УДК 004.89
DOI: 10.26102/2310-6018/2025.48.1.021

Метод генерации вопросов закрытого типа с использованием LLM

А.Е. Дагаев✉

Московский политехнический университет, Москва, Российская Федерация

Резюме. В исследовании представлен метод генерации вопросов закрытого типа, использующий большие языковые модели (LLM) для повышения качества и релевантности создаваемых вопросов. Предложенная структура объединяет этапы генерации, верификации и корректировки, что позволяет не исключать некачественные вопросы, а улучшать их с использованием обратной связи. Метод был протестирован на трех популярных наборах данных: SQuAD, Natural Questions и RACE. Ключевые метрики оценки ROUGE, BLEU и METEOR стабильно показывали улучшения производительности на всех протестированных моделях. В исследовании использовались четыре варианта LLM: O1, O1-mini, GPT-4o и GPT-4o-mini, при этом O1 достигла наивысших результатов по всем наборам данных и метрикам. Экспертная оценка показала увеличение точности до 14,4 % по сравнению с генерацией без верификации и корректировки. Полученные результаты подчеркивают эффективность метода в обеспечении большей ясности, фактической корректности и контекстуальной релевантности в сгенерированных вопросах. Сочетание автоматизированной верификации и корректировки дополнительно улучшает результаты, демонстрируя потенциал LLM в совершенствовании задач генерации текста. Результаты работы будут полезны исследователям в области обработки естественного языка, образовательных технологий, а также специалистам, работающим над адаптивными системами обучения и программным обеспечением корпоративного обучения.

Ключевые слова: генерация вопросов, большие языковые модели, искусственный интеллект, обработка естественного языка, O1, O1-mini, GPT-4o, GPT-4o-mini.

Для цитирования: Дагаев А.Е. Метод генерации вопросов закрытого типа с использованием LLM. *Моделирование, оптимизация и информационные технологии.* 2025;13(1). URL: https://moitvivt.ru/ru/journal/pdf?id=1799 DOI: 10.26102/2310-6018/2025.48.1.021

A method for generating closed-type questions using LLMs

A.E. Dagaev✉

Moscow Polytechnic University, Moscow, the Russian Federation

Abstract. This study presents a method for closed-ended question generation leveraging large language models (LLM) to improve the quality and relevance of generated questions. The proposed framework combines the stages of generation, verification, and refinement, which allows for the improvement of low-quality questions through feedback rather than simply discarding them. The method was tested on three widely recognized datasets: SQuAD, Natural Questions, and RACE. Key evaluation metrics, including ROUGE, BLEU, and METEOR, consistently showed performance gains across all tested models. Four LLM configurations were used: O1, O1-mini, GPT-4o, and GPT-4o-mini, with O1 achieving the highest results across all datasets and metrics. Expert evaluation revealed an accuracy improvement of up to 14.4% compared to generation without verification and refinement. The results highlight the method's effectiveness in ensuring greater clarity, factual correctness, and contextual relevance in generated questions. The combination of automated verification and refinement further enhances outcomes, showcasing the potential of LLMs to refine text generation tasks. These findings will benefit researchers in natural language processing, educational technology, and professionals working on adaptive learning systems and corporate training software.

Keywords: question generation, large language models, artificial intelligence, natural language processing, O1, O1-mini, GPT-4o, GPT-4o-mini.

For citation: Dagaev A.E. A method for generating closed-type questions using LLMs. *Modeling, Optimization and Information Technology*. 2025;13(1). (In Russ.). URL: https://moitvivt.ru/ru/journal/pdf?id=1799 DOI: 10.26102/2310-6018/2025.48.1.021

Введение

Актуальность темы подтверждается растущим спросом на автоматизированные образовательные инструменты, персонализированные платформы обучения и эффективные системы оценки знаний. С учетом возрастающей интеграции искусственного интеллекта (ИИ) в сферу образования, существует необходимость в надежных методах генерации качественных вопросов, которые способствуют улучшению учебного процесса и одновременно сокращают время и затраты человеческих ресурсов на подготовку таких материалов.

Системы автоматической генерации вопросов (QG) значительно эволюционировали, перейдя от подходов, основанных на правилах, к передовым нейронным архитектурам [1, 2]. Хотя в генерации открытых вопросов достигнуты существенные успехи, генерация закрытых вопросов, включая форматы с предполагаемым ответом «да / нет», остается сравнительно менее изученной областью [3]. Благодаря простоте и легкости автоматической оценки такие типы вопросов необходимы для стандартизированных тестов, быстрых проверок знаний и крупных по объему опросов.

Существующие методы генерации закрытых вопросов в основном опираются на генерацию с последующей фильтрацией или ранжированием для обеспечения соответствия качеству вопросов [4, 5]. Однако такой подход не предполагает итеративную доработку для корректировки характеристик вопросов, таких как ясность и фактическая корректность. Системы без постгенерационной проверки могут создавать вопросы, которые не будут соответствовать стандартам качества [6]. Исследование [7] показывает, что без итеративной доработки могут сохраняться фактические ошибки, снижая надежность сгенерированного контента. В статье [8] предлагают механизм Know-Filter, который эффективно удаляет нерелевантные вопросы. Однако эта система не дорабатывает и не генерирует вопросы повторно, оставляя пробелы в цепочке обеспечения качества.

Оригинальность и новизна данного подхода обусловлены выбором и применением структуры «генерация – верификация – корректировка», при которой присутствует возможность корректировать исходные формулировки с использованием обратной связи и повышать качество генерируемого контента. В отличие от традиционных методов генерации вопросов, данная структура использует LLM не только для генерации вопросов, но и для оценки их качества с последующей переработкой.

Целью данного исследования является разработка эффективного метода генерации закрытых вопросов с использованием больших языковых моделей. Для достижения цели были решены следующие задачи:

1. Разработан и реализован структурированный трехэтапный метод генерации вопросов, основанный на LLM.
2. Оценена производительность с использованием современных LLM.
3. Проведены сравнительные эксперименты эффективности предложенного метода, демонстрирующие улучшение ключевых показателей на популярных наборах данных.

Материалы и методы

Исходя из специфики стоящих задач, для исследования были взяты наборы данных:

– SQuAD. Один из самых популярных наборов данных в задачах генерации вопросов, выступающий в качестве стандартного эталона для оценки производительности моделей, предназначенных для создания качественных вопросов. Набор данных включает более 100 000 вопросов, сгенерированных на основе материалов по широкому спектру тем, что делает его богатым источником для фактологических и контекстно обоснованных вопросов. Также SQuAD используется для задач генерации текста, требующих высокой точности и выделения наиболее важных смысловых фрагментов. Указанный набор показал эффективность в адаптивных системах генерации вопросов в работе [9], где он использовался для повышения сложности автоматически созданных вопросов с помощью методов ИИ.

– Natural Questions [10]. Является эталонным набором данных, разработанным Google AI. Набор включает более 300 000 вопросов, связанных с реальными поисковыми запросами Google и соответствующими документами. В отличие от SQuAD, NQ содержит как длинные, так и короткие ответы, что делает его подходящим для генерации вопросов. Natural Questions используется для оценки устойчивости моделей генерации вопросов на разнообразных данных, в том числе содержащих двусмысленные формулировки и вопросы с разным уровнем детализации.

– RACE [11]. Представляет собой набор данных, специально разработанный для генерации вопросов. Набор включает более 100 000 вопросов, взятых из экзаменов. RACE имеет широкое распространение в задачах оценки качества генерации вопросов благодаря включенным в него вопросам, требующим высокоуровневых рассуждений и умозаключений.

При выборе метрик для оценки качества генерации вопросов с использованием искусственного интеллекта важно учитывать несколько ключевых показателей, каждый из которых имеет свои особенности и сферы применения. В исследовании использовались следующие:

– BLEU. Одна из самых известных и широко используемых метрик, которая измеряет степень совпадения n-грамм между сгенерированным текстом и эталонным. Например, исследование [12] демонстрирует эффективность BLEU в контексте сопоставления текстов. Эта метрика подходит для сценариев, где важно лексическое содержание. В то же время, получаемая оценка может не соответствовать семантической точности.

– METEOR. Метрика устраняет некоторые недостатки BLEU, оценивает семантическое сходство, принимая во внимание синонимы, стемминг и порядок слов, что делает ее особенно полезной для задач, требующих более глубокого лингвистического понимания. В работе [13] подчеркивается способность METEOR выявлять тонкие лингвистические различия в узкоспециализированных контекстах.

– ROUGE. Используется для оценки текстов, в том числе в задаче генерации вопросов. Данная метрика измеряет пересечение n-грамм, последовательностей слов и наибольших общих подпоследовательностей между сгенерированным и эталонным текстами, делая акцент на полноте, чтобы учесть охват ключевых фрагментов текста. В задачах генерации вопросов ROUGE часто применяется для оценки качества вопросов, созданных ИИ путем их сравнения с вопросами, составленными человеком. Например, в работе [14] использовали ROUGE для оценки качества сгенерированных ИИ вопросов с вопросами, созданными человеком, и обнаружили, что ROUGE-L, который измеряет

наибольшую совпадающую последовательность, был особенно эффективен для оценки структурного сходства сгенерированных вопросов с эталонными.

Для получения более сбалансированного результата рекомендуется использовать комбинацию нескольких метрик. Так, исследование [15] подчеркнуло важность комплексного подхода и показало преимущества совместного применения ROUGE, BLEU и METEOR для оценки современных моделей.

В данном исследовании для подтверждения эффективности метода оценивание проводится на четырех вариантах LLM – gpt-4o-2024-11-20, gpt-4o-mini-2024-07-18, o1-2024-12-17 и o1-mini-2024-09-12.

Метод для генерации вопросов включает три взаимосвязанных этапа: генерация, верификация и корректировка. Каждый из них полностью выполняется с использованием LLM, что обеспечивает высокую производительность и интеграцию всех процессов в единое решение.

На первом этапе LLM используется для генерации вопросов на основе заданного контекста. Задача модели – сгенерировать вопрос. В общем виде расчет представлен в формуле 1.

$$q = \mathrm{LLM}_{\mathrm{Gen}}(x), \tag{1}$$

где q – итоговый сгенерированный вопрос; x – входной контекст (текстовый фрагмент); $\mathrm{LLM}_{\mathrm{Gen}}$ – языковая модель, которая генерирует вопрос q непосредственно на основе входного контекста x.

На этапе верификации каждый сгенерированный вопрос проходит проверку с использованием LLM, которая может быть представлена в виде:

$$R_{ver} = \mathrm{LLM}_{\mathrm{Gen}}\left(\mathrm{prompt}_{\mathrm{ver}}(x, q)\right), \tag{2}$$

где R_{ver} – результат верификации; $\mathrm{prompt}_{\mathrm{ver}}$ – инструкция для верификации.

Взаимодействие с большими языковыми моделями возможно через инструкции на естественном языке. Для верификации применяется запрос, отраженный в Таблице 1, в который подставляются сгенерированный вопрос и исходный текст.

Таблица 1 – Запрос для этапа верификации
Table 1 – Request for the verification stage

Необходимо оценить вопрос {ВОПРОС}, который был сгенерирован из текста {ИСХОДНЫЙ_ТЕКСТ} по пунктам: 1) Ясность. Насколько понятен смысл вопроса? Есть ли двусмысленность или нечеткость формулировок? 2) Грамматическая корректность. Соответствует ли вопрос грамматическим, орфографическим и пунктуационным нормам? 3) Релевантность. Соответствует ли вопрос контексту исходного текста? Формат ответа: По каждому пункту требуется поставить оценку: если критерий выполняется – «замечаний нет», если не выполняется – «есть замечания». При наличии замечаний необходимо указать их в формате: Замечания: {ОПИСАНИЕ}

На этапе корректировки вопросы, неудовлетворительно прошедшие верификацию, дорабатываются с учетом замечаний, полученных на предыдущем этапе. Смысл корректировки заключается в том, что исходный вопрос, на котором выявлены недостатки, не исключается, а корректируется с помощью дополнительного запроса. Данный этап представлен в формуле (3).

$$q_{\text{new}} = \text{LLM}_{\text{Gen}}\left(\text{prompt}_{\text{corr}}(x, q, R_{ver})\right), \quad (3)$$

где q_{new} – скорректированный вопрос; x – контекст; q – исходный вопрос, который необходимо исправить; $\text{prompt}_{\text{corr}}$ – инструкция для корректировки; R_{ver} – замечания.

При подаче инструкции на корректировку в модель LLM добавляются выявленные замечания. Данный запрос представлен в Таблице 2.

Таблица 2 – Запрос для этапа корректировки
Table 2 – Request for the adjustment stage

Необходимо исправить вопрос {ВОПРОС}, который был сгенерирован из текста {ИСХОДНЫЙ_ТЕКСТ}, исходя из замечаний {ЗАМЕЧАНИЯ}

Следует отметить, в данном случае корректировка представляет собой целенаправленный процесс доработки, а не случайную перегенерацию. В результате ожидается получение улучшенного нового вопроса.

Результаты

Для набора данных SQuAD результаты представлены в Таблице 3.

Таблица 3 – Показатели метрик для набора данных SQuAD
Table 3 – Metrics for the SQuAD Dataset

LLM Model	Configuration	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	METEOR
O1	Base	45,3	21,1	43,7	41,7	40,3
	Optimized	51,9	24,2	49,6	47,8	46,2
	Improvement	**6,6**	**3,2**	**5,9**	**6,1**	**5,9**
O1-mini	Base	43,2	20,1	41,6	39,6	38,7
	Optimized	49,0	22,9	46,7	44,9	43,4
	Improvement	**5,8**	**2,8**	**5,1**	**5,3**	**4,7**
GPT-4o	Base	42,3	19,6	40,2	38,6	37,3
	Optimized	47,6	22,0	44,8	43,3	41,6
	Improvement	**5,3**	**2,4**	**4,6**	**4,7**	**4,3**

Для набора данных Natural Questions результаты представлены в Таблице 4.

Таблица 4 – Показатели метрик для набора данных Natural Questions
Table 4 – Metrics for the Natural Questions Dataset

LLM Model	Configuration	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	METEOR
O1	Base	44,6	20,5	42,5	40,5	39,9
	Optimized	50,6	23,6	48,4	46,5	45,4
	Improvement	**6,0**	**3,1**	**5,9**	**6,0**	**5,5**
O1-mini	Base	42,1	19,8	40,5	38,2	37,2
	Optimized	47,8	22,4	45,9	44,1	42,5
	Improvement	**5,7**	**2,6**	**5,4**	**5,9**	**5,3**
GPT-4o	Base	41,8	19,2	39,5	37,5	36,0
	Optimized	47,0	21,8	44,7	43,6	41,5
	Improvement	**5,2**	**2,6**	**5,2**	**6,1**	**5,5**
GPT-4o-mini	Base	40,5	18,9	38,5	36,5	35,4
	Optimized	45,4	20,9	43,0	41,5	40,1
	Improvement	**4,9**	**2,1**	**4,5**	**5,0**	**4,7**

Для набора данных RACE результаты представлены в Таблице 5.

Таблица 5 – Показатели метрик для набора данных RACE
Table 5 – Metrics for the RACE Dataset

LLM Model	Configuration	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	METEOR
O1	Base	46,0	21,3	44,0	41,8	40,2
	Optimized	52,9	24,5	50,5	48,1	46,8
	Improvement	**6,9**	**3,3**	**6,5**	**6,3**	**6,6**
O1-mini	Base	43,8	20,1	42,0	39,5	38,5
	Optimized	50,1	23,2	47,7	45,6	44,4
	Improvement	**6,3**	**3,1**	**5,7**	**6,1**	**5,9**
GPT-4o	Base	42,5	19,8	40,8	38,2	37,4
	Optimized	48,5	22,4	46,1	43,9	42,8
	Improvement	**6,0**	**2,7**	**5,3**	**5,7**	**5,4**
GPT-4o-mini	Base	41,3	18,9	39,2	36,7	35,5
	Optimized	46,7	21,5	44,6	42,1	40,7
	Improvement	**5,4**	**2,6**	**5,4**	**5,4**	**5,2**

На Рисунке 1 изображена тепловая карта улучшений производительности по всем метрикам и наборам данных.

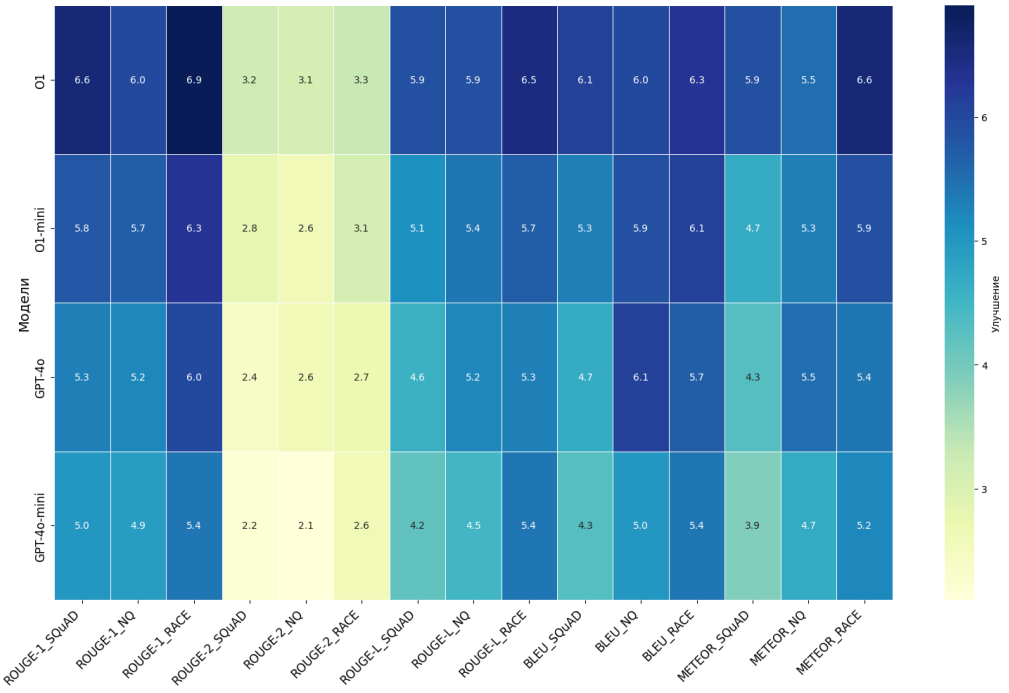

Рисунок 1 – Тепловая карта улучшений производительности по метрикам и наборам данных
Figure 1 – Heatmap of performance improvements across metrics and datasets

Обсуждение

Результаты экспериментов демонстрируют, что применение этапа корректировки повышает качество генерируемых вопросов. В отличие от тех методов, где некачественные вопросы исключаются, предложенный подход позволяет использовать обратную связь, полученную на этапе верификации, для доработки исходного вопроса.

Для набора данных SQuAD результаты показывают улучшения для всех моделей, причем O1 достигает наибольшего прироста, включая увеличение ROUGE-1 на 6,6 п. и

BLEU на 6,1 п. Модель O1-mini также продемонстрировала высокую производительность, улучшив ROUGE-1 на 5,8 п. В сравнении с ними, GPT-4o и GPT-4o-mini показали относительно меньшие, но все же заметные улучшения, особенно по BLEU и METEOR.

В наборе данных Natural Questions тенденция остается схожей. Модель O1 вновь лидирует с приростом ROUGE-1 и BLEU на 6 п., за ней следует O1-mini. Хотя GPT-4o и GPT-4o-mini показывают меньшие улучшения, но, тем не менее, демонстрируют стабильный рост, особенно по BLEU и METEOR. Однако общий прирост на этом наборе данных немного ниже, чем на SQuAD, что, вероятно, связано с большей сложностью содержащегося контента в Natural Questions.

Набор данных RACE показывает наиболее значительные улучшения для всех моделей, O1 достигает увеличения ROUGE-1 на 6,9 п. и METEOR на 6,6 п. O1-mini также демонстрирует положительные результаты, с существенными улучшениями по BLEU и ROUGE-1. GPT-4o и GPT-4o-mini показывают улучшения до 6,0 п. по различным метрикам.

В целом, O1 превосходит другие модели по всем наборам данных и метрикам, что свидетельствует о ее способности генерировать качественные вопросы. Верификация и корректировка стабильно улучшают производительность, при этом степень улучшений варьируется в зависимости от модели и набора данных.

Тепловая карта (Рисунок 1) наглядно иллюстрирует эти улучшения, где более темные оттенки указывают на больший прирост производительности. O1 достигает наиболее значительных улучшений, особенно на наборе данных RACE, тогда как GPT-4o-mini демонстрирует меньшие, но все же значимые приросты производительности.

Этап верификации был эффективен для выявления проблем, связанных с фактической точностью и контекстуальным соответствием, что важно для образовательных и оценочных задач.

Этап корректировки продемонстрировал свою важность, исправляя вопросы, отклоненные на этапе верификации, и позволил улучшить контекстуальную связность и синтаксическую ясность. Особенно заметно это было в наборе данных RACE, где корректировка повысила показатель ROUGE-L на 6,5 п., что подчеркивает его эффективность в сохранении структурного и семантического соответствия исходному материалу.

На Рисунке 2 изображена диаграмма оценивания экспертом сгенерированных вопросов.

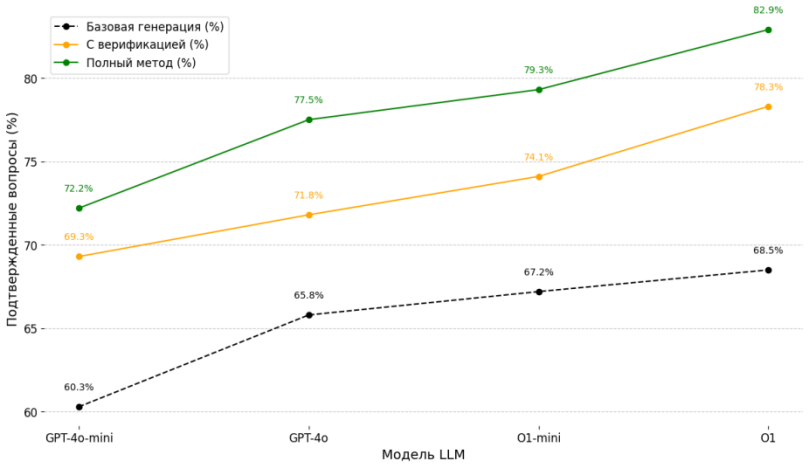

Рисунок 2 – Результаты подтверждений тестовых вопросов экспертом
Figure 2 – Results of test question validation by the expert

Процесс оценки включал три этапа: базовая генерация, с верификацией и с полным методом. На каждом этапе было выполнено 1000 генераций. Наибольший достигнутый показатель составил 14,4 %.

O1 стабильно показывала лучшие результаты на всех этапах, достигнув 82,9 % одобрения с применением полного метода. GPT-4o-mini при наиболее низкой базовой производительности также продемонстрировала улучшение, достигнув 72,2 % одобрения с полным методом.

Метод был протестирован на четырех LLM: O1, O1-mini, GPT-4o и GPT-4o-mini, при этом O1 превосходила остальные. Данная модель достигла наивысших показателей, особенно в ROUGE-1, BLEU и METEOR, благодаря улучшенной способности обрабатывать сложные контексты и генерировать семантически богатые тексты. GPT-4o показала наибольшую результативность от этапа корректировки (5,7 %), что привело как к росту показателей метрик, так и к общему количеству подтверждений экспертом. Аналогично, облегченная версия O1-mini продемонстрировала хорошую производительность, обеспечивая баланс между вычислительной эффективностью и качеством генерации текста.

Заключение

С научной точки зрения, новизна результатов заключается в разработке метода, основанного на использовании больших языковых моделей, которые улучшают качество генерируемых вопросов. Предложенный метод генерации вопросов закрытого типа с использованием LLM, объединяющий этапы генерации, верификации и корректировки, демонстрирует, что применение корректирующего запроса для доработки вопросов повышает их качество. Такой подход, отличающийся от методов фильтрования, позволяет сохранить потенциально ценные варианты и улучшить их лингвистическую и контекстуальную точность. Эффективность метода была подтверждена автоматической оценкой на наборах данных SQuAD, Natural Questions и RACE. Использованные ключевые метрики ROUGE, BLEU и METEOR показали рост производительности на всех протестированных моделях, при этом точность относительно генерации без применения данного метода повысилась до 14,4 %.

В исследовании показано, что предложенный метод позволяет обеспечить более высокую ясность, фактологическую точность и контекстуальную релевантность вопросов. Также можно отметить потенциал использования LLM для улучшения задач генерации текста. Сочетание автоматизированной верификации и корректировки обеспечивает более высокое качество результатов, делая метод полезным инструментом для развития ИИ-приложений в образовании, исследованиях и индустрии.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Huang J.-H., Zhu H., Shen Yi., et al. Image2Text2Image: A Novel Framework for Label-Free Evaluation of Image-to-Text Generation with Text-to-Image Diffusion Models. arXiv. URL: https://doi.org/10.48550/arXiv.2411.05706 [Accessed 3rd January 2025].
2. Chen Q., Wang Y., Wang F., et al. Decoding text from electroencephalography signals: A novel Hierarchical Gated Recurrent Unit with Masked Residual Attention Mechanism. *Engineering Applications of Artificial Intelligence.* 2025;139. https://doi.org/10.1016/j.engappai.2024.109615
3. Zakareya S., Alsaleem N., Alnaghmaish A., et al. Evaluating the Discrimination Index of AI-Generated vs. Human-Generated Multiple-Choice Questions: Action Research. In: *ICERI2024 Proceedings: 17th annual International Conference of Education,*

Research and Innovation, 11–13 November 2024, Seville, Spain. IATED; 2024. pp. 221–226. https://doi.org/10.21125/iceri.2024.0137

4. Shetty N., Li Yo. Detailed Image Captioning and Hashtag Generation. *Future Internet*. 2024;16(12). https://doi.org/10.3390/fi16120444
5. Kwiatkowski T., Palomaki J., Redfield O., et al. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*. 2019;7:453–466. https://doi.org/10.1162/tacl_a_00276
6. Lai G., Xie Q., Liu H., et al. RACE: Large-Scale ReAding Comprehension Dataset from Examinations. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 09–11 September 2017, Copenhagen, Denmark.* Association for Computational Linguistics; 2017. pp. 785–794. https://doi.org/10.18653/v1/D17-1082
7. Thorne W., Robinson A., Peng B., et al. Increasing the Difficulty of Automatically Generated Questions via Reinforcement Learning with Synthetic Preference. In: *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, 16 November 2024, Miami, USA.* Association for Computational Linguistics; 2024. pp. 450–462. https://doi.org/10.18653/v1/2024.nlp4dh-1.43
8. Ribeiro M.T., Singh S., Guestrin C. Semantically Equivalent Adversarial Rules for Debugging NLP Models. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Volume 1: Long Papers, 15–20 July 2018, Melbourne, Australia*. Association for Computational Linguistics; 2018. pp. 856–865. https://doi.org/10.18653/v1/P18-1079
9. Brown T., Mann B., Ryder N., et al. Language Models Are Few-Shot Learners. In: *Advances in Neural Information Processing Systems 33: 34th Conference on Neural Information Processing Systems 2020, NeurIPS 2020, 06–12 December 2020, Vancouver, Canada.* 2020. pp. 1877–1901.
10. Bian Yu., Huang J., Cai X., et al. On Attention Redundancy: A Comprehensive Study. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, 06–11 June 2021, Online.* Association for Computational Linguistics; 2021. pp. 930–945. https://doi.org/10.18653/v1/2021.naacl-main.72
11. Jiang N., De Marneffe M.-C. He Thinks He Knows Better than the Doctors: BERT for Event Factuality Fails on Pragmatics. *Transactions of the Association for Computational Linguistics*. 2021;9:1081–1097. http://doi.org/10.1162/tacl_a_00414
12. Lafkiar S., En Nahnahi N. An End-to-End Transformer-Based Model for Arabic Question Generation. *Multimedia Tools and Applications*. 2024. https://doi.org/10.1007/s11042-024-19958-3
13. Balepur N., Gu F., Ravichander A., et al. Reverse Question Answering: Can an LLM Write a Question so Hard (or Bad) that it Can't Answer? [Preprint]. arXiv. URL: https://doi.org/10.48550/arXiv.2410.15512 [Accessed 3rd January 2025].
14. Ye W., Zhang Q., Zhou X., et al. Correcting Factual Errors in LLMs via Inference Paths Based on Knowledge Graph. In: *Proceedings of the 2024 International Conference on Computational Linguistics and Natural Language Processing (CLNLP), 19–21 July 2024, Yinchuan, China.* IEEE; 2024. pp. 12–16. https://doi.org/10.1109/CLNLP64123.2024.00011
15. Wei X., Chen H., Yu H., et al. Guided Knowledge Generation with Language Models for Commonsense Reasoning. In: *Findings of the Association for Computational Linguistics: EMNLP 2024, 12–16 November 2024, Miami, USA*. Association for Computational Linguistics; 2024. pp. 1103–1136. http://doi.org/10.18653/v1/2024.findings-emnlp.61

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Дагаев Александр Евгеньевич, аспирант кафедры информатики и информационных технологий, Московский политехнический университет, Москва, Российская Федерация.
e-mail: a.e.dagaev@mospolytech.ru

Alexander E. Dagaev, postgraduate student of the Department of Informatics and Information Technologies, Moscow Polytechnic University, Moscow, the Russian Federation.

Статья поступила в редакцию 13.01.2025; одобрена после рецензирования 14.02.2025; принята к публикации 18.02.2025.

The article was submitted 13.01.2025; approved after reviewing 14.02.2025; accepted for publication 18.02.2025.