УДК 004.89
DOI: 10.26102/2310-6018/2025.49.2.028

Искусственный интеллект в задаче генерации дистракторов для тестовых заданий

А.Е. Дагаев✉

Московский политехнический университет, Москва, Российская Федерация

Резюме. Создание качественных дистракторов для тестовых заданий представляет собой трудоемкий процесс, имеющий большое значение для точной оценки знаний. Существующие подходы часто генерируют неправдоподобные или не отражающие типичные ошибки учащихся варианты. В данной статье предлагается алгоритм генерации дистракторов на основе искусственного интеллекта. Он использует LLM для построения сначала правильной цепочки рассуждений для заданного вопроса и ответа, а затем внедряет типичные ошибки для получения неверных, но в то же время убедительных вариантов ответа, стремясь отразить распространенные заблуждения учащихся. Алгоритм был протестирован на вопросах из русскоязычных наборов данных RuOpenBookQA и RuWorldTree. Оценка проводилась как с использованием автоматических метрик, так и экспертами. Результаты показывают, что алгоритм превосходит базовые методы (прямых запросов и семантических изменений), генерируя дистракторы с более высоким уровнем правдоподобия, релевантности, разнообразия и сходства с эталонными дистракторами, созданными человеком. Данная работа вносит вклад в область автоматизированной генерации контрольно-измерительных материалов, предоставляя способствующий созданию более эффективных оценочных материалов инструмент для преподавателей, разработчиков образовательных платформ и исследователей в сфере обработки естественного языка.

Ключевые слова: генерация дистракторов, искусственный интеллект, большие языковые модели, оценка знаний, тестовые задания, автоматическая генерация тестов, NLP.

Для цитирования: Дагаев А.Е. Искусственный интеллект в задаче генерации дистракторов для тестовых заданий. *Моделирование, оптимизация и информационные технологии.* 2025;13(2). URL: https://moitvivt.ru/ru/journal/pdf?id=1915 DOI: 10.26102/2310-6018/2025.49.2.028

Artificial intelligence in distractor generation for test items

A.E. Dagaev✉

Moscow Polytechnic University, Moscow, the Russian Federation

Abstract. Creating high-quality distractors for test items is a labor-intensive task that plays a crucial role in the accurate assessment of knowledge. Existing approaches often produce implausible alternatives or fail to reflect typical student errors. This paper proposes an AI-based algorithm for distractor generation. It employs a large language model (LLM) to first construct a correct chain of reasoning for a given question and answer, and then introduces typical misconceptions to generate incorrect but plausible answer choices, aiming to capture common student misunderstandings. The algorithm was evaluated on questions from the Russian-language datasets RuOpenBookQA and RuWorldTree. Evaluation was conducted using both automatic metrics and expert assessment. The results show that the proposed algorithm outperforms baseline methods (such as direct prompting and semantic modification), generating distractors with higher levels of plausibility, relevance, diversity, and similarity to human-authored reference distractors. This work contributes to the field of automated assessment material generation, offering a tool that supports the development of more effective evaluation resources for educators, educational platform developers, and researchers in natural language processing.

Keywords: distractor generation, artificial intelligence, large language models, knowledge assessment, test items, automated test generation, NLP.

For citation: Dagaev A.E. Artificial intelligence in distractor generation for test items. *Modeling, Optimization and Information Technology*. 2025;13(2). (In Russ.). URL: https://moitvivt.ru/ru/journal/pdf?id=1915 DOI: 10.26102/2310-6018/2025.49.2.028

Введение

Тестовые материалы представляют собой важный элемент в оценке знаний учащихся, позволяя результативно выявлять уровень подготовки в разных предметных областях. Эффективность тестов во многом зависит от качества дистракторов – неверных вариантов ответа, которые должны быть достаточно правдоподобными, чтобы выявлять заблуждения учащихся, но не настолько очевидными, чтобы их можно было легко исключить [1]. Создание таких дистракторов представляет собой сложный и трудоемкий процесс, требующий глубокого педагогического понимания и экспертных знаний в предметной области [2]. Разработка дистракторов вручную часто сопровождается ошибками и несоответствиями, что существенно снижает возможность масштабирования генерации качественных тестов [3].

Появление искусственного интеллекта, в частности, больших языковых моделей (LLM), таких как GPT-4, существенно изменило подходы к генерации контента и создало новые возможности для автоматизации генерации дистракторов [4]. LLM способны обрабатывать большие объемы текста, понимать нюансы контекста и генерировать ответы, близкие к человеческим, что делает их подходящим инструментом для образовательных задач. Однако существующие методы, основанные на LLM, часто используют упрощенные стратегии, такие как семантическое сходство или случайные изменения, которые нередко приводят к созданию дистракторов с низкой образовательной ценностью или не соответствующих типичным ошибкам учащихся [5]. Одним из вариантов решения этой проблемы является возможность генерировать дистракторы, которые одновременно правдоподобны и педагогически значимы, а также адаптируемы к различным языковым и культурным контекстам.

Цель исследования – разработать алгоритм на основе ИИ для автоматической генерации дистракторов на русском языке для вопросов с множественным выбором.

Научная новизна работы заключается в моделировании потенциальных ошибок и заблуждений учащихся, что позволяет на их основе формировать дистракторы, исходя из логических цепочек рассуждений.

Материалы и методы

Использование искусственного интеллекта, особенно больших языковых моделей (LLM), стало важным направлением в автоматизации процесса генерации заданий с множественным выбором. В последние годы было предложено множество подходов, в которых LLM применяются для создания как самих вопросов, так и дистракторов. Несмотря на высокую языковую адаптивность моделей, обеспечение педагогической релевантности и когнитивной правдоподобности дистракторов по-прежнему остается нерешенной задачей.

В ряде исследований дистракторы генерируются на основе прогнозирования выбора ответа учащихся, что позволяет моделировать типичные ошибки и заблуждения, как показано в работе [6]. Такой подход обеспечивает более точное соответствие дистракторов когнитивным особенностям целевой аудитории и способствует формированию оценочных материалов более высокого качества. Исследование [7] также

показало, что разделение процесса генерации вопросов и дистракторов позволяет добиться лучшей тематической согласованности и семантического разнообразия.

Тем не менее, при генерации дистракторов напрямую через LLM без дополнительного структурирования дистракторы нередко оказываются поверхностными или чрезмерно очевидными [8]. Отдельное направление связано с итеративным улучшением дистракторов путем внутренней самопроверки и ранжирования вариантов по различным метрикам качества, что позволяет повысить их семантическую точность и педагогическую ценность [9].

Для специализированных дисциплин, таких как медицина, было показано, что LLM способны генерировать дистракторы, не уступающие по правдоподобию вариантам, созданным людьми с начальным уровнем подготовки. Однако достичь уровня экспертных решений им пока не удается, особенно в контексте задач, требующих глубокого учета предметной специфики [10]. Для повышения качества таких дистракторов предлагается использовать методы обучения с подкреплением, основанные на предпочтениях пользователей и педагогических критериях [11].

Другие исследования демонстрируют потенциал так называемого «предиктивного промтинга», при котором генерация дистракторов опирается на логическую структуру рассуждений. Это позволяет уменьшить семантическое совпадение с правильным ответом и повысить диагностическую ценность [12]. Также отмечено, что без явной настройки запроса к модели, LLM часто создают нефункциональные или слишком очевидные дистракторы, что снижает надежность таких материалов в образовательной оценке [13].

Совокупность этих исследований демонстрирует как потенциал, так и ограничения современных подходов при генерации дистракторов. Основные проблемы включают поверхностную семантическую трансформацию, отсутствие моделирования ошибок учащихся и ограниченные возможности адаптации к языкам, отличным от английского. Предлагаемый в данной работе алгоритм направлен на преодоление таких ограничений за счет поэтапного моделирования правильных цепочек рассуждений и внедрения типичных когнитивных ошибок. Подход позволяет создавать дистракторы, которые одновременно правдоподобны, разнообразны, тематически релевантны и приближены по качеству к вариантам, созданным экспертами вручную.

Для оценки предложенного алгоритма были выбраны два русскоязычных набора данных, которые хорошо подходят для задачи генерации дистракторов:

RuOpenBookQA представляет собой русскоязычную версию набора данных OpenBookQA, содержащий вопросы по естественно-научным дисциплинам на русском языке в формате вопросов на множественный выбор (MCQ). Каждый вопрос сопровождается четырьмя вариантами ответа, из которых один правильный и три дистрактора, созданные человеком. Это позволяет проводить их прямое сравнение с автоматически создаваемыми дистракторами.

RuWorldTree, в свою очередь, является адаптацией на русский язык корпуса WorldTree, разработанного для задач оценки понимания фактов из научных областей. Он включает в себя не только вопросы с выбором ответа по естественным наукам, но и подробные графы объяснений, состоящие из взаимосвязанных предложений, которые шаг за шагом приводят к правильному ответу.

Оба набора данных были предварительно обработаны для обеспечения консистентности с вопросами и ответами, отформатированными для совместимости с запросами алгоритма к LLM. Из каждого набора проводилась случайная выборка 200 вопросов.

Качество сгенерированных дистракторов оценивалось с использованием комбинации автоматических и экспертных метрик, каждая из которых предназначена для оценки различных аспектов эффективности дистракторов:

– Оценка правдоподобия экспертом. Важный показатель, поскольку создание полноценных автоматизированных систем для определения правдоподобия дистракторов остается актуальной задачей. Эксперты классифицировали степень правдоподобия каждого дистрактора по трехуровневой шкале («неправдоподобно» – 0, «частично правдоподобно» – 1, «правдоподобно» – 2) после ознакомления с вопросом и правильным ответом. Итоговым показателем является средний балл, рассчитанный по ответам пяти экспертов.

– Семантическое сходство дистрактора с вопросом. Измеряется, насколько тематически дистрактор связан с исходным вопросом. Показатель рассчитывается как среднее косинусное сходство между эмбеддингами дистрактора и вопроса.

– Семантическое различие дистрактора с ответом. Оценивается, насколько семантически дистрактор отличается от правильного варианта. Сгенерированные дистракторы не должны быть простыми перефразированиями или вариациями правильного ответа [14].

– Сходство с эталонными дистракторами. Метрика оценивает, насколько сгенерированные дистракторы похожи на созданные человеком. Высокое сходство может указывать на то, что сгенерированные материалы отражают ключевые характеристики, которые в них закладывали разработчики тестов. Расчет производится по среднему значению BERTScore (F1) между каждым сгенерированным дистрактором и эталонным (созданным человеком) для того же вопроса. BERTScore был выбран вместо метрик BLEU и ROUGE из-за его семантической насыщенности.

– Внутригрупповое семантическое разнообразие. Анализируется семантическая вариативность дистракторов, сгенерированных для одного вопроса. Высокая степень разнообразия указывает на способность алгоритма охватывать разные типы потенциальных ошибок.

Алгоритм сравнивался с генерацией без явного моделирования рассуждений, при которой LLM предлагала неправильные ответы на основе контекста вопроса. Также сравнение было выполнено с подходом семантической трансформации правильного ответа, предполагающего создание дистракторов путем замены отдельных элементов на синонимы, его переформулировки либо синтаксической перестройки исходного варианта.

Алгоритм сравнивался с двумя альтернативными подходами. Первый – генерация без явного моделирования рассуждений, при которой LLM формировала неправильные ответы, опираясь исключительно на контекст вопроса. Второй – метод семантической трансформации правильного ответа, при котором дистракторы формируются путем замены отдельных компонентов на синонимичные выражения, переформулировки утверждения или его синтаксической перестройки.

Для каждого из 200 вопросов из RuOpenBookQA и RuWorldTree по три дистрактора генерировались с использованием предлагаемого алгоритма, базовой генерации и семантической трансформации. В качестве модели использовалась LLM GPT o3-mini.

Результаты

Алгоритм разработан для создания релевантных дистракторов с использованием LLM в структурированной, основанной на рассуждениях форме. Его работа проходит в три этапа, которые заключаются в следующем:

1. Генерация правильной цепочки рассуждений:

– Входные данные: Вопрос (Q) и его правильный ответ (A).

– Процесс: LLM используется для создания подробного объяснения каждого шага, как прийти к (A), обеспечивая соответствие намерению вопроса и предметной области. Запрос к LLM выглядит так: «Необходимо пошагово объяснить, как получить ответ {A} из вопроса {Q}».

– Выходные данные: Ожидаемая правильная цепочка рассуждений (C), которая служит основой для последующего внедрения ошибок.

– Представление этапа:

$$C = LLM(P_{correct}),$$

где $P_{correct} = f(Q, A)$, а f – функция, конструирующая запрос путем конкатенации вопроса, ответа и инструктивного текста в виде промта из описанного ранее процесса.

2. Генерация ошибочных цепочек рассуждений:

– Входные данные: Правильная цепочка рассуждений (C) и желаемое количество дистракторов (N).

– Процесс: LLM используется для выявления (N) распространенных ошибок на основе (C), каждая из которых приводит к неверному ответу. Запрос выглядит следующим образом: «На основе правильного рассуждения {C}, необходимо перечислить {N} распространенных ошибок, которые можно совершить в ходе рассуждения».

– Выходные данные: Ответ (R), содержащий (N) ошибочных цепочек рассуждений, каждая из которых связана с неверным ответом.

– Представление этапа:

$$R = \text{LLM}(P_{\text{mistakes}}),$$

где $P_{\text{mistakes}} = g(C, N)$, а g – функция запроса, интегрирующая (C) и указывающая (N).

3. Извлечение и валидация дистракторов:

– Входные данные: Ответ LLM (R).

– Процесс: Из ответа (R) извлекаются результаты рассуждений (N) неверных цепочек, которые становятся дистракторами. Валидация позволяет подтвердить, что дистракторы уникальны и отличаются от изначально правильного ответа (A). Если извлечено меньше (N) уникальных дистракторов, процесс генерации повторяется.

– Выходные данные: Набор дистракторов $D = \{d_1, d_2, \ldots, d_N\}$

– Представление этапа:

$$D = \text{parse}(R).$$

Результаты для набора данных RuOpenBookQA представлены в Таблице 1.

Таблица 1 – Оценка качества дистракторов на RuOpenBookQA
Table 1 – Distractor quality evaluation on RuOpenBookQA

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

Сводные результаты для набора данных RuWorldTree представлены в Таблице 2.

Таблица 2 – Оценка качества дистракторов на наборе RuWorldTree
Table 2 – Distractor quality evaluation on RuWorldTree

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

Обсуждение

Полученные результаты демонстрируют эффективность предложенного алгоритма генерации дистракторов, основанного на моделировании рассуждений, по сравнению с базовыми методами семантических изменений и прямой генерации. Преимущество алгоритма особенно заметно по метрикам экспертной оценки правдоподобия (1,66 на RuOpenBookQA, 1,60 на RuWorldTree) и сходства с эталонными дистракторами (0,58 и 0,56 соответственно). Высокие оценки указывают на то, что сгенерированные дистракторы воспринимаются экспертами как значительно более правдоподобные и содержательные по сравнению с результатами других подходов.

Ключевым фактором такого результата является механизм явного построения корректной цепочки рассуждений и последующего внедрения в нее типичных ошибок. В этом состоит отличие от метода семантической трансформации, в котором правильный ответ подвергается поверхностным изменениям (синонимы, переформулировки и т. д.), нередко приводящим к семантически близким, но педагогически менее значимым дистракторам. Базовая (прямая) генерация, в свою очередь, опирается исключительно на обобщенные знания LLM и не учитывает причинно-следственные связи между вопросом и ответом, что снижает глубину полученных итоговых альтернативных ответов.

Высокие значения по показателям тематической релевантности (0,72 и 0,69) подтверждают, что дистракторы остаются связанными с контекстом вопроса. Это объясняется тем, что генерация опирается на логическую структуру рассуждений, построенную в отношении конкретного вопроса. Одновременно достигается наибольшее семантическое различие с правильным ответом (0,68 и 0,65), что указывает на отсутствие простого перефразирования и тем самым повышает диагностическую ценность дистракторов.

Метрика внутригруппового семантического разнообразия (0,74 и 0,64) также демонстрирует, что для одного вопроса создаются дистракторы, отражающие различные типы ошибок, что особенно важно при проектировании тестов с множественным выбором. Наличие разных, но правдоподобных дистракторов позволяет эффективнее проверять глубину и устойчивость знаний учащихся.

Устойчивость результатов на двух наборах данных – RuOpenBookQA и RuWorldTree подтверждает универсальность предложенного подхода. В то же время несколько более низкие показатели на RuWorldTree могут быть связаны с повышенной сложностью самих вопросов, основанных на графах объяснений. Это указывает на возможные ограничения LLM при работе с более глубокими и сложными структурами знаний.

Результаты подчеркивают положительное влияние включения этапа явного моделирования рассуждений и ошибок в процесс генерации дистракторов. Это позволяет получить дистракторы, которые не только неверны, но и более правдоподобны, релевантны, разнообразны и близки к созданным человеком примерам.

Несмотря на свои сильные стороны, алгоритм потенциально имеет ряд ограничений. Зависимость алгоритма от знаний предметной области LLM может приводить к несоответствиям в узкоспециализированных областях, где у LLM было недостаточно обучающих данных. Кроме того, получение ответа от большой языковой модели для извлечения дистракторов бывает осложнено в условиях неструктурированного вывода модели. В проведенных экспериментах около 6 % ответов требовали повторного запроса для достижения необходимого количества дистракторов.

Заключение

Предложенный алгоритм генерации дистракторов на основе искусственного интеллекта демонстрирует значительное улучшение по сравнению с базовыми методами для вопросов с множественным выбором. Благодаря явному моделированию рассуждений и правдоподобных ошибок, алгоритм создает дистракторы, которые оцениваются автоматическими показателями оценок и экспертами как более правдоподобные, близкие к эталонным вариантам, при этом сохраняющие релевантность вопросу, имеющие отличия от правильного ответа и демонстрирующие большее разнообразие. Статья вносит вклад в область автоматизированной генерации контрольно-измерительных материалов. Такой подход представляет перспективное направление для использования LLM с целью создания более эффективных и диагностически ценных образовательных средств.

Работа вносит вклад в область автоматизированной генерации контрольно-измерительных материалов и может быть полезна преподавателям, разработчикам образовательных платформ, исследователям в области образовательных технологий и обработки естественного языка.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Awalurahman H.W., Budi I. Automatic Distractor Generation in Multiple-Choice Questions: A Systematic Literature Review. *PeerJ Computer Science*. 2024;10. https://doi.org/10.7717/peerj-cs.2441
2. Kumar A.P., Nayak A., K. M.Sh., Goyal Sh., Chaitanya. A Novel Approach to Generate Distractors for Multiple Choice Questions. *Expert Systems with Applications*. 2023;225. https://doi.org/10.1016/j.eswa.2023.120022
3. Bitew S.K., Hadifar A., Sterckx L., Deleu J., Develder Ch., Demeester Th. Learning to Reuse Distractors to Support Multiple-Choice Question Generation in Education. *IEEE Transactions on Learning Technologies*. 2022;17:375–390. https://doi.org/10.1109/tlt.2022.3226523
4. Artsi Ya., Sorin V., Konen E., Glicksberg B.S., Nadkarni G., Klang E. Large Language Models for Generating Medical Examinations: Systematic Review. *BMC Medical Education*. 2024;24. https://doi.org/10.1186/s12909-024-05239-y
5. Shi F., Chen X., Misra K., et al. Large Language Models Can Be Easily Distracted by Irrelevant Context. In: *Proceedings of the 40th International Conference on Machine Learning, ICML 2023: Volume 202, 23–29 July 2023, Honolulu, Hawaii, USA*. PMLR; 2023. P. 31210–31227.
6. Lee Yo., Kim S., Jo Yo. Generating Plausible Distractors for Multiple-Choice Questions via Student Choice Prediction. arXiv. URL: https://arxiv.org/abs/2501.13125 [Accessed 12th April 2025].

7. De-Fitero-Dominguez D., Garcia-Lopez E., Garcia-Cabot A., Del-Hoyo-Gabaldon J.-A., Moreno-Cediel A. Distractor Generation Through Text-to-Text Transformer Models. *IEEE Access*. 2024;12:25580–25589. https://doi.org/10.1109/access.2024.3361673
8. Zhang L., Zou B., Aw A.T. Empowering Tree-Structured Entailment Reasoning: Rhetorical Perception and LLM-driven Interpretability. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 20–25 May 2024, Torino, Italy*. ELRA and ICCL; 2024. P. 5783–5793.
9. Feng W., Lee J., McNichols H., et al. Exploring Automated Distractor Generation for Math Multiple-choice Questions via Large Language Models. In: *Findings of the Association for Computational Linguistics: NAACL 2024, 16–21 June 2024, Mexico City, Mexico.* Association for Computational Linguistics; 2024. P. 3067–3082. https://doi.org/10.18653/v1/2024.findings-naacl.193
10. Cai X., Wang Ch., Long Q., Zhou Yu., Xiao M. Knowledge Hierarchy Guided Biological-Medical Dataset Distillation for Domain LLM Training. arXiv. URL: https://arxiv.org/abs/2501.15108 [Accessed 12th April 2025].
11. Wang R., Jiang Yu., Tao Yu., Li M., Wang X., Ge Sh. High-Quality Distractors Generation for Human Exam Based on Reinforcement Learning from Preference Feedback. In: *Natural Language Processing and Chinese Computing: 13th National CCF Conference, NLPCC 2024: Proceedings: Part IV, 01–03 November 2024, Hangzhou, China*. Singapore: Springer; 2024. P. 94–106. https://doi.org/10.1007/978-981-97-9440-9_8
12. Maity S., Deroy A., Sarkar S. A Novel Multi-Stage Prompting Approach for Language Agnostic MCQ Generation Using GPT. In: *Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024: Proceedings: Part III, 24–28 March 2024, Glasgow, UK*. Cham: Springer; 2024. P. 268–277. https://doi.org/10.1007/978-3-031-56063-7_18
13. Shen Ch.-H., Kuo Yi-L., Fan Ya.-Ch. Personalized Cloze Test Generation with Large Language Models: Streamlining MCQ Development and Enhancing Adaptive Learning. In: *Proceedings of the 17th International Natural Language Generation Conference, 23–27 September 2024, Tokyo, Japan*. Association for Computational Linguistics; 2024. P. 314–319.
14. Wang H.-J., Hsieh K.-Yu, Yu H.-Ch., et al. Distractor Generation Based on Text2Text Language Models with Pseudo Kullback-Leibler Divergence Regulation. In: *Findings of the Association for Computational Linguistics: ACL 2023, 09–14 July 2023, Toronto, Canada.* Association for Computational Linguistics; 2023. P. 12477–12491. https://doi.org/10.18653/v1/2023.findings-acl.790

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 at the Department of Informatics and Information Technologies, Moscow Polytechnic University, Moscow, the Russian Federation.

Статья поступила в редакцию 21.04.2025; одобрена после рецензирования 13.05.2025; принята к публикации 22.05.2025.

The article was submitted 21.04.2025; approved after reviewing 13.05.2025; accepted for publication 22.05.2025.