Научная статья
УДК 004.09; DOI: 10.61260/2218-13X-2026-1-30-42

АЛГОРИТМ ПОДДЕРЖКИ ИНДИВИДУАЛЬНОГО ТЕСТИРОВАНИЯ ЗНАНИЙ НА ОСНОВЕ СИСТЕМ ГЕНЕРАТИВНОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

✉Коцюба Игорь Юрьевич;
Лайок Олег Владимирович;
Валдайцева Мария Викторовна.
Университет ИТМО, Санкт-Петербург, Россия
✉igor.kotciuba@gmail.com

Аннотация. Рассмотрен алгоритм автоматической генерации тематических тестов на примере тестов по английскому языку с использованием метода контрфактного анализа для повышения их качества на базе мобильного приложения.

В ходе детального анализа предметной области языкового тестирования были выстроены четкие требования к будущему сервису, классифицированы ключевые форматы контроля знаний с описанием типовых упражнений и уровней сложности, на которых они применяются, что помогло собрать целостную картину навыков, требующих автоматизированной проверки. Выделены сложные точки существующих тестов: двусмысленные формулировки, множественность корректных ответов, трудоёмкий подбор.

Разработан и апробирован комплексный подход к оценке эффективности промптов для генерации грамматических тестов на базе больших языковых моделей. В качестве ядра предложен контрфактный алгоритм, позволяющий выявлять латентные признаки, реально влияющие на выбор грамматических структур модели, точечно модифицировать промпт и оценивать изменения по трём взаимодополняющим метрикам. Применение алгоритма показало, что добавление явных указаний на самые значимые скрытые признаки повышает восприимчивость модели к ключевым факторам задания. Дальнейшая переоценка качества по разработанным метрикам и независимая экспертная проверка подтвердили статистически значимый прирост ($p < 0,01$) как в грамматическом соответствии, так и в соответствии структуре заданий: средняя оценка повысилась с 0,91 до 0,95. Таким образом, контрфактный анализ действительно является эффективным инструментом тонкой настройки промптов; предложенный улучшенный промпт обеспечивает более надёжную генерацию тестовых материалов, соответствующих образовательным стандартам, и закладывает основу для масштабирования алгоритма на другие типы заданий и языковые навыки.

Ключевые слова: качество образования, искусственный интеллект, Large Language Models, промт, контрфактный анализ, латентные признаки, грамматический тест, контрфактный алгоритм, восприимчивость модели, генерация тестов

Для цитирования: Коцюба И.Ю., Лайок О.В, Валдайцева М.В. Алгоритм поддержки индивидуального тестирования знаний на основе систем генеративного искусственного интеллекта // Научно-аналитический журнал «Вестник Санкт-Петербургского университета Государственной противопожарной службы МЧС России». 2026. № 1. С. 30–42. DOI: 10.61260/2218-13X-2026-1-30-42

Scientific article

ALGORITHM FOR SUPPORTING INDIVIDUAL KNOWLEDGE TESTING BASED ON A GENERATIVE ARTIFICIAL INTELLIGENCE SYSTEM

✉Kotsyuba Igor Yu.;
Laiok Oleg V.;
Valdaitceva Maria V.
ITMO University, Saint-Petersburg, Russia
✉igor.kotciuba@gmail.com

Abstract. The paper presents algorithm for the automatic generation of thematic tests using the example of English language tests using the counterfactual analysis method to improve their quality based on a mobile application.

A detailed analysis of the language domain led to the development of clear requirements for the future service. Key forms of assessment knowledge were classified, along with descriptions of typical exercises and the difficulty levels in which they are used, helping to create a comprehensive picture of the skills requiring step-by-step assessment. The challenges of existing tests are highlighted: ambiguous wording, multiple correct answers, and labor-intensive selection.

This paper develops and tests a comprehensive approach to assessing the effectiveness of prompts for generating grammar tests based on Large Language Models. A counterfactual algorithm is proposed as a core, which allows identifying latent features that actually influence the choice of grammatical structures of the model, selectively modifying the prompt, and evaluating changes using three complementary metrics. The application of the algorithm showed that adding explicit indications of the most significant hidden features increases the model's sensitivity to key factors of the task. Further re-evaluation of quality using the developed metrics and independent expert review confirmed a statistically significant increase ($p < 0.01$) in both grammatical compliance and compliance with the structure of tasks: the average score increased from 0,91 to 0,95. Thus, counterfactual analysis is indeed an effective tool for fine-tuning prompts; the proposed improved prompt ensures more reliable generation of test materials that meet educational standards and lays the foundation for scaling the algorithm to other types of tasks and language skills.

Keywords: quality of education, artificial intelligence, Large Language Models, prompt, counterfactual analysis, latent signs, grammar test, counterfactual algorithm, model sensitivity, test generation

For citation: Kotsyuba I.Yu., Laiok O.V., Valdaitceva M.V. Algorithm for supporting individual knowledge testing based on a generative artificial intelligence system // Scientific and analytical journal «Vestnik Saint-Petersburg university of State fire service of EMERCOM of Russia». 2026. № 1. P. 30–42. DOI: 10.61260/2218-13X-2026-1-30-42

Введение

В последние годы лидирующей технологией, направленной на повышение качества образования и его соответствие нуждам и требованиям каждого индивидуального студента, является искусственный интеллект. Различные модели машинного обучения с начала 20-х гг. активно применяются в создании индивидуального учебного плана и автоматической оценки работ учеников [1], однако активное развитие больших языковых моделей (Large Language Models, LLM) в 2022–2023 гг. дало новые возможности использовать искусственный интеллект в образовании. LLM по своему определению способна не только выступать в роли рекомендательной и оценочной модели, но и генерировать новый текстовый контент, что открывает множество сценариев использования в образовательном процессе.

Для оптимизации LLM и их обучения на узкоспециализированных задачах за последнее время появилось множество различных механик дообучения, однако все они требуют определенного объема вычислительных ресурсов, поэтому в первую очередь был сделан выбор сфокусироваться на prompt-engineering для достижения оптимального качества ответов модели на поставленной задаче. Однако оценка выбора промптов для модели может быть неочевидной задачей, так как генеративные нейронные сети представляют из себя черный ящик с малой интерпретируемостью результатов. В данной работе рассматривается метод контрфактного анализа как один из способов решения данной проблемы. Этот шаг необходим для создания интеллектуального решения по прохождению тематических тестов по английскому языку, отвечающего требованиям к создаваемым образовательным материалам.

Семейство LLM моделей начало активно развиваться с конца 2022 г. с выходом в общий доступ модели GPT-3.5, также известной как ChatGPT. Данный тип моделей основан на архитектуре трансформеров [2]. Модель «генеративный предобученный трансформер» (Generative Pre-trained Transformer, GPT) состоит из множества слоев трансформера. Каждый слой трансформера имеет две основные составляющие: механизм самовнимания (self-attention) и полносвязный нейронный слой. Входные данные проходят через несколько слоев трансформера, постепенно преобразуясь и уточняясь на каждом шаге. Механизм самовнимания позволяет модели обрабатывать и учитывать зависимости между различными элементами последовательности. Он основан на идее взаимодействия между отдельными элементами последовательности, где каждый элемент может «обратить внимание» на другие элементы для получения контекста. Это позволяет модели понимать долгосрочные зависимости и связи между словами в тексте. После применения механизма самовнимания данные проходят через полносвязные нейронные слои. Эти слои выполняют операции линейного преобразования и применения активационной функции, что помогает модели улавливать более сложные зависимости и выполнять предсказания на основе полученного контекста.

После прохождения через все слои трансформера данные проходят через процесс feeding-forward, который включает применение нескольких линейных преобразований и активационных функций для получения финального предсказания. Эти модели обучены на большом наборе текстовых данных, что позволяет с более высокой точностью выполнять генерацию текстов по запросу. К тому же, данные модели способны дообучаться на новых данных, которые может потребовать решение требуемой задачи, что делает их еще более адаптивными и расширяет сферу применения [3]. Благодаря большому количеству параметров (сотни миллиардов у лучших моделей в индустрии) и обширному набору данных, на котором обучаются LLM, они хорошо справляются с задачами обработки естественного языка в условиях нулевого и небольшого количества обучающих примеров [4]. Для понимания того, как работают большинство LLM, можно разобрать их на примере архитектуры GPT-3. В статье [5], опубликованной специалистами из OpenAI (USA, California), описываются ключевые моменты разработки и обучения модели, которые показывают, почему данная модель хорошо подходит для решаемой задачи. Подобные языковые модели развивают навыки распознавания шаблонов в данных, на которых они обучаются, что помогает им минимизировать потери при задаче моделирования языка. Позже эта способность помогает модели при постановке zero-shot задачи. Когда модели предоставляется несколько примеров и/или описание того, что ей нужно сделать, она сопоставляет шаблон примеров с тем, что она узнала в прошлом для аналогичных данных, и использует эти знания для выполнения задач. Это мощная возможность LLM, которая возрастает с увеличением количества параметров модели. Также авторами упоминается несколько методов обучения, описанных как Zero-Shot, One-Shot и Few-Shot Learning, − вместе с инструкцией в обучающем примере дается нулевое, одно и небольшое количество возможных примеров ответов соответственно. В случае небольшого количества примеров модели предоставляется описание задачи и столько примеров, сколько помещается в контекстное окно модели. В случае одного примера модели предоставляется ровно один пример, а в случае нулевого количества примеров примеры не предоставляются.

С увеличением мощности возможности модели с небольшим, одним и нулевым количеством примеров также улучшаются [5]. Данный паттерн обучения очень полезен при рассматриваемой задаче, так как способность модели генерировать данные более высокого качества только повышается при предоставлении нескольких примеров, что можно использовать при finetuning. GPT-3 обучалась на смешанном наборе из пяти различных корпусов, каждому из которых был назначен определенный вес. Наборы данных высокого качества выбирались чаще, и модель обучалась на них более одной эпохи. Пять использованных наборов данных: Common Crawl, WebText2, Books1, Books2 и Wikipedia. Наборы корпуса данных позволили модели выполнять не только задачу продолжения текста или генерацию текста по теме, но и полноценно отвечать на вопросы, выполнять инструкции и т.д. GPT-3 имеет 96 слоев, каждый слой содержит 96 модулей внимания. Размер эмбеддингов слов был увеличен до 12 888 для GPT-3 по сравнению с 1 600 для GPT-2. Размер контекстного окна был увеличен с 1 024 для GPT-2 до 2 048 токенов для GPT-3 [6].

За прошедший год появилось множество исследований [7−9], направленных на изучение вариантов применения LLM в образовательном процессе, из которых основными являются представление поддержки студенту по ходу образовательного процесса путем ответа на вопросы по предмету, формирование индивидуального учебного плана согласно запросам студента, предоставление образовательных материалов как для преподавателя, так и по запросу студента, оценка работ студента, создание интерактивных занятий, тренировка студентов к устным и письменным экзаменам. Кроме того, отдельно выделяется способность развивать различные навыки иностранного языка, так как LLM способна оценивать грамматическую и лексическую структуру речи, при этом все вышеперечисленные методы применения можно интегрировать и в изучение языков.

После анализа литературы можно выделить три основных направления, в которых используют LLM в образовании – вопросно-ответные системы, системы оценивания работ и генерация новых образовательных материалов. Данные варианты рассмотрены подробнее. Так, можно увидеть успешные случаи создания диалоговых систем для тренировки навыков устной речи студентов [10, 11]. В таких вариантах модель способна не только симулировать диалог с носителем языка, но отвечать на вопросы о речи, давая комментарии по ответам ученика. Компетентность таких моделей в сфере образования достигается путем составления специальных инструкций для модели под каждую задачу (prompt engineering), однако при разработке таких интеллектуальных диалоговых систем часто используется комбинация LLM с информационно-поисковой системой [12, 13], что существенно повышает компетентность модели в узкой предметной области, позволяет давать более экспертные ответы. Кроме того, для улучшения качества выдачи модели используется тонкая настройка (fine-tuning) на материалах, покрывающих больше диалоговых сценариев, а не стандартные обучающие датасеты, а также на различные учебники и методические пособия для уменьшения галлюцинаций модели.

Принимая во внимание существующий тренд на персонификацию учебного процесса [10], разработка и проведение тематических тестов может быть трудоемким процессом, а также не всегда способствовать полноценной и объективной оценке знаний студентов, так как персонализация теста под каждого конкретного ученика может быть довольно непростой задачей. К тому же, в общеобразовательных учреждениях существует большая проблема списывания на тестах, из-за чего страдает образовательный процесс студентов, а преподавателям приходится тратить еще больше времени на разработку уникальных тестов. Однако за последний год появилось несколько исследований, подтверждающих возможность использования GPT-моделей для решения этой задачи.

Для добавления интерактивных элементов в индивидуальную образовательную траекторию ученикам после глав курса предлагались задания на пройденную тему в формате множественного выбора, заполнения пропусков и сопоставления термина с определением. По оценкам экспертных преподавателей, сформированные тестовые вопросы соответствовали тематике курса и были корректными с лингвистической точки зрения.

Анализируя данные варианты тестовых заданий, можно сделать вывод, что, несмотря на явные отличия от тестов по другим предметам, LLM способны адаптироваться к такого вида тестам при правильном формулировании инструкций. Принимая во внимания данные успешные применения LLM в образовании стоит также не забывать о рисках, сопряженных с внедрением данного рода технологий. Образовательные стандарты имеют очень жесткие рамки, а грамотно выстроенный образовательный процесс использует большое число различных методик, практик и литературных источников, что делает их полное использование в интеллектуальной системе крайне затруднительным. Кроме того, LLM – это довольно новая технология, имеющая ряд нерешенных недостатков, таких как галлюцинации, требование больших вычислительных мощностей для обучения и размещения, отсутствие прозрачности в ее обучении и работа с персональными данными [14]. К тому же не каждый преподаватель имеет достаточный уровень компьютерной грамотности для внедрения данных технологий в свою образовательную практику. Однако LLM является очень мощным инструментом для задач, стоящих перед современными преподавателями и студентами, вследствие чего возникает потребность в разработке интеллектуальных систем, направленных на их решение. В процессе изучения вопроса ответственного использования данного семейства моделей в образовании исследователями был сформулирован ряд принципов, по которым предлагается разрабатывать интеллектуальные системы на основе LLM, среди которых выделяется прозрачность и открытость, уменьшение предвзятости и контроль человека над результатами работы системы [15].

Промпты – это текстовые инструкции, с помощью которых пользователи взаимодействуют с языковыми моделями. Промпт содержит указание модели, что именно нужно сгенерировать или какую задачу выполнить. Рассмотрен пример генерации тестов на знание грамматики по английскому языку. Для получения корректного теста нужно тонко настроить промпт для модели, что может быть непростой задачей по ряду причин, связанных как с особенностями работы генеративных моделей, так и с природой самого языка. Прежде всего, одной из главных трудностей является необходимость точного указания грамматических маркеров, которые должны быть протестированы. Английский язык, как и многие другие, обладает сложной системой времен и других грамматических категорий, которые могут выражаться различными способами – через форму глаголов, порядок слов, наречия времени и так далее. Если в промпте недостаточно точно обозначены эти грамматические структуры, модель может предложить варианты, которые не соответствуют задачам теста. Например, вместо нужной формы прошедшего времени модель может сгенерировать предложение в настоящем времени, что не будет полезным для проверки знаний учащихся. Проблема вариативности в генерации также играет важную роль. Даже при одинаковом запросе модель может сгенерировать разные предложения. Это свойство полезно, когда необходимо разнообразие в тестах, однако оно затрудняет стандартизацию примеров для проверки конкретных грамматических структур.

Другой трудностью является то, что модели могут ограниченно интерпретировать контекст. Если в промпте не указаны четкие ограничения на тип контекста, временные рамки или другие ключевые детали, модель может создавать предложения, не соответствующие нужному уровню абстракции или сложности. Так же одним из ключевых аспектов является необходимость учитывать уровень знаний учащихся. Для тестов, предназначенных для разных уровней, важно генерировать предложения, которые соответствуют не только грамматическим требованиям, но и лексической сложности и синтаксической структуре, характерным для этих уровней.

Для эффективной работы модели исследователи сформулировали множество различных правил и методов, с помощью которых нужно составлять инструкции [16−18]. Еще одним фактором, осложняющим генерацию, является наличие различных признаков в предложениях. Такие признаки, далее именуемые латентными, – это скрытые элементы, которые влияют на структуру и смысл предложения, но не всегда явно выражены. Эти признаки могут включать уровень сложности предложения, варианты применения требуемых грамматических структур или предполагаемый контекст. Модель может

интерпретировать задачу шире, чем предполагалось, и включить в предложение неявные грамматические или лексические элементы, что усложняет контроль за процессом генерации. Внедрение объяснимости в языковую модель служит для улучшения прозрачности ее работы и повышает доверие среди экспертов в области. Ввиду особенностей архитектуры LLM сложно восстановить процесс принятия решений. Объяснимость позволяет заглянуть внутрь черного ящика модели, выделяя значимые для вывода признаки и логику обработки информации. Это особенно важно для выявления ошибок и предвзятости, что способствует корректировке модели и предотвращению негативных последствий от неправильных интерпретаций. В контексте создания учебных материалов объяснимость позволяет убедиться, что модель корректно распознает и обрабатывает сложные грамматические структуры, гарантируя, что генерируемые предложения соответствуют требованиям и целям задания.

Говоря об объяснимости в языковых моделях, ученые исследуют несколько подходов в изучении причин генерации моделей или их предсказания в задачах классификации для выявления основных факторов, влияющих на результаты работы алгоритмов. В частности, рассмотрен доминирующий в научном сообществе на сегодняшний день метод контрфактной генерации (counterfactual generation) [19−21]. Это инструмент, используемый для повышения объяснимости языковых моделей, который помогает исследовать, как небольшие изменения во входных данных могут повлиять на результат, позволяя лучше понять, какие факторы наиболее важны для модели при принятии решения. Кроме того, контрфактная генерация не только предоставляет объяснение тому, как работает модель, но и помогает корректировать ее поведение, изменяя промпты или данные для обучения с целью улучшения качества и точности модели.

Таким образом, хотя LLM и начинают применяться в сфере образования, для предмета изучения языков не существует готового решения, которое представляло бы собой объединение вопросно-ответной системы и генерации разнообразных тестовых заданий с их проверкой для помощи студенту в тренировке различных языковых навыков, а учителю – в построении комфортного учебного процесса.

Методы исследования

Для промпта, с помощью которого будет генерироваться тест, использовался ансамбль из нескольких подходов − Instruction-based, Role prompting и Contextual prompting.

В ходе предыдущих экспериментов над генерацией тестовых материалов по английскому языку была замечена особенность модели понижать качество и вариативность при добавлении слишком большого объема контекста или числа примеров. Это связанно с тем, что у LLM снижается креативность, и генерируемые примеры становятся идентичными в промпте, что в неминуемом исходе ведет к повторяемости.

Для контрфактного анализа был использован следующий алгоритм (рис).

На первом шаге был сгенерирован тест согласно базовому промпту. Затем был подключен LLM-эксперт для анализа сгенерированных предложений и выявления латентных признаков, которые повлияли на выбор грамматических структур, указывая при этом, на какие аспекты задания обращать внимание (уровень, грамматическая конструкция, дополнительный контекст и т.д.).

Для каждого из выявленных латентных признаков выполняютя следующие действия:

1) сначала LLM-эксперт ищет конкретные токены (слова или фразы), которые влияют на определённый латентный признак;

2) далее LLM-эксперт меняет промпт так, чтобы выбранный латентный признак в новом предложении изменился;

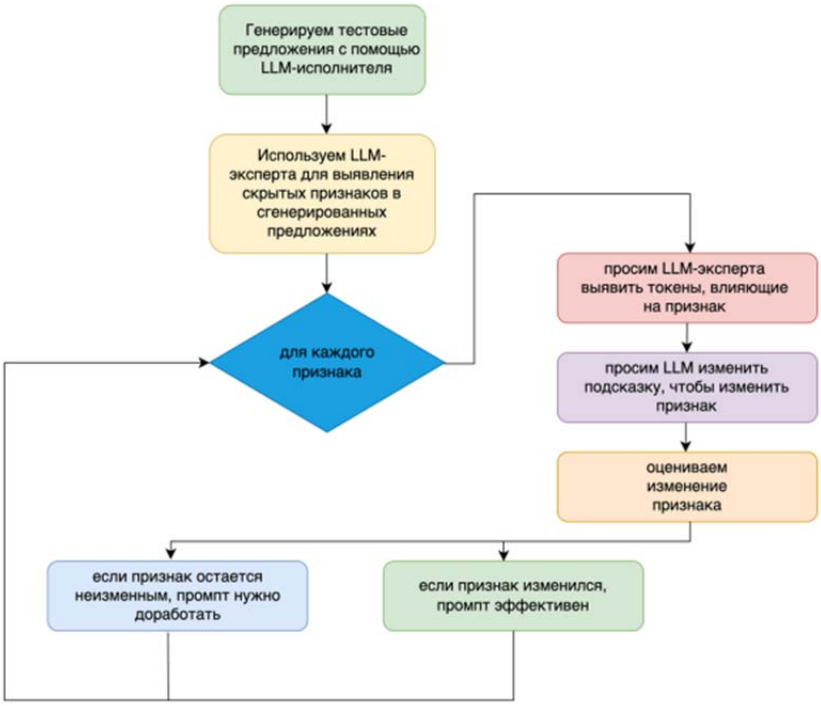

Рис. **Алгоритм контрфактного анализа**

3) оценка результата изменений следующим образом:

– если признак не изменился, это значит, что модель не распознала нужный латентный признак, и промпт неэффективно его захватывает. В таком случае нужно доработать инструкцию согласно данному признаку;

– если признак изменился, это указывает на то, что промпт успешно учитывает этот латентный признак, и его можно считать эффективным.

Перед тем как перейти к разбору примера, на котором тестировался данный алгоритм оценки эффективности промпта, рассматриваются выбранные метрики качества для промпта. Для оценки изменения качества промпта была выработана следующая система оценок, основанная на присутствии латентных признаков в тексте:

1) процент изменений латентных признаков (Latent Feature Change Rate) – оценивает процент случаев, когда латентные признаки изменились в соответствии с ожиданиями при модификации предложений. Для каждого изменённого предложения сравниваются исходные и новые значения латентных признаков;

2) изменение по количеству токенов, связанных с латентным признаком (Token Change for Latent Feature) − данный параметр показывает, насколько точно изменение предложения затронуло только те токены, которые связаны с изменением конкретного латентного признака. После изменения предложения LLM-эксперт указывает токены, которые влияют на целевой латентный признак (например, временные указатели или глаголы, связанные с грамматическим временем). Затем подсчитывается, сколько этих токенов было изменено между двумя версиями предложений, и берется среднее относительно всех токенов в предложениях;

3) чувствительность к изменениям латентных признаков (Sensitivity to Latent Feature Changes) – оценивает, насколько изменения латентных признаков влияют на финальное предложение и соответствуют ожидаемым результатам. После изменения одного конкретного латентного признака оценивается, насколько предложение изменилось с точки зрения остальных признаков. Идеально, если модификация затрагивает только один признак, не влияя на другие.

Далее рассматривается применение объяснимости для вышеуказанного базового промпта и его настройку согласно оценкам контрфактного анализа. Экспертной моделью были выделены следующие латентные признаки в сгенерированном тексте:

1) Temporal Context (Временной контекст) – слова-маркеры, которые указывают на время;

2) Aspect of Duration or Completion (Указание длительности или завершения) – слова-маркеры, указывающие на то, какое действие продолжительное, а какое завершенное;

3) Simultaneity vs. Sequentiality (Одновременность или последовательность) – происходят ли действия одновременно или последовательно друг за другом;

4) Clause Type (Тип предложения) – тип части предложения (основное, придаточное и т.д.), в которой находится глагол;

5) Interruptions (Прерывания) – наличие прерывающих действий в предложении;

6) Action Type (Dynamic vs. Stative тип глагола) – показывает тип глагола, от которого зависит употребление в continuous;

7) Language Proficiency Level (Уровень владения языком) – определенные слова, характеризующие уровень сложности понимания предложения;

8) Task Format (Тип задания) – формат теста, в котором создано предложение.

После нахождения латентных признаков были вычислены метрики по каждому из них. Было определено, что у 1, 7 и 8 признака самые высокие показатели, а значит, они больше всего влияют на результат генерации. У 5 признака также имелся довольно большой вес, поэтому на основе этих четырех признаков нужно изменить промпт – «You are an English teacher, who is preparing a grammar test for his students. Generate 10 sentences, that can be in this task: put the verbs in brackets into the simple past or past continuous tense. Use the context of each sentence to select the correct tense, considering whether the action was completed or ongoing, simultaneous or sequential, or interrupted by another event. Each sentence is designed for B1-level proficiency, so focus on familiar vocabulary and straightforward sentence structures while carefully choosing the correct verb form».

Теперь нужно оценить, насколько измененный промпт улучшил генерацию. В данном случае нет четкого исходного набора данных, с которым можно сравнить генерируемый текст. Модель должна выдавать текст, похожий на исходный только грамматически и структурно (например, если задание звучит как «Put the verbs in brackets into the right form», то в полученном результате глаголы должны стоять в скобочках в начальной форме и никак иначе). Следовательно, в данном случае нужно было разработать подходы именно по оценке грамматического и структурного соответствия оригиналу. Следует заметить, что в данной работе реализованы именно грамматические тесты, преимущественно на времена, поэтому для других тестов, которые в будущем можно добавить в данный сервис, например, на знание лексики, потребуется разработка уже других методик тестирования.

Грамматическое соответствие двух предложений можно оценить по соответствию времен в них. Если в теме задания требуется указать конкретное время или выбрать из нескольких, то корректным сгенерированным предложением будет считаться то, время которого соответствует времени в теме. Для автоматического определения времени в предложении следует разработать алгоритм, который определяет части речи в предложении, выделяет глаголы и на основе правил времен определяет, какие глаголы соответствуют какому времени. Правила времен можно описать регулярным выражением как последовательность определенных форм глагола, наличие которой в предложении показывает его время. Соответственно, если времена в сгенерированном предложении полностью соответствуют временам в теме, предложение получает оценку 1; если в предложении встретились все времена из темы, но добавились еще другие, то оценкой будет являться отношение правильных времен к количеству времен в сгенерированном предложении; если количество времен в предложении меньше или равно количеству в теме и при этом хотя бы одно время указано неправильно, то данное предложение является неверным и получает оценку 0.

Результаты исследования и их обсуждение

Данный метод оценки был применен к результатам генерации с базовым промптом и с измененным промптом. Для повышения качества оценки по данной метрике были опрошены эксперты – учителя английского, которые проставили оценки от 0 до 10. Результаты были переведены в диапазон [0,1] и усреднены вместе с результатами автоматической проверки. Итоговые метрики можно видеть в табл. 1, 2.

Таблица 1

Результат оценки грамматического соответствия до контрфактного анализа

Тема	Оценка для количества сгенерированных предложений				Итого
	5 предл.	10 предл.	15 предл.	20 предл.	
Present Simple or Present Continuous	0,83	0,98	0,93	0,99	0,93
Past Simple or Past Continuous	0,83	0,94	0,99	0,95	0,93
Past Simple or Present Perfect	0,99	0,90	0,86	0,99	0,94
Present Perfect or Present Perfect Continuous	0,79	0,86	0,99	0,90	0,89
Past Perfect or Past Simple	0,84	0,90	0,86	0,85	0,86
Итог	0,86	0,92	0,93	0,94	0,91

Таблица 2

Результат оценки грамматического соответствия после контрфактного анализа

Тема	Оценка для количества сгенерированных предложений				Итого
	5 предл.	10 предл.	15 предл.	20 предл.	
Present Simple or Present Continuous	0,98	0,99	0,93	0,99	0,97
Past Simple or Past Continuous	0,88	0,98	0,99	0,99	0,96
Past Simple or Present Perfect	0,99	0,95	0,99	0,99	0,98
Present Perfect or Present Perfect Continuous	0,99	0,80	0,99	0,96	0,94
Past Perfect or Past Simple	0,80	0,90	0,86	0,99	0,89
Итог	0,93	0,92	0,95	0,98	0,95

Для оценки соответствия по структуре довольно проблематично разработать средства автоматизации, поэтому было принято решение провести человеческую оценку. Сгенерированные предложения были оценены по соответствию структуре заданий. Если предложение не соответствовало структуре, оно получало оценку 0, в противном случае – оценку 1.

Данный метод оценки был также применен к результатам с оригинальным промптом (табл. 3) и с измененным промптом (табл. 4).

Таблица 3

Оценка соответствия структуре до контрфактного анализа

Тип упражнения	Оценка для количества сгенерированных предложений				Итого
	5 предл.	10 предл.	15 предл.	20 предл.	
Put in the right form	0,80	1,00	0,93	1,00	0,93
Choose the right form	0,80	0,90	0,93	1,00	0,91
Fill in the blanks	0,60	0,90	0,86	0,90	0,82
Итог	0,73	0,93	0,91	0,97	0,89

Таблица 4

Оценка соответствия структуре после контрфактного анализа

Тип упражнения	Оценка для количества сгенерированных предложений				Итого
	5 предл.	10 предл.	15 предл.	20 предл.	
Put in the right form	0,95	1,00	1,00	0,97	0,98
Choose the right form	1,00	0,88	1,00	1,00	0,97
Fill in the blanks	0,80	0,90	0,93	1,00	0,91
Итог	0,92	0,93	0,98	0,99	0,95

Из таблиц явно видно, что метрики улучшились по обоим критериям с изменением промпта. Однако нужно провести статистические тесты, чтобы убедиться, что прирост метрик не случаен. Сформулированы следующие гипотезы:

– $H_0 - \mu_d$iff ≤ 0 – средний прирост не наблюдается или отсутствует;

– $H_1 - \mu d_i$ff > 0 – средний прирост присутствует.

Метрики парные (имеются значения до и после), поэтому используется парный тест. Для того чтобы можно было провести односторонний парный t-тест, проверяются нормальности разностей метрик до-после через тест Шапиро-Уилка. Сначала проверяются метрики грамматического соответствия. Для полученных разностей пар и уровня значимости 0,05 p-value = 0,010, а значит гипотеза о нормальности распределения отклоняется. Тем не менее, при выборке из 20 пар t-тест с большой вероятностью сохраняет уровень ошибки и высокую мощность [21], а значит, можно его использовать, но на всякий случай проверяется исходная гипотеза через непараметрический t-критерий Уилкоксона, не требующий нормальности распределения. По нему, с тем же уровнем значимости, p-value = 0,0105 что позволяет отвергнуть гипотезу H0. Проведя парный t-тест при 19 степенях свободы с тем же уровнем значимости, получен p-value = 0,0097, что значительно меньше уровня значимости. Таким образом, можно утверждать, что средний прирост метрики по грамматическому соответствию присутствует и является статистически значимым.

Далее проводятся те же шаги для метрик соответствия типу. С тем же уровнем значимости p-value по тесту Шапиро-Уилка равно 0,125, что показывает нормальность распределения, поэтому можно применить t-тест. При нем получается p-value = 0,0078, что также позволяет отвергнуть гипотезу H0. При критерии Уилкоксона p-value = 0,0098, что так же меньше уровня значимости и явно показывает наличие статистически значимого среднего прироста и по метрике соответствия типу теста. Таким образом, прирост метрик является подтверждённым, а значит метод контрфактного анализа действительно работает для улучшения результата генерации через дообогащение промпта.

Заключение

Проведённое исследование показало, что применение LLM в сфере изучения иностранных языков не только возможно, но и отвечает стандартам качества в данной области. При использовании свойства объяснимости, а конкретно метода контрфактного анализа для тонкой настройки промптов LLM, приводит к статистически значимому росту качества автоматически сгенерированных грамматических тестов (средний комплексный показатель повысился с 0,91 до 0,95). Предложенный набор метрик – Latent Feature Change Rate, Relative Token Change и Sensitivity – проявил себя как надёжный инструмент экспресс-диагностики без необходимости ручной разметки, что особенно важно при масштабировании методики на новые типы заданий.

Проделанная работа открывает ряд перспектив для дальнейшего развития. Предложенный подход можно распространить на задания по аудированию, чтению и лексике, дополнив систему модулями синтеза и распознавания речи для письменных

и устных упражнений. Кроме того, портирование на другие языки потребует адаптации латентных признаков, но заложенные принципы сохранятся. Перспективным видится и переход к локальным или облегчённым LLM-моделям, что позволит организовать офлайн-режим и обеспечить по-настоящему адаптивную траекторию обучения.

Список источников

1. Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education / T.K.F. Chiu [et al.] // Computers and Education: Artificial Intelligence. 2023. Vol. 4. P. 100118. DOI: 10.1016/j.caeai.2022.100070

2. Kalyan K.S., Rajasekharan A., Sangeetha S. AMMUS: A Survey of Transformer-based Pretrained Models in Natural Language Processing // arXiv preprint. 2021. DOI: 10.48550/arXiv.2108.05542

3. Training language models to follow instructions with human feedback / L. Ouyang [et al.] // arXiv preprint. 2022. DOI:10.48550/arXiv.2203.02155

4. Language Models are Few-Shot Learners / T.B. Brown [et al.] // arXiv preprint. 2020. DOI: 10.48550/arXiv:2005.14165

5. GPT-3 family: Diverse applications of a large language model / T.B. Brown [et al.] // arXiv preprint. 2021. DOI: 10.48550/arXiv:2105.14208

6. Text-davinci: A large language model for diverse and creative text generation / A. Radford [et al.] // arXiv preprint. 2022. DOI: 10.48550/arXiv:2201.12136

7. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education / E. Kasneci [et al.] // arXiv preprint. 2023. DOI: 10.48550/arXiv:2304.11208

8. Adapting Large Language Models for Education: Foundational Capabilities, Potentials, and Challenges / Q. Li [et al.] // arXiv preprint. 2023. DOI: 10.48550/arXiv:2401.08664

9. Practical and Ethical Challenges of Large Language Models in Education: A Systematic Scoping Review / L. Yan [et al.] // arXiv preprint. 2023. DOI: 10.48550/arXiv:2303.13379

10. Nitze A. Future-proofing Education: A Prototype for Simulating Oral Examinations Using Large Language Models // arXiv preprint. 2023. DOI: 10.48550/arXiv:2401.06160

11. Peng L., Nuchged B., Gao Y. Spoken Language Intelligence of Large Language Models for Language Learning // arXiv preprint. 2023. DOI: 10.48550/arXiv:2308.14536

12. Wang K., Ramos J., Lawrence R. ChatEd: A Chatbot Leveraging ChatGPT for an Enhanced Learning Experience in Higher Education // arXiv preprint. 2023. DOI: 10.48550/arXiv:2401.00052

13. Castleman B., Turkcan M.K. Examining the Influence of Varied Levels of Domain Knowledge Base Inclusion in GPT-based Intelligent Tutors // arXiv preprint. 2023. DOI: 10.48550/arXiv:2309.12367

14. Large Language Models in Education: Vision and Opportunities / W. Gan [et al.] // arXiv preprint. 2023. DOI: 10.48550/arXiv:2311.13160

15. Challenges and Opportunities of Generative AI for Higher Education as Explained by ChatGPT / R. Michel-Villarreal [et al.] // Education Sciences. 2023. Vol. 13. № 9. P. 856. DOI: 10.3390/educsci13090856

16. A systematic survey of prompt engineering in large language models: Techniques and applications / P. Sahoo [et al.] // arXiv preprint. 2024. DOI: 10.48550/arXiv:2402.07927

17. Luo H., Specia L. From understanding to utilization: A survey on explainability for large language models // arXiv preprint. 2024. DOI: 10.48550/arXiv:2309.01029

18. Analyzing Chain-of-Thought Prompting in Large Language Models via Gradient-based Feature Attributions / S. Wu [et al.] // arXiv preprint. 2023. DOI: 10.48550/arXiv:2309.01029

19. Larger language models do in-context learning differently / J. Wei [et al.] // arXiv preprint. 2024. DOI: 10.48550/arXiv:2405.19592

20. Madsen A., Chandar S., Reddy S. Can Large Language Models Explain Themselves? // arXiv preprint. 2024. DOI: 10.48550/arXiv:2401.07927

21. LLMs as Counterfactual Explanation Modules: Can ChatGPT Explain Black-box Text Classifiers? / A. Bhattacharjee [et al.] // arXiv preprint. 2023. DOI: 10.48550/arXiv:2309.13340

References

1. Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education / T.K.F. Chiu [et al.] // Computers and Education: Artificial Intelligence. 2023. Vol. 4. P. 100118. DOI: 10.1016/j.caeai.2022.100070

2. Kalyan K.S., Rajasekharan A., Sangeetha S. AMMUS: A Survey of Transformer-based Pretrained Models in Natural Language Processing // arXiv preprint. 2021. DOI: 10.48550/arXiv.2108.05542

3. Training language models to follow instructions with human feedback / L. Ouyang [et al.] // arXiv preprint. 2022. DOI:10.48550/arXiv.2203.02155

4. Language Models are Few-Shot Learners / T.B. Brown [et al.] // arXiv preprint. 2020. DOI: 10.48550/arXiv:2005.14165

5. GPT-3 family: Diverse applications of a large language model / T.B. Brown [et al.] // arXiv preprint. 2021. DOI: 10.48550/arXiv:2105.14208

6. Text-davinci: A large language model for diverse and creative text generation / A. Radford [et al.] // arXiv preprint. 2022. DOI: 10.48550/arXiv:2201.12136

7. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education / E. Kasneci [et al.] // arXiv preprint. 2023. DOI: 10.48550/arXiv:2304.11208

8. Adapting Large Language Models for Education: Foundational Capabilities, Potentials, and Challenges / Q. Li [et al.] // arXiv preprint. 2023. DOI: 10.48550/arXiv:2401.08664

9. Practical and Ethical Challenges of Large Language Models in Education: A Systematic Scoping Review / L. Yan [et al.] // arXiv preprint. 2023. DOI: 10.48550/arXiv:2303.13379

10. Nitze A. Future-proofing Education: A Prototype for Simulating Oral Examinations Using Large Language Models // arXiv preprint. 2023. DOI: 10.48550/arXiv:2401.06160

11. Peng L., Nuchged B., Gao Y. Spoken Language Intelligence of Large Language Models for Language Learning // arXiv preprint. 2023. DOI: 10.48550/arXiv:2308.14536

12. Wang K., Ramos J., Lawrence R. ChatEd: A Chatbot Leveraging ChatGPT for an Enhanced Learning Experience in Higher Education // arXiv preprint. 2023. DOI: 10.48550/arXiv:2401.00052

13. Castleman B., Turkcan M.K. Examining the Influence of Varied Levels of Domain Knowledge Base Inclusion in GPT-based Intelligent Tutors // arXiv preprint. 2023. DOI: 10.48550/arXiv:2309.12367

14. Large Language Models in Education: Vision and Opportunities / W. Gan [et al.] // arXiv preprint. 2023. DOI: 10.48550/arXiv:2311.13160

15. Challenges and Opportunities of Generative AI for Higher Education as Explained by ChatGPT / R. Michel-Villarreal [et al.] // Education Sciences. 2023. Vol. 13. № 9. P. 856. DOI: 10.3390/educsci13090856

16. A systematic survey of prompt engineering in large language models: Techniques and applications / P. Sahoo [et al.] // arXiv preprint. 2024. DOI: 10.48550/arXiv:2402.07927

17. Luo H., Specia L. From understanding to utilization: A survey on explainability for large language models // arXiv preprint. 2024. DOI: 10.48550/arXiv:2309.01029

18. Analyzing Chain-of-Thought Prompting in Large Language Models via Gradient-based Feature Attributions / S. Wu [et al.] // arXiv preprint. 2023. DOI: 10.48550/arXiv:2309.01029

19. Larger language models do in-context learning differently / J. Wei [et al.] // arXiv preprint. 2024. DOI: 10.48550/arXiv:2405.19592

20. Madsen A., Chandar S., Reddy S. Can Large Language Models Explain Themselves? // arXiv preprint. 2024. DOI: 10.48550/arXiv:2401.07927

21. LLMs as Counterfactual Explanation Modules: Can ChatGPT Explain Black-box Text Classifiers? / A. Bhattacharjee [et al.] // arXiv preprint. 2023. DOI: 10.48550/arXiv:2309.13340

Информация о статье:
Статья поступила в редакцию: 12.01.2026; одобрена после рецензирования: 25.03.2026;
принята к публикации: 27.03.2026

Information about the article:
The article was submitted to the editorial office: 12.01.2026; approved after review: 25.03.2026;
accepted for publication: 27.03.2026

Информация об авторах:
Коцюба Игорь Юрьевич, доцент факультета прикладной информатики университета ИТМО (197101, Санкт-Петербург, Кронверкский пр., д. 49), кандидат технических наук, e-mail: igor.kotciuba@gmail.com, https://orcid.org/0000-0002-1680-5597, SPIN-код: 5296-3099
Лайок Олег Владимирович, магистрант факультета прикладной информатики университета ИТМО (197101, Санкт-Петербург, Кронверкский пр., д. 49), e-mail: laolvl@mail.ru
Валдайцева Мария Викторовна, преподаватель факультета технологического менеджмента и инноваций университета ИТМО (197101, Санкт-Петербург, Кронверкский пр., д. 49), кандидат технических наук, e-mail: mvvaldaitceva@itmo.ru, SPIN-код: 8895-0962

Information about authors:
Kotsyuba Igor Yu., associate professor of the faculty of applied informatics of ITMO university (197101, Saint-Petersburg, Kronverksky ave., 49), candidate of technical sciences, e-mail: igor.kotciuba@gmail.com, https://orcid.org/0000-0002-1680-5597, SPIN: 5296-3099
Laiok Oleg M., master's student at the faculty of applied informatics of ITMO university (197101, Saint-Petersburg, Kronverksky ave., 49), e-mail: laolvl@mail.ru
Valdaitceva Maria V., lecturer of the faculty of technological management and innovation of ITMO university (197101, Saint-Petersburg, Kronverksky ave., 49), candidate of technical sciences, e-mail: mvvaldaitceva@itmo.ru, SPIN: 8895-0962