 1. Независимая параллельная оценка.

Все агенты запускаются одновременно. Каждый агент A_j принимает локальные входные данные I_j, вызывает необходимые инструменты T_j, выполняет рассуждение и формирует выход первой фазы:

$$O_j^{(1)} = (s_j^{(1)}, r_j^{(1)}, \mathcal{F}_j^{(1)}), \qquad j = 1, \dots, 7 \qquad (3.14)$$

где $s_j^{(1)} \in \{0,1\}$ – предварительное бинарное решение, $r_j^{(1)}$ – обоснование, $\mathcal{F}_j^{(1)}$ – множество выявленных проблем. По завершении фазы 1 все агенты публикуют результаты на доску объявлений:

$$\mathcal{B} \leftarrow \{j, s_j^{(1)}, r_j^{(1)}, \mathcal{F}_j^{(1)}\}_{j=1}^{7} \qquad (3.15)$$

Фаза 2. Избирательная коммуникация и пересмотр решений.

После публикации результатов на $\mathcal{B}$ активируется граф коммуникации $\mathcal{G}_{\text{com}} = (\mathcal{A}, \mathcal{E}_{\text{com}})$. Содержательное обоснование каждого ребра приведено в таблице 3.3.

Для предотвращения избыточных вычислений пересмотр выполняется не по отдельным сообщениям, а по совокупности всех сообщений, адресованных одному агенту. После завершения Фазы 1 для каждого агента инициализируется промежуточный результат:

$$O_j^{(2)} \leftarrow O_j^{(1)}, \qquad j = 1, \dots, 7 \qquad (3.16)$$

Далее для каждого агента-получателя $A_{j'}$ формируется множество входящих сообщений:

$$M_{j'} = \left\{ (s_j^{(1)}, r_j^{(1)}, F_j^{(1)}) \mid (A_j \to A_{j'}) \in E_{\text{com}} \wedge \left(s_j^{(1)} = 0 \vee F_j^{(1)} \neq \emptyset \right) \right\} \qquad (3.17)$$

Условие активации означает, что сообщение передаётся только тогда, когда агент-источник выявил дефект. Либо сформировано отрицательное решение, либо

зафиксировано непустое множество замечаний. При отсутствии дефектов соответствующий канал не активируется, что снижает вычислительную нагрузку второй фазы.

Если $M_{j'} \neq \emptyset$, агент-получатель выполняет однократный пересмотр своего решения:

$$O_{j'}^{(2)} = A_{j'}(I_{j'}, M_{j'}) = \left(s_{j'}^{(2)}, r_{j'}^{(2)}, F_{j'}^{(2)}\right) \qquad (3.18)$$

Если $M_{j'} = \emptyset$, сохраняется результат первой фазы:

$$O_{j'}^{(2)} = O_{j'}^{(1)}. \qquad (3.19)$$

Таблица 3.3 – Содержание графа коммуникации

Ребро	**Агент-источник**	**Агент-получатель**	**Сжатая форма сообщения**
$A_1 \rightarrow A_2$	FactChecker	ContextGrounding	Выявлено утверждение, не подтверждаемое контентом; необходимо переоценить, решается ли задание исключительно на основе C
$A_1 \rightarrow A_3$	FactChecker	AnswerValidity	Помеченный правильный ответ содержит фактическую ошибку или неподтверждаемое утверждение; требуется переоценить валидность структуры ответа
$A_4 \rightarrow A_3$	Clarity	AnswerValidity	Формулировка допускает неоднозначную интерпретацию; требуется проверить, не возникает ли несколько допустимых ответов
$A_1 \rightarrow A_5$	FactChecker	DifficultyMatch	Фактическая ошибка искажает содержательную основу задания; вывод о соответствии уровню сложности требует пересмотра
$A_3 \rightarrow A_5$	AnswerValidity	DifficultyMatch	Структурная неоднозначность ответа делает неочевидной реальную когнитивную операцию, необходимую для решения
$A_7 \rightarrow A_5$	Originality	DifficultyMatch	Почти дословное копирование контента может искусственно снижать фактическую когнитивную сложность задания

Однократность пересмотра является важным свойством протокола. Оно исключает конфликт множественных переопределений, предотвращает зацикленные пересмотры решений и обеспечивает предсказуемость вычислительной нагрузки. Каждый агент может быть вызван на второй фазе не более одного раза.

Двухфазовый протокол сочетает независимость первичных выводов и контролируемое межагентное взаимодействие. Первая фаза обеспечивает автономность локальных решений, вторая позволяет учесть имеющиеся зависимости только в тех случаях, когда они действительно существенны для качества задания.

После завершения второй фазы формируется вектор окончательных бинарных решений:

$$s(q) = \left(s_1^{(2)}, s_2^{(2)}, s_3^{(2)}, s_4^{(2)}, s_5^{(2)}, s_6^{(2)}, s_7^{(2)}\right) \in \{0,1\}^7 \qquad (3.20)$$

Маска дефектных аспектов задания определяется как

$$f(q) = \mathbf{1}_7 - s(q), \qquad f_j = 1 - s_j^{(2)}, \qquad j = 1, \dots, 7, \qquad (3.21)$$

где $\mathbf{1}_7 = (1,1,1,1,1,1,1)^T$ – единичный вектор размерности 7. Компонента $f_j = 1$ означает нарушение j-го критерия, а $f_j = 0$ – его выполнение. Носитель $\mathrm{supp}(f(q)) = \{j: f_j = 1\}$ задаёт перечень нарушенных критериев и является основой для формирования обратной связи ϕ при повторной генерации.

Степень дефектности задания определяется как

$$\delta(q) = \sum_{j=1}^{7} f_j = 7 - \sum_{j=1}^{7} s_j^{(2)} \qquad (3.22)$$

Задание является бездефектным при $\delta(q) = 0$ и дефектным по всем критериям при $\delta(q) = 7$.

На основе вектора решений применяется строгое конъюнктивное правило допуска:

$$\mathrm{Accept}(q) = \prod_{j=1}^{7} s_j^{(2)} \qquad (3.23)$$

Поскольку $s_j^{(2)} \in \{0,1\}$, произведение равно 1 тогда и только тогда, когда все семь агентов приняли положительное решение. Следовательно, задание допускается в итоговый набор банка заданий только при выполнении всех требований одновременно; наличие хотя бы одного дефекта влечёт отклонение задания и запуск повторной генерации.

Использование именно конъюнктивного правила обусловлено спецификой образовательных тестовых материалов. В отличие от задач, где допустим компромисс между отдельными критериями, в контексте контрольно-измерительных материалов, единичный дефект способен сделать задание непригодным для диагностики независимо от того, насколько хорошо выполнены остальные требования. Так, фактическая ошибка в правильном ответе не компенсируется ясной формулировкой, а структурная неоднозначность не устраняется тематической релевантностью.

Для обоснования выбора правила принятия задания в таблице 3.4 приведено сравнение альтернативных стратегий агрегации решений.

Таблица 3.4 – Сравнение стратегий агрегации решений агентов

Стратегия	**Формула допуска**	**Точность принятых заданий**	**Полнота выявления дефектных заданий**	**Риск пропуска критического дефекта**
Строгое конъюнктивное	$\prod s_j^{(2)} = 1$	Высокая	Высокая	Минимальный
Большинство голосов	$\sum s_j^{(2)} \geq 4$	Умеренная	Умеренная	Высокий
Взвешенное голосование	$\sum w_j s_j^{(2)} \geq \theta$	Настраиваемая	Настраиваемая	Зависит от w_j
Иерархическое	$s_1^{(2)} \cdot s_3^{(2)} = 1$ и $\sum_{j \notin \{1,3\}} s_j^{(2)} \geq 4$	Высокая	Умеренная	Низкий

В рамках таблицы 3.4 под точностью понимается доля заданий, принятых системой и действительно свободных от критических дефектов, а под полнотой – доля дефектных заданий, корректно отклонённых системой. Строгое конъюнктивное правило обеспечивает максимальную чувствительность к любым дефектам. Одиночный дефект гарантированно приводит к отклонению задания. Возможное снижение доли автоматически принимаемых заданий в этом случае компенсируется механизмом повторной генерации с адресной обратной связью.

При отрицательном результате верификации текстовая обратная связь $\phi = \text{generate\_feedback}(\text{supp}(f(q)))$ включается в промпт следующей итерации через блок $\text{FEEDBACK}(\phi)$ в формуле (3.5), направляя генератор к устранению конкретных выявленных дефектов. Все семь агентов первой фазы работают параллельно, что снижает задержку верификации до времени одного LLM-запроса плюс задержка на запуск и агрегацию. Компоненты вычислительной схемы алгоритма представлены в таблице 3.5.

Таблица 3.5 – Компоненты вычислительной схемы алгоритма

Этап	**Компонент**	**Тип операции**	**Число LLM-запросов**	**Зависимости**
1	Конструирование промпта	Детерминированная	0	C, τ, σ, ϕ
2	Генерация задания	Стохастическая	1	Промпт p
3	Синтаксический разбор JSON	Детерминированная	0	Вывод M
4	Фаза 1: параллельная верификация	Стохастическая	7 (параллельно)	I_j для каждого A_j
5	Публикация на B	Детерминированная	0	$O_j^{(1)}$
6	Фаза 2: избирательная коммуникация	Стохастическая	0–3	B, G_{com}
7	Агрегация решений	Детерминированная	0	$s(q)$
8	Генерация обратной связи	Детерминированная	0	$\text{supp}(f(q))$

Примечание. На второй фазе запросы активируются только для агентов-получателей A_2, A_3, A_5; следовательно, максимально возможное число дополнительных LLM-запросов на второй фазе равно трём.

Оценка среднего числа LLM-запросов на одно принятое задание строится эмпирически на основе логов выполнения алгоритма. Базовое число запросов на одну итерацию равно $1 + 7 = 8$ (один запрос на генерацию и семь параллельных запросов агентов первой фазы). Среднее число дополнительных запросов второй фазы обозначается через n_{com}. По данным экспериментальных запусков оно оценивается как $n_{\text{com}} \approx 0{,}63$. Среднее ожидаемое число итераций обозначается через $E[\text{attempt}]$. По эмпирическим данным о прохождении верификации и повторной генерации оно составляет $E[\text{attempt}] \approx 1{,}42$. Тогда среднее число LLM-запросов на одно итоговое принятое задание оценивается как

$$E[N_{\text{LLM}}] = (8 + n_{\text{com}}) \cdot E[\text{attempt}] \approx (8 + 0{,}63) \cdot 1{,}42 \approx 12{,}3. \qquad (3.24)$$

Таким образом, предложенный алгоритм использует в среднем около 12,3 LLM-запросов на одно итоговое принятое задание. По сравнению с базовым алгоритмом генерации, использующим 8 LLM-запросов в однопроходной схеме без второй фазы и без адресной перегенерации, прирост вычислительных затрат составляет около 54%. Такой рост обусловлен двумя факторами, связанными с дополнительными запросами второй фазы и повторной генерацией с целевой обратной связью. Как будет показано в экспериментах, данный прирост вычислительных затрат компенсируется существенным повышением качества выходного массива заданий.

3.3. Экспериментальное исследование алгоритма

Экспериментальное исследование проводилось на корпусе из 1000 автоматически сгенерированных тестовых заданий, равномерно распределённых по пяти форматам – по 200 заданий каждого типа и трём уровням сложности. Такой объём выборки обеспечивает статистически устойчивую оценку показателей качества и сопоставим с объёмами выборок, используемых в смежных исследованиях [19, 26].

Задания генерировались на базе текстов, сформированных алгоритмом главы 2 с использованием языковой модели. Для получения репрезентативного распределения дефектов алгоритм запускался с обязательным сохранением всех промежуточных результатов до применения верификационного модуля, что позволило сформировать полный массив черновых вариантов заданий первой попытки для анализа.

Оценка качества проводилась в два этапа. Первый этап представлял собой автоматическую верификацию всего массива из 1000 заданий. Второй этап включал экспертную оценку всех заданий корпуса. Каждое задание оценивалось двумя предметными специалистами по пятибалльной шкале по четырём критериям: содержательная точность, педагогическая ценность, ясность формулировки и адекватность уровня сложности. Межэкспертная согласованность контролировалась по коэффициенту каппы Коэна κ [35]. Распределение заданий по форматам и уровням сложности представлено в таблице 3.6.

Таблица 3.6 – Распределение выборки по форматам и уровням сложности

Формат	$\sigma = 1$	$\sigma = 2$	$\sigma = 3$	**Итого**
single	80	80	40	200
multiple	60	80	60	200
open	50	80	70	200
match	70	80	50	200
sequence	70	80	50	200
Итого	**330**	**400**	**270**	**1000**

Распределение уровней сложности ($\sigma = 1$: 33%; $\sigma = 2$: 40%; $\sigma = 3$: 27%) отражает типичное соотношение в образовательных тестах, где преобладают задания среднего уровня, базовый уровень составляют около трети, а более сложные – около четверти общего числа.

До верификации из 1000 заданий дефектными являлись 419 (41,9%). После прохождения верификационного модуля число заданий с дефектами в принятом потоке составило 65 (6,5%). Снижение доли дефектных заданий составляет

$$\Delta_{\text{def}} = \frac{419 - 65}{419} \times 100\% \approx 84{,}5\%. \qquad (3.25)$$

В таблице 3.7 каждому дефектному заданию сопоставлен один доминирующий тип дефекта, определяемый по основному нарушенному критерию.

Таблица 3.7 – Распределение заданий по доминирующему типу дефекта

Тип дефекта	**Агент**	**До**	**После**	**Снижение, %**
Задания с дефектами (всего)	–	419 (41,9%)	65 (6,5%)	84,5
Фактические ошибки	FactChecker	106	7	93,4
Тематическая нерелевантность	ContextGrounding	76	17	77,6
Некорректная структура ответа	AnswerValidity	54	9	83,3
Неясные формулировки	Clarity	64	13	79,7
Несоответствие уровня сложности	DifficultyMatch	47	10	78,7
Наличие предвзятости	BiasDetection	41	5	87,8
Копирование исходного текста	Originality	31	4	87,1

Наибольший эффект фильтрации достигается для фактических ошибок (93,4%). Это объясняется тем, что задача обладает более высокой степенью формальной определённости, поскольку проверка отдельных утверждений по заданному исходному контенту допускает более строгую формализацию по сравнению с оценкой релевантности или сложности. Дополнительный вклад в снижение дефектности вносит ребро $A_1 \rightarrow A_3$. Часть заданий, которые на первой фазе выглядели как структурно корректные, после сигнала от «FactChecker» были переоценены агентом «AnswerValidity» как дефектные.

Высокие показатели снижения для предвзятости и копирования также связаны с относительной определённостью соответствующих критериев. Более низкие значения показателей тематической нерелевантности и несоответствия уровня сложности обусловлены тем, что границы между релевантным и нерелевантным содержанием, а также между соседними когнитивными уровнями,

не поддаются столь же строгой формализации, как задачи фактической проверки. Распределение степени дефектности заданий в исследуемом массиве представлено в таблице 3.8.

Таблица 3.8 – Распределение степени дефектности $\delta(q)$ в массиве заданий

$\delta(q)$	**Заданий**	**Доля от дефектных, %**	**Доля от всей выборки, %**
0 (бездефектные)	581	–	58,1
1	219	52,3	21,9
2	112	26,7	11,2
3	58	13,8	5,8
≥ 4	30	7,2	3,0
Итого дефектных	**419**	**100**	**41,9**

Большинство дефектных заданий характеризуются наличием одного (52,3%) или двух (26,7%) нарушений, что свидетельствует о достаточно высоком исходном качестве генерации, обеспечиваемом базовой языковой моделью. Задания, содержащие четыре и более дефектов, составляют 3% выборки и преимущественно обусловлены структурными ошибками при формировании заданий сложных форматов. Доля дефектных заданий по форматам до и после верификации приведена в таблице 3.9.

Таблица 3.9 – Доля дефектных заданий по форматам (%)

Формат	**До верификации**	**После верификации**	**Снижение, п.п.**
single	38,0	5,5	−32,5
multiple	47,0	8,0	−39,0
open	39,5	7,0	−32,5
match	44,0	6,0	−38,0
sequence	41,0	6,0	−35,0
Среднее	**41,9**	**6,5**	**−35,4**

Наибольшая исходная дефектность характерна для формата «multiple» (47%), что обусловлено сложностью одновременной генерации нескольких правильных и нескольких неправильных вариантов с чётким разграничением их статуса. Для

форматов «match» и «sequence» исходная доля дефектных заданий также остаётся повышенной и составляет 44% и 41% соответственно, что связано с необходимостью соблюдения структурной однозначности соответствий и порядка элементов. После верификации разброс между форматами существенно сокращается и составляет 5,5–8,0%, а средняя доля дефектных заданий уменьшается с 41,9% до 6,5%, то есть на 35,4 п.п. Это свидетельствует о применимости предложенной схемы контроля качества ко всем использованным форматам заданий.

Экспертная оценка верифицированных заданий по пятибалльной шкале представлена в таблице 3.10.

Таблица 3.10 – Экспертная оценка качества верифицированных заданий

Формат	**Содержательная точность**	**Педагогическая ценность**	**Ясность**	**Адекватность сложности**	**Среднее**
single	4,3	4,1	4,2	4,0	4,15
multiple	4,1	4,2	4,0	3,9	4,05
open	4,4	4,3	4,3	4,1	4,28
match	4,2	4,0	4,1	4,0	4,08
sequence	4,3	4,1	4,2	4,1	4,18
Среднее	**4,26**	**4,14**	**4,16**	**4,02**	**4,15**

Средняя экспертная оценка верифицированных заданий составила 4,15 из 5, что соответствует уровню «хорошо». Наивысшие значения зафиксированы для формата «open», при этом языковая модель демонстрирует высокое качество генерации развернутых эталонных ответов и критериев оценивания. Наименьшие оценки характерны для формата «multiple», преимущественно по критерию адекватности уровня сложности, что указывает на сохраняющуюся трудность корректного балансирования числа верных и неверных вариантов ответа.

Для сопоставления была проведена аналогичная экспертная оценка отвергнутых заданий, среднее значение которой составило 3,12 из 5. Разница в 1,03

балла свидетельствует о том, что верификационный модуль отбирает задания существенно более низкого качества, а не отклоняет их случайным образом.

Согласованность между автоматической верификацией и экспертной бинарной оценкой составила $\kappa = 0{,}71$. Основной источник расхождений связан с заданиями, имеющими единичный дефект по агенту «DifficultyMatch», эксперты оценивали такие задания выше порога, несмотря на формальное несоответствие уровню сложности.

Для сравнительной оценки использовались подходы, реализованные на той же большой языковой модели. Использование единой базовой модели обеспечивает методологически корректное сравнение, позволяя изолировать эффект архитектуры верификации от эффекта выбора генератора.

Рассматривались следующие базовые решения:

- Baseline-NoVerif – прямая генерация без верификации (1 LLM-запрос);
- Baseline-SingleAgent – одноагентная верификация, где единый LLM-промпт проверяет все 7 критериев одновременно (2 LLM-запроса: генерация и верификация);
- Baseline-Majority – мультиагентная верификация с правилом большинства $\sum s_j^{(1)} \geq 4$, без второй фазы и без механизма адресной обратной связи (8 LLM-запросов: 1 + 7 параллельных).

Введём показатель эффективности верификации:$\eta = 1 - \frac{n_{\text{def}}^{\text{out}}}{n_{\text{out}}}$, где $n_{\text{def}}^{\text{out}}$ – число дефектных заданий в выходном потоке, n_{out} – объём выходного потока. Результаты сравнения предложенного алгоритма с базовыми решениями представлены в таблице 3.11. Предложенный алгоритм обеспечивает наибольшую эффективность верификации, доля бездефектных заданий в выходном потоке достигает 93,5%. По сравнению с генерацией без верификации прирост вычислительной эффективности достигается при помощи повторной генерации с адресной обратной связью и запросами второй фазы. Оба указанных фактора непосредственно направлены на повышение качества итогового набора заданий.

Таблица 3.11 – Сравнение подходов к контролю качества тестовых заданий

Алгоритм	**Доля дефектных** $(1-\eta)$	η	**Среднее число LLM-запросов**
Baseline-NoVerif	41,9%	58,1%	1
Baseline-SingleAgent	18,4%	81,6%	2
Baseline-Majority	12,1%	87,9%	8
Предложенный алгоритм	6,5%	93,5%	$\approx 12{,}3$

Преимущество по сравнению с одноагентной верификацией обусловлено разделением критериев между специализированными агентами. При объединении всех критериев в одном промпте эффективность снижается с 93,5% до 81,6%, что подтверждает необходимость локализации экспертных ролей. Также это согласуется с общим выводом о том, что мультиагентное взаимодействие превосходит механизм единого агента при решении сложных многокритериальных задач [31].

Преимущество относительно мультиагентной системы с правилом большинства определяется совокупностью механизмов строгого конъюнктивного правила допуска, которое исключает принятие задания при наличии даже одного нарушения. Вторая фаза верификации обеспечивает возможность уточнения решений в противоречивых по содержанию случаях, а перегенерация с адресной обратной связью снижает вероятность повторного возникновения ранее выявленных нарушений. Существенный вклад в повышение качества вносит ребро $A_1 \to A_3$, обеспечивающее выявление класса заданий, в которых агент AnswerValidity на первой фазе не фиксирует проблему, однако после получения сигнала о фактической ошибке в помеченном правильном ответе пересматривает своё решение.

Для оценки вклада каждого агента было проведено абляционное исследование. Алгоритм запускался с последовательным исключением одного агента при сохранении строгого конъюнктивного правила для оставшихся. Результаты абляционного исследования приведены в таблице 3.12. Абляционное исследование показывает, что исключение любого из семи агентов приводит к

увеличению доли дефектных заданий в итоговом потоке. Следовательно, каждый агент вносит самостоятельный вклад в общую эффективность системы.

Таблица 3.12 – Результаты абляционного исследования

Конфигурация	**Доля дефектных**	**Прирост дефектности**
Все 7 агентов (базовая конфигурация)	6,5%	–
Без FactChecker	15,4%	+8,9 п.п.
Без AnswerValidity	13,9%	+7,4 п.п.
Без ContextGrounding	11,8%	+5,3 п.п.
Без Clarity	11,2%	+4,7 п.п.
Без BiasDetection	9,7%	+3,2 п.п.
Без DifficultyMatch	9,4%	+2,9 п.п.
Без Originality	9,1%	+2,6 п.п.

Наибольший вклад обеспечивают агенты FactChecker и AnswerValidity, что содержательно согласуется с педагогической значимостью наиболее критичных дефектов. Фактические ошибки и некорректная структура ответа непосредственно снижают диагностическую ценность задания.

Обобщённые выводы о механизмах верификации могут быть сформулированы следующим образом.

- Специализация каждого агента на единственном критерии является важным условием эффективности. Объединение критериев в один оценочный запрос приводит к снижению качества отбора.
- Двухфазный протокол с межагентным взаимодействием, реализуемым по ориентированному ациклическому графу, обеспечивает уточнение решений в неоднозначных случаях без нарушения независимости предыдущей фазы и без возникновения циклических пересмотров.
- Строгое конъюнктивное правило допуска исключает принятие задания при наличии хотя бы одного нарушения. За счёт этого обеспечивается высокая чувствительность к дефектным заданиям, что методически оправдано в контексте формирования контрольно-измерительных материалов.

- Алгоритм демонстрирует стабильную эффективность для всех поддерживаемых форматов. Это подтверждает универсальность предложенной архитектуры и отсутствие выраженных ограничений для отдельных типов заданий.

3.4. Выводы по третьей главе

В главе представлены результаты разработки и исследования алгоритма генерации многоформатных тестовых заданий с мультиагентной верификацией. Разработана формальная постановка задачи генерации многоформатных тестовых заданий. Она включает в себя множество поддерживаемых форматов, параметр сложности, формализованную структуру ответа для каждого формата, а также систему из критериев качества, которые отражают взаимодополняющие аспекты корректности тестового задания.

Предложен алгоритм, интегрирующий LLM-генератор и мультиагентную систему верификации. Каждый агент определяется как специализированная экспертная сущность, использующая собственную системную инструкцию, локальное подмножество входных данных и набор инструментов.

Разработан двухфазный протокол верификации. Первая фаза обеспечивает независимую параллельную оценку задания по критериям. Вторая фаза реализует избирательное межагентное взаимодействие по ориентированному ациклическому графу. Показано, что такая организация взаимодействия позволяет учитывать зависимости между критериями без утраты формальной управляемости протокола. Введено строгое конъюнктивное правило допуска, исключающее принятие задания при наличии хотя бы одного критического нарушения.

Экспериментальное исследование на корпусе из заданий пяти форматов показало снижение доли дефектных заданий с 41,9% до 6,5%. Сравнительный анализ с базовыми решениями показал увеличение доли заданий без выявленных нарушений до 93,5% при среднем числе около 12,3 LLM-запросов на одно принятое задание. Абляционное исследование также подтвердило самостоятельный вклад каждого из семи агентов в общую эффективность системы.

Глава 4. Алгоритм генерации дистракторов на основе когнитивного моделирования ошибок обучающихся

Среди компонентов тестового задания с выбором ответа дистракторы обладают наибольшей диагностической значимостью. Их качество в значительной мере определяет способность задания дифференцировать обучающихся с различным уровнем освоения материала, где явно некорректные варианты не обеспечивают диагностически значимого эффекта отвлечения от правильного ответа даже для слабо подготовленных испытуемых, тогда как варианты, воспроизводящие типичные заблуждения, позволяют выявить характер затруднений и скорректировать образовательный процесс [82, 52]. Как отмечено в исследовании [52], разработка высококачественных дистракторов является наиболее трудоёмкой частью разработки заданий на множественный выбор и выполняется без достаточного методологического обоснования.

Несмотря на значительный прогресс в области автоматической генерации дистракторов, особенно связанный с применением LLM, главной нерешённой задачей остаётся педагогическая обоснованность генерируемых вариантов ответа [46, 104]. Большинство алгоритмов ограничиваются семантической близостью к правильному ответу или прямой генерацией. В таком случае формируются варианты, фактически являющиеся ответами в альтернативной формулировке или тривиально неверные. Исследование [46] на математических задачах MCQ показало, что LLM, несмотря на формальную корректность сгенерированных вариантов, показывает низкое соответствие типичным ошибкам обучающихся.

Данная глава посвящена алгоритму генерации дистракторов, отличающемуся от существующих решений способом формирования неверных вариантов. В отличие от подходов, основанных на поиске семантически близких к правильному ответу выражений или прямой генерации неверных вариантов ответа, предлагается формировать корректную цепочку рассуждений. Она приводит к правильному ответу, с последующим целенаправленным внесением когнитивных ошибок, характерных для типичных заблуждений обучающихся. Дистракторы, полученные

таким образом, воспроизводят логику ошибочных рассуждений и обеспечивают высокую педагогическую значимость при сохранении тематической релевантности. Идея использования ошибочных рассуждений в качестве основы для генерации дистракторов опирается на принцип Chain-of-Exemplar [78] и расширяет его за счёт явной типизации когнитивных ошибок применительно к русскоязычному образовательному контенту.

Глава включает формальную постановку задачи, описание трёхэтапного алгоритма, результаты экспериментального исследования на двух русскоязычных наборах данных, а также описание программного комплекса, реализующего полный алгоритмический конвейер.

4.1. Формальная постановка задачи генерации дистракторов

Задача автоматической генерации дистракторов формализуется как отображение входного представления (Q, A, C) в множество неверных вариантов ответа D^*, удовлетворяющих системе требований.

Определение 4.1 (Вопрос). Вопросом Q называется текстовая формулировка, отражающая проверяемый содержательный аспект и однозначно задающая требуемый тип ответа. В рамках рассматриваемого конвейера $Q \equiv q$ – формулировка задания, полученная в результате работы алгоритма главы 3.

Определение 4.2 (Правильный ответ). Правильным ответом A называется краткое или развёрнутое утверждение, корректно отвечающее на Q в соответствии с контентом C. В структуре $V(q)$ это означает верный элемент для формата «single» или один из верных вариантов для формата «multiple».

Определение 4.3 (Контент). Контент $C \equiv S^*$ – оптимальная аннотация, выбранная алгоритмом главы 2 по критерию (2.4). Он служит фактической основой для проверки корректности цепочек рассуждений и тематической релевантности дистракторов.

Определение 4.4 (Корректная цепочка рассуждений). Корректной цепочкой рассуждений R_c называется упорядоченная последовательность логических шагов

$(r_1, r_2, \dots, r_m)$, связывающих вопрос Q с ответом A через последовательно обоснованные промежуточные выводы:

$$R_c = (r_1 \to r_2 \to \cdots \to r_m \to A) \qquad (4.1)$$

где каждый шаг r_i является текстовым утверждением, логически следующим из предыдущих шагов, вопроса Q и соответствующим содержанию C.

Определение 4.5 (Ошибочная цепочка рассуждений). Ошибочной цепочкой рассуждений R_e называется последовательность, полученная из R_c путём целенаправленной модификации одного или нескольких шагов, приводящая к неверному итоговому выводу:

$$R_e = (r_1 \to \cdots \to r_{k-1} \to \hat{r}_k \to \hat{r}_{k+1} \to \cdots \to \hat{r}_m \to d) \qquad (4.2)$$

где $\hat{r}_k$ – модифицированный шаг, содержащий когнитивную ошибку конкретного типа, $d \neq A$ – неверный итоговый вывод.

Определение 4.6 (Дистрактор). Дистрактором d_i называется краткое утверждение, сделанное на основе итогового вывода ошибочной цепочки $R_e^{(i)}$ и являющееся неверным, но правдоподобным с точки зрения обучающегося, совершившего характерную когнитивную ошибку, ответом на Q.

Определение 4.7 (Задача ADG). Задача автоматической генерации дистракторов формулируется как вычисление:

$$D^* = \{d_1, d_2, \dots, d_N\} = \mathrm{ADG}(Q, A, C;\ N) \qquad (4.3)$$

где N – требуемое число дистракторов ($N = 3$ для стандартного MCQ с четырьмя вариантами), D^* – множество валидных дистракторов, удовлетворяющих используемым ограничениям.

На основе анализа педагогической и психометрической литературы [82, 52, 112] сформулированы три группы требований.

Группа 1. Правдоподобие. Дистрактор должен восприниматься как потенциально верный ответ обучающихся с недостаточным уровнем освоения материала. Он не должен содержать явных логических противоречий и содержательно некорректных утверждений, распознаваемых без привлечения специальных знаний. При наличии соответствующего уровня подготовки дистрактор должен однозначно идентифицироваться как неверный. Согласно

обзору [52], именно правдоподобие является ключевым критерием качества дистрактора с точки зрения психометрии. Нефункционирующие дистракторы, выбираемые менее чем 5% испытуемых, не обеспечивают различия обучающихся по уровню подготовки и снижают эффективность теста.

Правдоподобие $\mathrm{Pl}(d_i) \in \{0,1,2\}$ оценивается по трёхбалльной экспертной шкале (таблица 4.3). Целевое значение: $\bar{\mathrm{Pl}} \geq 1.5$ в среднем по набору.

Группа 2. Тематическая релевантность. Дистрактор должен принадлежать той же предметной области и тематическому контексту, что Q и A. Тематически нерелевантный вариант отвергается обучающимися без специальных знаний на основании несоответствия теме. Релевантность измеряется через BERTScore [132]:

$$\mathrm{Rel}(d_i, Q) = \mathrm{BERTScore}(d_i, Q) \geq \theta_{\mathrm{rel}} \qquad (4.4)$$

Группа 3. Диагностическая ценность. Дистрактор должен отражать существующие когнитивные заблуждения обучающихся, характерные для данной предметной области. Диагностически ценный дистрактор позволяет идентифицировать характер ошибки, включая неверное обобщение, смешение понятий, ошибочное применение правила и т.д. [52]. Данное требование является ключевым отличием предлагаемого подхода от алгоритмов семантической трансформации и согласуется с теорией когнитивной нагрузки [112], согласно которой качественный дистрактор активирует те же когнитивные процессы, что и правильный ответ, создавая ситуацию подлинного выбора.

Вместе с требованиями к отдельным дистракторам вводятся ограничения на совокупность $D^* = \{d_1, d_2, \dots, d_N\}$:

– Семантическое различие с правильным ответом. Каждый дистрактор должен быть достаточно удалён от правильного ответа:

– Внутригрупповое разнообразие. Дистракторы в наборе должны быть достаточно различны между собой, охватывая широкий спектр возможных заблуждений:

– Уникальность и нетождественность правильному ответу.

Совокупность ограничений определяет пространство допустимых множеств дистракторов и учитывается на этапе валидации. Пороговые значения $\theta_{\text{rel}} = 0.50$, $\theta_{\text{dist}} = 0.25$, $\theta_{\text{div}} = 0.20$ определены на валидационной выборке.

В таблице 4.1 приведено сводное представление элементов задачи.

Таблица 4.1 – Представление элементов задачи генерации дистракторов

Наименование	**Роль в алгоритме**	**Источник**
Вопрос (формулировка задания)	Входные данные	Алгоритм Гл. 3, формула (3.4)
Правильный ответ	Входные данные	Алгоритм Гл. 3, $V(q)$
Контент	Входные данные	Алгоритм Гл. 2, S^*, формула (2.4)
Корректная цепочка	Промежуточный	Этап 1, формула (4.5)
Ошибочная цепочка i	Промежуточный	Этап 2, формула (4.6)
Дистрактор-кандидат	Промежуточный	Итоговый вывод $R_e^{(i)}$
Множество валидных дистракторов	Выход	После валидации, этап 3

4.2. Описание алгоритма

Алгоритм реализует трёхэтапную процедуру, основанную на моделировании педагогической ошибки. На первом этапе формируется корректная цепочка рассуждений (этап 1), на втором этапе в неё вносятся типизированные искажения, соответствующие определённым когнитивным заблуждениям (этап 2), после чего из полученных ошибочных цепочек извлекаются и проходят валидацию кандидаты в дистракторы (этап 3).

На начальном этапе языковая модель формирует цепочку рассуждений R_c, приводящую к правильному ответу A на вопрос Q с опорой на контент C:

$$R_c = LLM(P_{\text{correct}}), P_{\text{correct}} = f(Q, A, C) \qquad (4.5)$$

где f – представляет собой функцию формирования инструкции, объединяющую системные указания, задающие роль модели, текстовый контент C, вопрос Q и правильный ответ A. LLM обозначает языковую модель с режимом рассуждений, который на данном этапе имеет важное значение, поскольку модель формирует развёрнутую цепочку последовательности шагов перед выдачей итогового ответа,

что способствует повышению логической связности и фактической корректности r_i.

Инструкция задаёт представление рассуждения в виде нумерованной последовательности из 3–5 шагов, которые завершаются правильным ответом. Выбор указанного количества обоснован тем, что цепочка из 1–2 шагов не обеспечивает достаточного числа позиций для внесения ошибок на следующем этапе. Цепочка из 6 и более шагов оказывается избыточной вследствие экспоненциального увеличения числа возможных ошибочных ветвей и снижения связности формируемых дистракторов. Указанный диапазон согласуется с рекомендациями по проектированию цепочек рассуждений для сохранения логической прозрачности.

Корректность цепочки R_c проверяется по двум критериям, включающим соответствие итогового вывода правильному ответу A через $\text{BERTScore}(R_c.\text{conclusion}, A) \geq 0.85$; а также логическую согласованность последовательности шагов, определяемую посредством дополнительного запроса к языковой модели. В случае несоблюдения хотя бы одного из критериев выполняется повторная генерация цепочки.

Второй этап является основным с точки зрения научной новизны. M принимает корректную цепочку R_c и генерирует $N_{\text{cand}} \geq N$ ошибочных вариантов с явным указанием типа вносимой ошибки:

$$R_e = LLM(P_{\text{mistakes}}), P_{\text{mistakes}} = g(R_c, N_{\text{cand}}) \qquad (4.6)$$

где g – функция формирования промпта, предписывающего сгенерировать N_{cand} ошибочных вариантов с явной типизацией ошибки.

Система типов когнитивных ошибок разработана на основе теории когнитивной нагрузки J. Sweller [112] и исследований диагностики учебных ошибок [82, 52]. В рамках данной системы выделяются следующие типы ошибок:

– Неверное обобщение (overgeneralization) представляет собой применение правила, корректного в ограниченном контексте, к более широкому классу явлений. В качестве примера можно привести ситуацию, в которой закон, справедливый для замкнутой системы, используется без учёта граничных условий.

– Подмена смежных понятий (concept confusion) заключается в подмене одного концепта другим, внешне сходным, но отличающимся по смыслу. Характерным примером является смешение понятий «скорость» и «ускорение».

– Ошибка порядка следования (sequence error) выражается в нарушении причинно-следственной или хронологической последовательности шагов. Данный тип ошибок наиболее значим для заданий, требующих восстановления причинно-следственных цепочек, поскольку именно в них нарушение последовательности шагов приводит к принципиально иному выводу.

– Неполная цепочка (incomplete reasoning) характеризуется пропуском существенного промежуточного шага, что приводит к формированию неверного вывода.

– Ошибочная количественная оценка (quantitative error) связана с использованием численно близкого, но некорректного значения параметра.

Требование явного указания типа ошибки выполняет две взаимосвязанные функции. С одной стороны, оно направляет процесс генерации к педагогически значимым типам заблуждений. С другой стороны, оно формирует диагностическую метку дистрактора, используемую преподавателем при интерпретации результатов тестирования. Параметр $N_{\text{cand}} = 2N$ обеспечивает формирование резервного множества кандидатов для последующей валидации.

В отличие от подхода «overgenerate-and-rank», предложенного Scarlatos и соавторами [104], реализованный алгоритм управляет типом ошибки на этапе генерации, а не ограничивается последующим ранжированием кандидатов, что обеспечивает более высокую педагогическую обоснованность формируемого набора дистракторов.

На третьем этапе из ответа R_e извлекается множество кандидатов:

$$D_{\text{cand}} = \{d_1^{\text{cand}}, d_2^{\text{cand}}, \dots, d_{N_{\text{cand}}}^{\text{cand}}\} = \text{parse}(R_e) \qquad (4.7)$$

Каждый кандидат проходит трёхступенчатую валидацию.

Шаг 1 характеризует проверку уникальности и нетождественности. Обеспечивается лексическое различие с A и отсутствие дубликатов. Проверка выполняется детерминированно без привлечения языковой модели.

Шаг 2 соответствует проверке тематической релевантности. Для каждого кандидата вычисляется значение $\mathrm{BERTScore}(d_i, Q)$, после чего кандидаты, для которых $\mathrm{BERTScore}(d_i, Q) < \theta_{\mathrm{rel}}$, исключаются.

Шаг 3 связан с оценкой семантического различия с правильным ответом. Вычисляется значение $\mathrm{BERTScore}(d_i, A)$, и кандидаты, для которых $\mathrm{BERTScore}(d_i, A) > 1 - \theta_{\mathrm{dist}}$, исключаются.

Множество кандидатов, прошедших все три этапа проверки, обозначается D_{val}. Из него отбирается N дистракторов с максимальным внутригрупповым разнообразием в соответствии с

$$D^* = \arg \max_{D \subseteq D_{\mathrm{val}},\, |D|=N} \mathrm{Div}(D) \qquad (4.8)$$

При $|D_{\mathrm{val}}| \geq N$ задача (4.8) при $N = 3$ и $|D_{\mathrm{val}}| \leq 12$ решается полным перебором; при большем числе кандидатов используется алгоритм отбора, направленный на максимизацию разнообразия. При $|D_{\mathrm{val}}| < N$ весь цикл повторяется с расширенным набором параметров g.

4.3. Экспериментальное исследование алгоритма

В качестве экспериментальной базы использованы открытые русскоязычные наборы данных с вручную сформированными дистракторами, что обеспечивает возможность количественной верификации результатов на основе эталонных ответов.

RuOpenBookQA является русскоязычной версией набора OpenBookQA [82] и содержит вопросы по естественно-научным дисциплинам начального уровня. Структура каждого задания предполагает наличие набора опорных фактов, на основе которых формируется вопрос, а также четырёх вариантов ответа. Среди них один является корректным, тогда как три остальных представляют собой дистракторы, полученные в результате ручной разметки. Дистракторы эталонного набора характеризуются высоким уровнем правдоподобия с точки зрения содержания, что требует наличия предметных знаний для их корректного исключения. Согласно исследованию [82], точность ответов человека на данном

наборе составляет около 92%, указывая на умеренную сложность вопросов и высокое качество дистракторов.

RuWorldTree представляет собой русскоязычную адаптацию набора WorldTree V2.0 [126]. Набор ориентирован на задачи последовательного рассуждения. Каждый вопрос сопровождается аннотированным графом объяснений, представляющим собой связную цепочку. Она включает в среднем около шести фактов из таблицы объяснений, обосновывающих правильный ответ. Такая структура позволяет исследовать применимость предложенного алгоритма к вопросам составного причинно-следственного вывода.

Характеристики наборов данных приведены в таблице 4.2.

Таблица 4.2 – Характеристика экспериментальных наборов данных

Характеристика	**RuOpenBookQA**	**RuWorldTree**
Предметная область	Естественные науки	Естественные науки
Использовано в экспериментах	200 (случайная выборка)	200 (случайная выборка)
Число дистракторов на вопрос	3	3
Сгенерировано дистракторов (на алгоритм)	600	600
Эталонные дистракторы	Да (вручную)	Да (вручную)
Граф объяснений	Нет	Да (ср. 6 фактов на вопрос)
Средняя длина вопроса (слов)	12,4	15,7
Характер рассуждений	Одношаговые / двухшаговые	Многошаговые, составные

Из каждого набора случайным образом отбирались 200 вопросов; для каждого вопроса все три алгоритма формировали по 3 дистрактора. Таким образом, для каждого алгоритма и набора оценивались 600 дистракторов.

Для сопоставления использовались два подхода, представляющие наиболее распространённые классы методов ADG:

– SM (Semantic Modification). Семантическое изменение, при котором правильный ответ A трансформируется заменой ключевых слов семантически

близкими альтернативными вариантами на основе векторных представлений. Определяет класс лексико-семантических методов (раздел 1.2).

– DG (Direct Generation). Прямая генерация, где M формирует дистракторы по инструкции, которая содержит указание на создание трёх вариантов неверных ответов без построения цепочки рассуждений. Этот подход представляет класс LLM-ориентированных методов без когнитивного моделирования и содержательно близок к подходу, рассмотренному в работе [46] для математических задач.

Во всех трёх алгоритмах (SM, DG и предложенном) использовалась одна и та же языковая модель и одинаковые входные данные. Постановка эксперимента позволяет изолировать влияние алгоритмического механизма от влияния выбора языковой модели.

Качество дистракторов оценивалось по пяти показателям:

– Экспертная оценка правдоподобия $\mathrm{Pl}(d_i) \in \{0,1,2\}$ проводилась по трёхбалльной шкале (таблица 4.3) пятью независимыми экспертами по каждому набору данных.

– Семантическое сходство с формулировкой задания $\mathrm{sim}(d_i, Q) = \mathrm{BERTScore}(d_i, Q)$ характеризующее тематическую релевантность.

– Различие с правильным ответом $\mathrm{dist}(d_i, A) = 1 - \mathrm{BERTScore}(d_i, A)$, отражающее степень семантического расхождения.

– Сходство с эталонными дистракторами $\mathrm{BERTScore}(d_i, d_{\mathrm{ref}})$ позволяющее оценить близость к вручную сформированным вариантам.

– Внутригрупповое разнообразие $\mathrm{Div}(D)$, характеризует охват различных типов заблуждений в выбранном наборе дистракторов.

Процедура экспертной оценки. Пять независимых экспертов оценивали каждый дистрактор по указанной шкале без информации об источнике его генерации. Межэкспертная согласованность контролировалась с использованием коэффициента каппа Флейса; при значении $\kappa < 0.60$ для конкретного задания оценки согласовывались в ходе обсуждения. Итоговый балл вычислялся как среднее арифметическое оценок всех пяти экспертов.

Таблица 4.3 – Шкала экспертной оценки дистракторов

Балл	**Категория**	**Критерий оценки**
0	Неправдоподобно	Дистрактор содержит явные противоречия; исключается любым обучающимся без специальных знаний
1	Частично правдоподобно	Может ввести в заблуждение при поверхностном уровне знаний; при внимательном анализе ошибка выявляется
2	Полностью правдоподобно	Требует чёткого знания предметного материала для исключения и воспроизводит типичное заблуждение

Результаты, представленные в таблице 4.4, ранее опубликованы автором в статье [4].

Таблица 4.4 – Оценка качества дистракторов на наборе RuOpenBookQA

Способ	**Экспертная оценка**	**Сходство дистрактора с вопросом**	**Различие дистрактора с ответом**	**Сходство с эталонными дистракторами**	**Внутригрупповое разнообразие**
Семантическое изменение	1,07	0,60	0,55	0,35	0,52
Базовая генерация	1,20	0,68	0,62	0,42	0,61
Предлагаемый алгоритм	**1,66**	**0,72**	**0,68**	**0,58**	**0,74**

Предложенный алгоритм превосходит оба базовых решения по всем пяти показателям. Наиболее выраженное преимущество наблюдается по ключевому показателю правдоподобия: 1,66 против 1,20 (DG) и 1,07 (SM). Значение показателя сходства с эталонными дистракторами (0,58) указывает на высокую близость автоматически генерируемых вариантов ответа к дистракторам, созданным вручную.

В таблице 4.5 приведено распределение типов когнитивных ошибок в сгенерированных дистракторах на наборе RuOpenBookQA. Наиболее распространёнными являются подмена понятий (29,0%) и неверное обобщение (26,0%), при этом именно они характеризуются наибольшими значениями экспертной оценки (1,73 и 1,68 соответственно).

Таблица 4.5 – Распределение типов когнитивных ошибок (RuOpenBookQA)

Тип ошибки	**Число**	**Доля, %**	**Средняя** $\bar{P}$l
Подмена смежных понятий	174	29,0	1,73
Неверное обобщение	156	26,0	1,68
Ошибочная количественная оценка	102	17,0	1,61
Неполная цепочка рассуждений	96	16,0	1,59
Ошибка порядка следования	72	12,0	1,57
Итого	**600**	**100,0**	**1,66**

Данный результат согласуется с выводами педагогических исследований [112, 52], согласно которым указанные типы ошибок отражают устойчивые искажения в понимании предметного материала, возникающие при его поверхностном усвоении.

Результаты на наборе RuWorldTree представлены в таблице 4.6. Ранее они были опубликованы автором в статье [4].

Таблица 4.6 – Оценка качества дистракторов на наборе RuWorldTree

Способ	**Экспертная оценка**	**Сходство дистрактора с вопросом**	**Различие дистрактора с ответом**	**Сходство с эталонными дистракторами**	**Внутригрупповое разнообразие**
Семантическое изменение	0,98	0,47	0,52	0,31	0,51
Базовая генерация	1,13	0,53	0,60	0,40	0,57
Предлагаемый алгоритм	**1,60**	**0,69**	**0,65**	**0,56**	**0,64**

Результаты, полученные на наборе RuWorldTree, согласуются с результатами RuOpenBookQA, при этом предложенный алгоритм устойчиво превосходит базовые решения по всем пяти показателям. Снижение абсолютных значений по сравнению с RuOpenBookQA обусловлено большей сложностью многошаговых вопросов, поскольку генерация правдоподобных дистракторов для заданий, требующих последовательного объяснения [126], предполагает более глубокое понимание причинно-следственных связей.

Снижение показателя разнообразия (0,64 против 0,74) объясняется более узкой тематической направленностью таких вопросов, поскольку для многошаговых рассуждений множество реалистичных ошибок оказывается более ограниченным по сравнению с вопросами, требующими одношагового ответа.

Показатель 22,7%, заявленный в научной новизне диссертации, вычисляется как среднее значение относительного прироста по пяти показателям качества относительно алгоритма прямой генерации (DG) на наборе RuOpenBookQA. Данный набор выбран как основной вследствие более стандартизированной структуры вопросов, обеспечивающей наибольшую сопоставимость результатов. Расчёт приведён в таблице 4.7.

Таблица 4.7 – Расчёт сводного показателя прироста качества дистракторов

Показатель	**DG**	**Предложенный алгоритм**	**Относительный прирост Δ_j, %**
Экспертная оценка	1,20	1,66	$\frac{1{,}66-1{,}20}{1{,}20}\times 100=+38{,}3$
Сходство с вопросом	0,68	0,72	$\frac{0{,}72-0{,}68}{0{,}68}\times 100=+5{,}9$
Различие с правильным ответом	0,62	0,68	$\frac{0{,}68-0{,}62}{0{,}62}\times 100=+9{,}7$
Сходство с эталонными дистракторами	0,42	0,58	$\frac{0{,}58-0{,}42}{0{,}42}\times 100=+38{,}1$
Внутригрупповое разнообразие	0,61	0,74	$\frac{0{,}74-0{,}61}{0{,}61}\times 100=+21{,}3$
Среднее $\bar{\Delta}$	–	–	$\frac{38{,}3+5{,}9+9{,}7+38{,}1+21{,}3}{5}\approx 22{,}7\%$

Формализованное определение сводного показателя прироста имеет вид:

$$\bar{\Delta}=\frac{1}{5}\sum_{j=1}^{5}\frac{m_j^{\text{alg}}-m_j^{\text{DG}}}{m_j^{\text{DG}}}\times 100\%\approx 22{,}7\% \qquad (4.9)$$

Таким образом, показатель 22,7 % отражает не частное улучшение по одному критерию, а усреднённый эффект по совокупности диагностически значимых характеристик дистракторов.

Следует выделить три аспекта интерпретации данного показателя. Во-первых, наибольший прирост достигается по наиболее значимым критериям, а именно по экспертной оценке (+38,3%) и сходству с эталонными дистракторами (+38,1%), что отражает содержательное улучшение диагностической ценности дистракторов. Во-вторых, умеренный прирост по показателю тематического сходства с вопросом (+5,9%) обусловлен тем, что базовый алгоритм DG уже обеспечивает достаточную тематическую близость, и потенциал дополнительного улучшения здесь ограничен. В-третьих, сводный показатель 22,7% отражает усреднённый эффект по всем пяти показателям, тогда как по ключевым диагностическим критериям прирост выше.

Статистическая значимость различий проверялась критерием Вилкоксона для парных сравнений ($n = 200$ вопросов × 3 дистрактора).

В таблице 4.8 приведено сравнение прироста относительно прямой генерации на обоих наборах данных.

Таблица 4.8 – Сравнение прироста относительно DG

Показатель	**Прирост, RuOpenBookQA**	**Прирост, RuWorldTree**
Экспертная оценка	+38,3%	+41,6%
Сходство с вопросом	+5,9%	+30,2%
Различие с правильным ответом	+9,7%	+8,3%
Сходство с эталоном	+38,1%	+40,0%
Разнообразие	+21,3%	+12,3%
Среднее	**+22,7%**	**+26,5%**

Устойчивость результатов на наборе RuWorldTree подтверждает, что предложенный алгоритм показывает преимущество не только на задачах с более простой структурой, но и на задачах последовательного рассуждения.

Можно выделять основные механизмы, обеспечивающие преимущество алгоритма:

Структурная согласованность. Каждый дистрактор формируется как результат цепочки рассуждений, которая совпадает с правильным ответом на всех шагах до момента внесения ошибки. Это обеспечивает тематическую релевантность, потому что дистрактор семантически близок и логически связан с тем же рассуждением. В отличие от прямой генерации [46], где тематическая связность достигается преимущественно за счёт контекста промпта, предложенный подход обеспечивает её за счёт внутренней структуры рассуждения.

Когнитивная обоснованность. Явная типизация ошибок направляет процесс генерации к формам ошибок, описанным в педагогических исследованиях [112, 52], в отличие от алгоритма DG, который нередко формирует семантически близкие, но слабо соответствующие реальным когнитивным искажениям варианты ответа. Полученные результаты согласуются с выводами Feng и соавторов [46], согласно которым языковые модели при прямой генерации обладают ограниченной способностью воспроизводить реальные ошибки обучающихся.

Абляционное исследование. Алгоритм запускался в двух конфигурациях, с полной типизацией ошибок и без указания типа ошибки в промпте. Результаты представлены в таблице 4.9.

Таблица 4.9 – Вклад компонентов алгоритма (RuOpenBookQA)

Конфигурация	**Экспертная оценка**	**Сходство с эталоном**	**Разнообразие**
Без цепочки рассуждений (= DG)	1,20	0,42	0,61
Цепочка без типизации ошибок	1,41	0,47	0,68
Полная (цепочка + типизация)	**1,66**	**0,58**	**0,74**
Прирост от цепочки (без типизации)	+17,5%	+11,9%	+11,5%
Прирост от типизации (при цепочке)	+17,7%	+23,4%	+8,8%

Абляционное исследование позволяет выделить два сопоставимых по величине компонента прироста. Использование цепочки рассуждений без типизации обеспечивает увеличение показателей примерно на 17–18%, а дополнительная типизация ошибок приводит к дальнейшему росту на 17–23%. Исключение любого из компонентов приводит к заметному снижению качества, что подтверждает необходимость их совместного использования и взаимодополняющий характер.

4.4. Программный комплекс генерации контрольно-измерительных материалов

Разработанный программный комплекс реализует последовательность взаимосвязанных этапов автоматизированного формирования контрольно-измерительных материалов и включает одиннадцать программ для ЭВМ, зарегистрированных в Федеральной службе по интеллектуальной собственности (Роспатент). Все программные компоненты комплекса реализованы на языке Python, что обеспечило единообразие средств разработки, повторное использование библиотек обработки естественного языка и машинного обучения, а также унификацию межмодульного взаимодействия.

Полный перечень программ с указанием их функциональных ролей:

1. «Автоматизированная система генерации контрольно-измерительных материалов». Управляющая система; генерация контрольно-измерительных материалов из текстовых источников; формирование заданий закрытого типа, с множественным выбором, открытой формы, на упорядочение и установление соответствия; формирование итогового документа с вопросами и вариантами ответов.

2. «Подсистема верификации контрольно-измерительных материалов». Автоматическая верификация сформированных заданий; структурная проверка в зависимости от типа вопроса; интеллектуальный анализ возможных ошибок в заданиях.

3. «Адаптивный модуль формирования дистракторов». Автоматизированная генерация неверных вариантов ответа на основе текста вопроса и правильного ответа.

4. «Интеллектуальная система анализа качества дистракторов». Оценка качества дистракторов по показателям релевантности, логической связности, грамматической корректности и уникальности.

5. «Инструмент оценки контентной валидности теста». Расчёт показателей контентной валидности, оценка внутренней согласованности и формирование рекомендаций по доработке структуры теста.

6. «Подсистема автоматической оценки правильности ответов». Автоматическая оценка ответов на задания различных типов с учётом контекста вопроса.

7. «Модуль анализа соответствия тестовых заданий образовательным стандартам». Сопоставление сформированных заданий с учебными целями и образовательными стандартами; выделение ключевых компетенций; формирование отчётов о покрытии стандартов.

8. «Инструмент автоматической категоризации тестовых вопросов». Анализ заданий, их тематическая классификация, оценка сложности, распределение по категориям и визуализация результатов.

9. «Модуль автоматического извлечения структурированной информации из графических данных». Извлечение формализованных данных из изображений, диаграмм и иных графических материалов для последующей передачи в управляющую систему.

10. «Инструмент генерации вопросов причинно-следственных связей». Специализированная генерация вопросов, направленных на выявление понимания причинно-следственных зависимостей в текстах.

11. «Модуль формирования тестовых заданий на основе табличных данных». Автоматизированное формирование заданий по структурированным числовым и категориальным данным, включая задания на интерпретацию данных.

Такое распределение функций позволяет объединить в единую систему алгоритмы диссертационной работы и прикладные инструменты, ориентированные на различные образовательные сценарии.

Взаимодействие с языковой моделью организовано через унифицированный программный слой. Это позволяет использовать стандартизированный механизм обращения к модели в различных модулях комплекса, включая генерацию заданий, их верификацию, построение дистракторов и автоматическую оценку ответов. Единый способ интеграции обеспечивает технологическую согласованность компонентов и упрощает сопровождение программного комплекса в процессе эксплуатации.

Программный комплекс охватывает не только генерацию заданий в узком смысле, но и последовательность технологических процедур, включающую подготовку входных данных, семантический анализ, сопоставление с образовательными требованиями, верификацию, генерацию дистракторов, оценку ответов и формирование итоговых материалов. За счёт модульной организации комплекс может использоваться как целиком, так и по отдельным функциональным направлениям, что расширяет возможности его использования в различных образовательных платформах и сценариях внедрения.

Сопоставление ручной разработки и автоматизированного формирования контрольно-измерительных материалов представлено в таблице 4.10. Полученные результаты показывают, что разработанный программный комплекс обеспечивает не только автоматизацию генерации тестовых заданий, но и повышение технологической эффективности их подготовки. Достигаемый эффект обусловлен не снижением требований к контролю качества, а распределением функций между специализированными программами, каждая из которых реализует отдельный этап обработки и тем самым обеспечивает согласованность работы системы. Реализация комплекса на языке Python предусматривает расширяемость архитектуры, упрощает сопровождение и позволяет интегрировать в единую программную среду модули обработки текста, анализа данных, генерации заданий и их автоматизированной проверки.

Таблица 4.10 – Сравнение с ручной разработкой КИМ

Характеристика	**Ручная разработка**	**Программный комплекс**
Время на одно задание с дистракторами	15–20 минут	38–50 секунд
Доля заданий без правок	–	91%
Многоформатность	Ресурсами эксперта	Все 5 форматов автоматически
Обновляемость при изменении контента	Ресурсами эксперта	Автоматически по изменённым разделам
Диагностические метки типов ошибок	Не формируются	Формируются автоматически для каждого дистрактора
Зависимость от предметного эксперта	Высокая	Минимальная, только на этапе итоговой верификации
Поддерживаемые форматы экспорта	Зависит от используемого инструмента	QTI 2.1, GIFT, JSON, XLSX

4.5. Выводы по четвёртой главе

В четвёртой главе сформирована формальная постановка задачи автоматической генерации дистракторов, включающая определения корректной и ошибочной цепочек рассуждений, формализацию задачи автоматической генерации дистракторов, систему требований к правдоподобию, тематической релевантности и диагностической ценности, а также ограничения на семантическое различие, разнообразие и уникальность дистракторов. Такая постановка позволяет различать дистракторы высокой и низкой диагностической ценности и задаёт основу для последующей алгоритмической реализации.

Предложен трёхэтапный алгоритм генерации дистракторов, включающий построение корректной цепочки рассуждений с опорой на исходный текстовый контент, внесение типизированных когнитивных ошибок пяти классов и последующее извлечение, валидацию и отбор дистракторов с максимальным разнообразием. Показано, что механизм явной типизации ошибок обеспечивает педагогическую обоснованность генерируемых вариантов ответа и их содержательное отличие от подходов прямой генерации.

Экспериментальная оценка на наборах RuOpenBookQA и RuWorldTree с участием пяти независимых экспертов подтвердила превосходство предложенного алгоритма над базовыми решениями DG и SM по всем пяти показателям качества. Экспертные оценки составили 1,66 и 1,60 против 1,07–1,20 у базовых подходов.

Абляционное исследование показало, что использование цепочки рассуждений и явной типизации ошибок вносит сопоставимый вклад в итоговый прирост качества, что подтверждает необходимость совместного использования обоих компонентов алгоритма. Также описан разработанный программный комплекс, включающий одиннадцать программ для ЭВМ.

Заключение

Полученные результаты в совокупности подтверждают достижение поставленной цели исследования и решение всех сформулированных задач.

1. Проведен комплексный анализ существующих алгоритмов автоматической генерации тестовых заданий, в результате которого выявлены их ключевые ограничения применительно к русскоязычному материалу (в частности, поверхностное перефразирование, отсутствие механизмов верификации и слабый учет дидактических требований). Доказана необходимость перехода к интегрированным контурам генерации.

2. Сформирован алгоритм семантического анализа образовательного контента и извлечения из него ключевой информации, релевантной для последующей генерации заданий. Доказано экспериментально, что применение параметрической метрики с доменно-специфичными весовыми коэффициентами позволило превзойти все рассмотренные базовые решения, при этом по сравнению с лучшей базовой моделью (O1) усреднённый по предметным областям прирост интегральной оценки качества аннотаций составил 13,2 %.

3. Создан алгоритм генерации многоформатных тестовых заданий, успешно формирующий задания с выбором одного или нескольких вариантов ответа, задания открытой формы, а также на установление соответствия и правильной последовательности. Внедрение в данный алгоритм разработанной мультиагентной системы верификации позволило автоматически исключать некорректные формулировки и снизить общую долю дефектных заданий на 84,5 %.

4. Разработан алгоритм интеллектуального формирования дистракторов на основе когнитивного моделирования ошибок обучающихся. Эксперименты на наборах данных RuOpenBookQA и RuWorldTree подтвердили, что имитация искажений в цепочке рассуждений превосходит подходы на основе семантической близости, повышая интегральное качество (правдоподобность и дидактическую ценность) неверных вариантов ответа на 22,7 %.

5. Созданные алгоритмы реализованы в виде программного комплекса для автоматизированной генерации контрольно-измерительных материалов (получено 11 свидетельств о государственной регистрации программ для ЭВМ). Проведенная экспериментальная апробация подтвердила высокую эффективность предложенных решений и дидактическое качество генерируемых материалов, готовых к интеграции в современные образовательные системы.

Перспективы дальнейших исследований связаны с расширением числа поддерживаемых предметных областей и форматов заданий, разработкой специализированных русскоязычных корпусов и наборов для оценки алгоритмов AQG и ADG, совершенствованием процедур автоматизированной оценки качества заданий, интеграцией внешних баз знаний в генеративный контур, а также с исследованием адаптивных сценариев генерации тестовых материалов с учётом уровня подготовки, профиля ошибок и образовательной траектории конкретного обучающегося.

Список литературы

1. Ананин Д. П., Комаров Р. В., Реморенко И. М. «Когда честно–хорошо, для имитации–плохо»: стратегии использования генеративного искусственного интеллекта в российском вузе // Высшее образование в России. 2025. Т. 34. № 2. С. 31–50.
2. Белый А. В., Митрофанова О. А., Дубинина Н. А. Автоматическая генерация лексико-грамматических заданий по русскому языку как иностранному с помощью предсказывающих языковых моделей // Мир русского слова. 2023. № 2. С. 108–118.
3. Васильев А. А., Нестеров А. С. Применение алгоритмов формирования вопросов по тексту для автоматической генерации тестов // Вестник кибернетики. 2023. Т. 22. № 3. С. 17–22.
4. Дагаев А. Е. Искусственный интеллект в задаче генерации дистракторов для тестовых заданий // Моделирование, оптимизация и информационные технологии. 2025. Т. 13. № 2. URL: https://moitvivt.ru/ru/journal/article?id=1915. DOI: 10.26102/2310-6018/2025.49.2.028.
5. Двинина С. Ю., Цвентух Т. С. Возможности gigachat при обучении русскому языку как иностранному: нейросеть, создание учебных текстов и развитие видов речевой деятельности на уровне b1 // Вестник Челябинского государственного университета. 2025. № 6 (500). С. 150–158.
6. Демухаметов П. Н., Зарембо Я. А., Аврискин М. В. Исследование методов генерации тестовых вопросов по лекционным материалам ИТ-дисциплин // Математическое и информационное моделирование: материалы Всероссийской конференции молодых ученых. Вып. 23. Тюмень : ТюмГУ-Press, 2025.
7. Клобукова Л. П., Майоров Н. Д., Кочеткова Ю. А. Использование технологий искусственного интеллекта при разработке систем упражнений и заданий по русскому языку для иностранных студентов-социологов и учащихся

подготовительных факультетов российских вузов // Педагогика. Вопросы теории и практики. 2025. Т. 10. № 6. С. 735-742.

8. Кручинин В. В. Генераторы в компьютерных учебных программах : монография. Томск : ТМЦДО, 2003. 200 с. ISBN 5-7511-1763-8.
9. Кручинин В. В., Кузовкин В. В. Обзор существующих методов автоматической генерации задач с условиями на естественном языке // Компьютерные инструменты в образовании. 2022. № 1. С. 85-96.
10. Кручинин В. В., Морозова Ю. В. Модели и алгоритмы генерации задач в компьютерном тестировании // Известия Томского политехнического университета. Инжиниринг георесурсов. 2004. Т. 307. № 5. С. 127-131.
11. Курганова Н. А., Лапчик Е. С. Приемы разработки учебных заданий педагогами высшей школы с помощью нейросети // Проблемы современного педагогического образования. 2024. № 85-3. С. 187-190.
12. Куртасов А. М., Швецов А. Н. Метод автоматизированной генерации заданий для тестов контроля знаний из текстов учебных пособий // Современные информационные технологии и ИТ-образование. 2013. № 9. С. 218-228.
13. Куртасов А. М., Швецов А. Н. Метод автоматизированной генерации контрольно-тестовых заданий из текста учебных материалов // Вестник Череповецкого государственного университета. 2014. № 7 (60). С. 7-11.
14. Маслова М. А. Автоматизированный подход к отбору предложений для генерации тестовых заданий // Computational nanotechnology. 2024. Т. 11. № 2. С. 29-34.
15. Маслова М. А. Обзор существующих методов автоматической генерации тестовых заданий на естественном языке // Computational nanotechnology. 2023. Т. 10. № 4. С. 46-55.
16. Патаракин Е. Д. и др. Экспериментальное использование учебных материалов, разработанных с применением искусственного интеллекта, в естественнонаучном образовании2 // Вестник МГПУ. 2024. С. 79.
17. Сошников Д. В., Буров В. В., Патаракин Е. Д. Генерация учебных задач при помощи генеративных моделей // Информатизация образования и методика

электронного обучения: цифровые технологии в образовании. 2023. С. 1350-1354.

18. Швецов А. Н. и др. Архитектура интеллектуального агентно-ориентированного учебного комплекса для подготовки специалистов технического профиля // Открытое образование. 2018. Т. 22. № 3. С. 14-24.
19. Achiam J. et al. Gpt-4 technical report // arXiv preprint arXiv:2303.08774. 2023.
20. Alhazmi E. et al. Distractor generation in multiple-choice tasks: A survey of methods, datasets, and evaluation // Proceedings of the 2024 conference on empirical methods in natural language processing. 2024. С. 14437-14458.
21. Ali M. et al. Judging Quality Across Languages: A Multilingual Approach to Pretraining Data Filtering with Language Models // Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025. С. 8870-8909.
22. Anderson L. W., Krathwohl D. R. A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition. Addison Wesley Longman, Inc., 2001.
23. Awalurahman H. W., Budi I. Automatic distractor generation in multiple-choice questions: A systematic literature review // PeerJ Computer Science. 2024. Т. 10. С. e2441.
24. Banerjee S., Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments // Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. 2005. С. 65-72.
25. Biancini G., Ferrato A., Limongelli C. Multiple-choice question generation using large language models: Methodology and educator insights // Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization. 2024. С. 584-590.
26. Bitew S. K. et al. Distractor generation for multiple-choice questions with predictive prompting and large language models // Joint European Conference on

Machine Learning and Knowledge Discovery in Databases. Cham : Springer Nature Switzerland, 2023. C. 48-63.

27. Braam M., Van Der Velde M., Van Rijn H. Generating Competitive Distractors from Student Error Data // Proceedings of the Twelfth ACM Conference on Learning@ Scale. 2025. C. 305-309.
28. Brown T. et al. Language models are few-shot learners // Advances in neural information processing systems. 2020. T. 33. C. 1877-1901.
29. Byun G., Choi J. D. D-GEN: Automatic Distractor Generation and Evaluation for Reliable Assessment of Generative Models // Findings of the Association for Computational Linguistics: ACL 2025. 2025. C. 3316-3349.
30. Callison-Burch C., Osborne M., Koehn P. Re-evaluating the role of BLEU in machine translation research // 11th conference of the european chapter of the association for computational linguistics. 2006. C. 249-256.
31. Chan C. M. et al. Chateval: Towards better llm-based evaluators through multi-agent debate // arXiv preprint arXiv:2308.07201. 2023.
32. Chiang S. H., Wang S. C., Fan Y. C. Cdgp: Automatic cloze distractor generation based on pre-trained language model // Findings of the Association for Computational Linguistics: EMNLP 2022. 2022. C. 5835-5840.
33. Chirkin A. et al. RusConText Benchmark: A Russian Language Evaluation Benchmark for Understanding Context // Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop). 2025. C. 1158-1170.
34. Circi R., Hicks J., Sikali E. Automatic item generation: Foundations and machine learning-based approaches for assessments // Frontiers in Education. Frontiers Media SA, 2023. T. 8. C. 858273.
35. Cohen J. A coefficient of agreement for nominal scales // Educational and psychological measurement. 1960. T. 20. №. 1. C. 37-46.
36. Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding // Proceedings of the 2019 conference of the North American

chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019. C. 4171-4186.

37. Dey K. et al. Better to ask in english: Evaluation of large language models on english, low-resource and cross-lingual settings // arXiv preprint arXiv:2410.13153. 2024.
38. Dijkstra R. et al. Reading Comprehension Quiz Generation using Generative Pre-trained Transformers // iTextbooks@ AIED. 2022. C. 4-17.
39. Do X. L. et al. Modeling what-to-ask and how-to-ask for answer-unaware conversational question generation // Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023. C. 10785-10803.
40. Doughty J. et al. A comparative study of AI-generated (GPT-4) and human-crafted MCQs in programming education // Proceedings of the 26th Australasian Computing Education Conference. 2024. C. 114-123.
41. Du X., Shao J., Cardie C. Learning to ask: Neural question generation for reading comprehension // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017. C. 1342-1352.
42. Eleragi A. M. S. et al. Evaluating the multiple-choice questions quality at the College of Medicine, University of Bisha, Saudi Arabia: a three-year experience // BMC Medical Education. 2025. Т. 25. № 1. С. 233.
43. Fabbri A. R. et al. Summeval: Re-evaluating summarization evaluation //Transactions of the Association for Computational Linguistics. – 2021. – Т. 9. – С. 391-409.
44. Falcão F. et al. A suggestive approach for assessing item quality, usability and validity of Automatic Item Generation // Advances in Health Sciences Education. 2023. Т. 28. № 5. С. 1441-1465.
45. Fei Z. et al. CQG: A simple and effective controlled generation framework for multi-hop question generation // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022. C. 6896-6906.

46. Feng W. et al. Exploring automated distractor generation for math multiple-choice questions via large language models // Findings of the Association for Computational Linguistics: NAACL 2024. 2024. С. 3067-3082.
47. Fenogenova A. et al. MERA: A comprehensive LLM evaluation in Russian // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024. С. 9920-9948.
48. Fu W. et al. Qgeval: benchmarking multi-dimensional evaluation for question generation // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024. С. 11783-11803.
49. Gao Y. et al. Retrieval-augmented generation for large language models: A survey // arXiv preprint arXiv:2312.10997. 2023. Т. 2. № 1. – С. 32.
50. Ghanem B., Fyshe A. DISTO: Textual Distractors for Multiple Choice Reading Comprehension Questions Using Negative Sampling // International Educational Data Mining Society. 2024.
51. Gierl M. J. et al. A method for generating educational test items that are aligned to the common core state standards // Journal of Applied Testing Technology. 2015. С. 1-18.
52. Gierl M. J. et al. Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review // Review of educational research. 2017. Т. 87. № 6. С. 1082-1116.
53. Gierl M. J., Haladyna T. M. (ed.). Automatic item generation: Theory and practice. Routledge, 2012.
54. Gorgun G., Bulut O. Current evaluation methods are a bottleneck in automatic question generation // AI for education: Bridging innovation and responsibility at the 38th AAAI annual conference on AI. 2024.
55. Grévisse C., Pavlou M. A. S., Schneider J. G. Docimological quality analysis of LLM-generated multiple choice questions in computer science and medicine // SN Computer Science. 2024. Т. 5. № 5. С. 636.

56. Grover K. et al. Deep learning based question generation using t5 transformer // International Advanced Computing Conference. – Singapore : Springer Singapore, 2020. C. 243-255.
57. Guo S. et al. A survey on neural question generation: Methods, applications, and prospects // arXiv preprint arXiv:2402.18267. 2024.
58. Haladyna T. M., Rodriguez M. C. Developing and validating test items. Routledge, 2013.
59. Hassany M. et al. Generating Effective Distractors for Introductory Programming Challenges: LLMs vs Humans // Proceedings of the 15th International Learning Analytics and Knowledge Conference. 2025. C. 484-493.
60. Holtzman A. et al. The curious case of neural text degeneration // arXiv preprint arXiv:1904.09751. 2019.
61. Huang L. et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions // ACM Transactions on Information Systems. 2025. Т. 43. № 2. С. 1-55.
62. Indran I. R. et al. Twelve tips to leverage AI for efficient and effective medical question generation: A guide for educators using Chat GPT // Medical Teacher. 2024. Т. 46. № 8. С. 1021-1026.
63. Ji Z. et al. Survey of hallucination in natural language generation // ACM computing surveys. 2023. Т. 55. № 12. С. 1-38.
64. Katinskaia A. et al. Semi-automatically annotated learner corpus for Russian // Proceedings of the Thirteenth Language Resources and Evaluation Conference. 2022. С. 832-839.
65. Kıyak Y. S. et al. ChatGPT for generating multiple-choice questions: evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam // European journal of clinical pharmacology. 2024. Т. 80. № 5. С. 729-735.
66. Kıyak Y. S., Kononowicz A. A. Using a hybrid of AI and template-based method in automatic item generation to create multiple-choice questions in medical education: Hybrid AIG // JMIR Formative Research. 2025. Т. 9. С. e65726.

67. Kosakin D. et al. Russian learner corpus: Towards error-cause annotation for L2 Russian // Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). 2024. С. 14240-14258.
68. Kurdi G. et al. A systematic review of automatic question generation for educational purposes // International journal of artificial intelligence in education. 2020. Т. 30. № 1. С. 121-204.
69. Lai G. et al. Race: Large-scale reading comprehension dataset from examinations // Proceedings of the 2017 conference on empirical methods in natural language processing. 2017. С. 785-794.
70. Law A. K. K. et al. AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination // BMC Medical Education. 2025. Т. 25. № 1. С. 208.
71. Leo J. et al. Ontology-based generation of medical, multi-term MCQs // International Journal of Artificial Intelligence in Education. 2019. Т. 29. № 2. С. 145-188.
72. Lewis M. et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension // Proceedings of the 58th annual meeting of the association for computational linguistics. 2020. С. 7871-7880.
73. Lewis P. et al. Retrieval-augmented generation for knowledge-intensive nlp tasks // Advances in neural information processing systems. 2020. Т. 33. С. 9459-9474.
74. Liang C. et al. Distractor generation for multiple choice questions using learning to rank // Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications. 2018. С. 284-290.
75. Lin C. Y. Rouge: A package for automatic evaluation of summaries // Text summarization branches out. 2004. С. 74-81.
76. Liu B. et al. Learning to generate questions by learningwhat not to generate // The world wide web conference. 2019. С. 1106-1118.

77. Liu Z. Automatic Distractor and Feedback Generation in Online AI Education: A Design-Based Research Study // Proceedings of the 2025 ACM Conference on International Computing Education Research V. 2. 2025. C. 34-35.
78. Luo H. et al. Chain-of-exemplar: Enhancing distractor generation for multimodal educational question generation // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024. C. 7978-7993.
79. Maity S., Deroy A., Sarkar S. A novel multi-stage prompting approach for language agnostic mcq generation using gpt // European conference on information retrieval. – Cham : Springer Nature Switzerland, 2024. C. 268-277.
80. Martin Kowal J. et al. Harnessing Generative AI for Assessment Item Development: Comparing AI-Generated and Human-Authored Items // International Journal of Selection and Assessment. 2025. Т. 33. № 3. С. e70021.
81. McNichols H. et al. Automated distractor and feedback generation for math multiple-choice questions via in-context learning // arXiv preprint arXiv:2308.03234. 2023.
82. Mihaylov T. et al. Can a suit of armor conduct electricity? a new dataset for open book question answering // Proceedings of the 2018 conference on empirical methods in natural language processing. 2018. C. 2381-2391.
83. Min S. et al. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023. C. 12076-12100.
84. Mitkov R., Le An H., Karamanis N. A computer-aided environment for generating multiple-choice test items // Natural language engineering. 2006. Т. 12. № 2. С. 177-194.
85. Mohammadshahi A. et al. RQUGE: Reference-free metric for evaluating question generation by answering the question // Findings of the Association for Computational Linguistics: ACL 2023. 2023. C. 6845-6867.

86. Moore S. et al. Assessing the quality of multiple-choice questions using gpt-4 and rule-based methods // European conference on technology enhanced learning. Cham : Springer Nature Switzerland, 2023. С. 229-245.
87. Mulla N., Gharpure P. Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications // Progress in Artificial Intelligence. 2023. Т. 12. № 1. С. 1-32.
88. Nguyen B. et al. Reference-based metrics disprove themselves in question generation // Findings of the Association for Computational Linguistics: EMNLP 2024. 2024. С. 13651-13666.
89. Ouyang L. et al. Training language models to follow instructions with human feedback // Advances in neural information processing systems. 2022. Т. 35. С. 27730-27744.
90. Pan S. et al. Unifying large language models and knowledge graphs: A roadmap // IEEE Transactions on Knowledge and Data Engineering. 2024. Т. 36. № 7. С. 3580-3599.
91. Papineni K. et al. Bleu: a method for automatic evaluation of machine translation // Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002. С. 311-318.
92. Patarakin E. D., Burov V. V., Soshnikov D. V. Experimental generation of educational tasks in natural science disciplines using artificial intelligence // Вестник Московского Городского Педагогического Университета. Серия: Педагогика И Психология. 2023. Т. 17. № 4. С. 28-41.
93. Peng B. et al. Graph retrieval-augmented generation: A survey // ACM Transactions on Information Systems. 2025. Т. 44. № 2. С. 1-52.
94. Perkoff E. M. et al. Comparing neural question generation architectures for reading comprehension // Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023). 2023. С. 556-566.
95. Pugachev A. et al. Repa: Russian error types annotation for evaluating text generation and judgment capabilities // Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025). 2025. С. 136-150.

96. Raffel C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer // Journal of machine learning research. 2020. Т. 21. № 140. С. 1-67.
97. Rathod M., Tu T., Stasaski K. Educational multi-question generation for reading comprehension // Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022). 2022. С. 216-223.
98. Remnev N. et al. A language model for grammatical error correction in l2 Russian // arXiv preprint arXiv:2307.01609. 2023.
99. Ren S., Zhu K. Q. Knowledge-driven distractor generation for cloze-style multiple choice questions // Proceedings of the AAAI conference on artificial intelligence. 2021. Т. 35. № 5. С. 4339-4347.
100. Rodriguez-Torrealba R., Garcia-Lopez E., Garcia-Cabot A. End-to-end generation of multiple-choice questions using text-to-text transfer transformer models // Expert Systems with Applications. 2022. Т. 208. С. 118258.
101. Rodriguez-Torrealba R., Garcia-Lopez E., Garcia-Cabot A. Joint generation of distractors for multiple-choice questions: A text-to-text approach // Computers, Materials, & Continua. 2025. Т. 83. № 2. С. 1683.
102. Rozovskaya A., Roth D. Grammar error correction in morphologically rich languages: The case of Russian // Transactions of the Association for Computational Linguistics. 2019. Т. 7. С. 1-17.
103. Sayin A., Gierl M. Using OpenAI GPT to generate reading comprehension items // Educational Measurement: Issues and Practice. 2024. Т. 43. № 1. С. 5-18.
104. Scarlatos A. et al. Improving automated distractor generation for math multiple-choice questions with overgenerate-and-rank // Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024). 2024. С. 222-231.
105. See A., Liu P. J., Manning C. D. Get to the point: Summarization with pointer-generator networks // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017. С. 1073-1083.

106. Sellam T., Das D., Parikh A. BLEURT: Learning robust metrics for text generation // Proceedings of the 58th annual meeting of the association for computational linguistics. 2020. С. 7881-7892.

107. Shavrina T. et al. RussianSuperGLUE: A Russian language understanding evaluation benchmark // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020. С. 4717-4726.

108. Song Y., Du J., Zheng Q. Automatic item generation for educational assessments: a systematic literature review // Interactive Learning Environments. 2025. Т. 33. № 9. С. 5386-5405.

109. Sorokin A., Nasyrova R. LORuGEC: the Linguistically Oriented Rule-annotated corpus for Grammatical Error Correction of Russian // Proceedings of the International Conference «Dialogue». 2025. Т. 2025.

110. Sun X. et al. Answer-focused and position-aware neural question generation // Proceedings of the 2018 conference on empirical methods in natural language processing. 2018. С. 3930-3939.

111. Susanti Y. et al. Automatic distractor generation for multiple-choice English vocabulary questions // Research and practice in technology enhanced learning. 2018. Т. 13. № 1. С. 15.

112. Sweller J. Cognitive load during problem solving: Effects on learning // Cognitive science. 1988. Т. 12. № 2. С. 257-285.

113. Taktasheva E. et al. RuBLiMP: Russian benchmark of linguistic minimal pairs // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024. С. 9268-9299.

114. Taktasheva E. et al. TAPE: Assessing few-shot Russian language understanding // Findings of the Association for Computational Linguistics: EMNLP 2022. 2022. С. 2472-2497.

115. Taslimipoor S. et al. Distractor generation using generative and discriminative capabilities of transformer-based models // Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). 2024. С. 5052-5063.

116. Touvron H. et al. Llama 2: Open foundation and fine-tuned chat models // arXiv preprint arXiv:2307.09288. 2023.

117. Trinh V. A., Rozovskaya A. New dataset and strong baselines for the grammatical error correction of Russian // Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021. C. 4103-4111.

118. Ushio A., Alva-Manchego F., Camacho-Collados J. An empirical comparison of LM-based question and answer generation methods // Findings of the Association for Computational Linguistics: ACL 2023. 2023. C. 14262-14272.

119. Vachev K. et al. Leaf: Multiple-choice question generation // European Conference on information retrieval. – Cham : Springer International Publishing, 2022. C. 321-328.

120. Vaswani A. et al. Attention is all you need // Advances in neural information processing systems. 2017. T. 30.

121. Wang H. J. et al. Distractor generation based on Text2Text language models with pseudo Kullback-Leibler divergence regulation // Findings of the Association for Computational Linguistics: ACL 2023. 2023. C. 12477-12491.

122. Wang Z., Funakoshi K., Okumura M. Automatic answerability evaluation for question generation // arXiv preprint arXiv:2309.12546. 2023.

123. Wei J. et al. Chain-of-thought prompting elicits reasoning in large language models // Advances in neural information processing systems. 2022. T. 35. C. 24824-24837.

124. Willard B. T., Louf R. Efficient guided generation for large language models // arXiv preprint arXiv:2307.09702. 2023.

125. Wu Q. et al. Autogen: Enabling next-gen LLM applications via multi-agent conversations // First conference on language modeling. 2024.

126. Xie Z. et al. WorldTree v2: A corpus of science-domain structured explanations and inference patterns supporting multi-hop inference // Proceedings of the twelfth language resources and evaluation conference. 2020. C. 5456-5473.

127. Xiromeriti M., Newton P. M. Solving not answering. Validation of guidance for writing higher-order multiple-choice questions in medical science education // Medical Science Educator. 2024. Т. 34. № 6. С. 1469-1477.

128. Xuan W. et al. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation // Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025. С. 1513-1532.

129. Yang A. et al. Qwen3 technical report // arXiv preprint arXiv:2505.09388. 2025.

130. Yao Z. et al. Mcqg-srefine: Multiple choice question generation and evaluation with iterative self-critique, correction, and comparison feedback // Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2025. С. 10728-10777.

131. Yu H. C. et al. Enhancing distractor generation for multiple-choice questions with retrieval augmented pretraining and knowledge graph integration // Findings of the Association for Computational Linguistics: ACL 2024. 2024. С. 11019-11029

132. Zhang T. et al. Bertscore: Evaluating text generation with bert // arXiv preprint arXiv:1904.09675. 2019.

133. Zhang X. et al. Diagram-Driven Course Questions Generation // Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025. С. 5995-6010.

134. Zhang Y. et al. P-mmeval: A parallel multilingual multitask benchmark for consistent evaluation of llms // Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025. С. 4809-4836.

135. Zhao Y. et al. Paragraph-level neural question generation with maxout pointer and gated self-attention networks // Proceedings of the 2018 conference on empirical methods in natural language processing. 2018. С. 3901-3910.

136. Zheng L. et al. Judging llm-as-a-judge with mt-bench and chatbot arena // Advances in neural information processing systems. 2023. Т. 36. С. 46595-46623.

Приложение А. Свидетельства о регистрации программ для ЭВМ

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2025615835

Автоматизированная система таксономического структурирования контента для генерации контрольно-измерительных материалов

Правообладатель: ***Дагаев Александр Евгеньевич (RU)***

Автор(ы): ***Дагаев Александр Евгеньевич (RU)***

Заявка № **2025614040**
Дата поступления **03 марта 2025 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***10 марта 2025 г.***

Руководитель Федеральной службы
по интеллектуальной собственности

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ
Сертификат 0692e7c1a6300bf54f240f670bca2026
Владелец **Зубов Юрий Сергеевич**
Действителен с 10.07.2024 по 03.10.2025

Ю.С. Зубов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2024612459

Подсистема верификации контрольно-измерительных материалов

Правообладатель: ***Дагаев Александр Евгеньевич (RU)***

Автор(ы): ***Дагаев Александр Евгеньевич (RU)***

Заявка № **2024610781**
Дата поступления **19 января 2024 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***01 февраля 2024 г.***

Руководитель Федеральной службы
по интеллектуальной собственности

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ
Сертификат 429b6a0fe3853164baf96f83b73b4aa7
Владелец Зубов Юрий Сергеевич
Действителен с 10.05.2023 по 02.08.2024

Ю.С. Зубов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2024691569

Адаптивный модуль формирования дистракторов

Правообладатель: ***Дагаев Александр Евгеньевич (RU)***

Автор(ы): ***Дагаев Александр Евгеньевич (RU)***

Заявка № **2024689800**
Дата поступления **09 декабря 2024 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***23 декабря 2024 г.***

Руководитель Федеральной службы
по интеллектуальной собственности

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ
Сертификат 0692e7c1a6300bf54f240f670bca2026
Владелец Зубов Юрий Сергеевич
Действителен с 10.07.2024 по 03.10.2025

Ю.С. Зубов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2024691716

Интеллектуальная система анализа качества дистракторов

Правообладатель: ***Дагаев Александр Евгеньевич (RU)***

Автор(ы): ***Дагаев Александр Евгеньевич (RU)***

Заявка № **2024691445**
Дата поступления **20 декабря 2024 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***24 декабря 2024 г.***

Руководитель Федеральной службы
по интеллектуальной собственности

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ
Сертификат 0692e7c1a6300bf54f240f670bca2026
Владелец Зубов Юрий Сергеевич
Действителен с 10.07.2024 по 03.10.2025

Ю.С. Зубов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2025610015

Инструмент оценки контентной валидности теста

Правообладатель: ***Дагаев Александр Евгеньевич (RU)***

Автор(ы): ***Дагаев Александр Евгеньевич (RU)***

Заявка № **2024691726**
Дата поступления **23 декабря 2024 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***09 января 2025 г.***

Руководитель Федеральной службы
по интеллектуальной собственности

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ
Сертификат 0692e7c1a6300bf54f240f670bca2026
Владелец Зубов Юрий Сергеевич
Действителен с 10.07.2024 по 03.10.2025

Ю.С. Зубов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2025614018

Подсистема автоматической оценки правильности ответов

Правообладатель: ***Дагаев Александр Евгеньевич (RU)***

Автор(ы): ***Дагаев Александр Евгеньевич (RU)***

Заявка № **2025610901**
Дата поступления **27 января 2025 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***18 февраля 2025 г.***

Руководитель Федеральной службы
по интеллектуальной собственности

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ
Сертификат 0692e7c1a6300bf54f240f670bca2026
Владелец Зубов Юрий Сергеевич
Действителен с 10.07.2024 по 03.10.2025

Ю.С. Зубов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2025661041

Модуль анализа соответствия тестовых заданий образовательным стандартам

Правообладатель: ***Дагаев Александр Евгеньевич (RU)***

Автор(ы): ***Дагаев Александр Евгеньевич (RU)***

Заявка № **2025619874**
Дата поступления **28 апреля 2025 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***29 апреля 2025 г.***

Руководитель Федеральной службы
по интеллектуальной собственности

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ
Сертификат 0692e7c1a6300bf54f240f670bca2026
Владелец Зубов Юрий Сергеевич
Действителен с 10.07.2024 по 03.10.2025

Ю.С. Зубов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2025661449

Инструмент автоматической категоризации тестовых вопросов

Правообладатель: ***Дагаев Александр Евгеньевич (RU)***

Автор(ы): ***Дагаев Александр Евгеньевич (RU)***

Заявка № **2025619943**
Дата поступления **28 апреля 2025 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***06 мая 2025 г.***

Руководитель Федеральной службы
по интеллектуальной собственности

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ
Сертификат 0692e7c1a6300bf54f240f670bca2026
Владелец Зубов Юрий Сергеевич
Действителен с 10.07.2024 по 03.10.2025

Ю.С. Зубов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2025661372

Модуль автоматического извлечения структурированной информации из графических данных

Правообладатель: ***Дагаев Александр Евгеньевич (RU)***

Автор(ы): ***Дагаев Александр Евгеньевич (RU)***

Заявка № **2025660126**
Дата поступления **29 апреля 2025 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***05 мая 2025 г.***

Руководитель Федеральной службы
по интеллектуальной собственности

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ
Сертификат 0692e7c1a6300bf54f240f670bca2026
Владелец Зубов Юрий Сергеевич
Действителен с 10.07.2024 по 03.10.2025

Ю.С. Зубов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2025666271

Инструмент генерации вопросов причинно-следственных связей

Правообладатель: ***Дагаев Александр Евгеньевич (RU)***

Автор(ы): ***Дагаев Александр Евгеньевич (RU)***

Заявка № **2025664225**
Дата поступления **08 июня 2025 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***24 июня 2025 г.***

Руководитель Федеральной службы
по интеллектуальной собственности

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ
Сертификат 0692e7c1a6300bf54f240f670bca2026
Владелец Зубов Юрий Сергеевич
Действителен с 10.07.2024 по 03.10.2025

Ю.С. Зубов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2025666095

Модуль формирования тестовых заданий на основе табличных данных

Правообладатель: ***Дагаев Александр Евгеньевич (RU)***

Автор(ы): ***Дагаев Александр Евгеньевич (RU)***

Заявка № **2025664224**
Дата поступления **08 июня 2025 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***23 июня 2025 г.***

Руководитель Федеральной службы
по интеллектуальной собственности

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ
Сертификат 0692e7c1a6300bf54f240f670bca2026
Владелец Зубов Юрий Сергеевич
Действителен с 10.07.2024 по 03.10.2025

Ю.С. Зубов

Приложение Б. Акты внедрения и апробации

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«МОСКОВСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»
(МОСКОВСКИЙ ПОЛИТЕХ)

УТВЕРЖДАЮ
И.О. Проректора по учебной работе
Московского Политеха
_____ А.Б. Максимов
«__» _____ 2025 г

АКТ
_____№_____
_____

О внедрении в учебный процесс университета результатов научной работы
Дагаева Александра Евгеньевича на тему «Разработка алгоритмов извлечения и обработки информации для генерации многоформатных тестовых заданий», представленной на соискание ученой степени кандидата технических наук по специальности 2.3.8. «Информатика и информационные процессы».

Мы, нижеподписавшиеся директор Департамента по образовательной политике А.Б. Максимов, декан факультета Информационных технологий Д.Г. Демидов, заведующий кафедрой «Информатика и информационные технологии» Е.В. Булатников, составили настоящий акт в том, что результаты диссертационной работы по теме «Разработка алгоритмов извлечения и обработки информации для генерации многоформатных тестовых заданий» имеют как теоретическую, так и практическую значимость для учебного процесса и внедрены в учебный процесс Московского политехнического университета по направлению подготовки 09.04.02 «Информационные системы и технологии» (дисциплина «Искусственный интеллект в мобильных системах»). В учебном процессе нашли отражение следующие алгоритмы:

- абстрактивного извлечения информации с формализованным критерием дидактической значимости в виде параметрической интегральной метрики. Предложенный алгоритм обеспечивает управляемую параметрическую адаптацию процесса извлечения к специфике предметной области, повышая релевантность получаемого материала.
- генерации многоформатных тестовых заданий с модулем верификации на основе мультиагентной системы, обеспечивающим комплексную проверку автоматически сгенерированных заданий.
- генерации дистракторов, основанный на когнитивном моделировании. Алгоритм симулирует ошибочные инференционные цепочки, что позволяет целенаправленно генерировать правдоподобные неверные ответы, отражающие типичные заблуждения учащихся и обладающие более высокой диагностической ценностью.

Кафедра, внедрившая результаты: кафедра «Информатика и информационные технологии».

Дата и номер протокола заседания кафедры, на котором рассмотрены результаты внедрения 22.09.2025 №2.

Начало использования объекта внедрения: осенний семестр 2025-2026 учебного года.

Директор Департамента по образовательной политике _____ / А.Б. Максимов /

Декан факультета Информационных технологий _____ / Д.Г. Демидов /

Заведующий кафедрой «Информатика и Информационные технологии» _____ / Е.В. Булатников /

ОПТИК СМАРТ
300001, Тульская область, г.о. Город Тула, г Тула,
ул Замочная, дом 105Б, квартира 3, тел. 7 920 273 77 47
ОГРН 1237100012940
ИНН 7100044131
КПП 710001001
р/с: 40702810611740004764
ФИЛИАЛ «ЦЕНТРАЛЬНЫЙ» БАНКА ВТБ (ПАО)
БИК: 044525411
Корр. счет: 30101810145250000411

Исх.217 от 10.10.2025

УТВЕРЖДАЮ
генеральный директор ООО «ОПТИК СМАРТ"

Плыкина Е.В.

М.П.

«10» октября 2025г.

АКТ

о внедрении результатов диссертационного исследования

Дагаева Александра Евгеньевича

на тему «Разработка алгоритмов извлечения и обработки информации для генерации многоформатных тестовых заданий»

в практическую деятельность ООО «ОПТИК СМАРТ»

Комиссия в составе генерального директора Плыкиной Екатерины Викторовны, технического директора к.м.н. Лихачева Владимира Владимировича и главного инженера, д.т.н. Макарова Николая Николаевича подтверждает, что результаты диссертационного исследования, проведенного Дагаевым Александром Евгеньевичем, по теме: «Разработка алгоритмов извлечения и обработки информации для генерации многоформатных тестовых заданий» на соискание ученой степени кандидата технических наук по специальности 2.3.8. Информатика и информационные процессы, используются в практической деятельности ООО «ОПТИК СМАРТ», в частности, с помощью следующих зарегистрированных автором программ для ЭВМ:

1. «Автоматизированная система генерации контрольно-измерительных материалов» (Свидетельство о государственной регистрации программы для ЭВМ №2023686997);

2. «Подсистема верификации контрольно-измерительных материалов» (Свидетельство о государственной регистрации программы для ЭВМ №2024612459);
3. «Адаптивный модуль формирования дистракторов» (Свидетельство о государственной регистрации программы для ЭВМ №2024691569).

генерируются многоформатные текстовые задания с использованием разработанной автором уникальной мультиагентной системы, обеспечивающей комплексную проверку сгенерированных заданий, а также используется в работе предприятия зарегистрированный авторский комплекс по генерации контрольно-измерительных материалов, что повышает эффективность технической и программной поддержки продукции, производимой на нашем предприятии.

Генеральный директор Плыкина Е.В.

Технический директор Лихачев В.В.

Гл.инженер Макаров Н.Н.