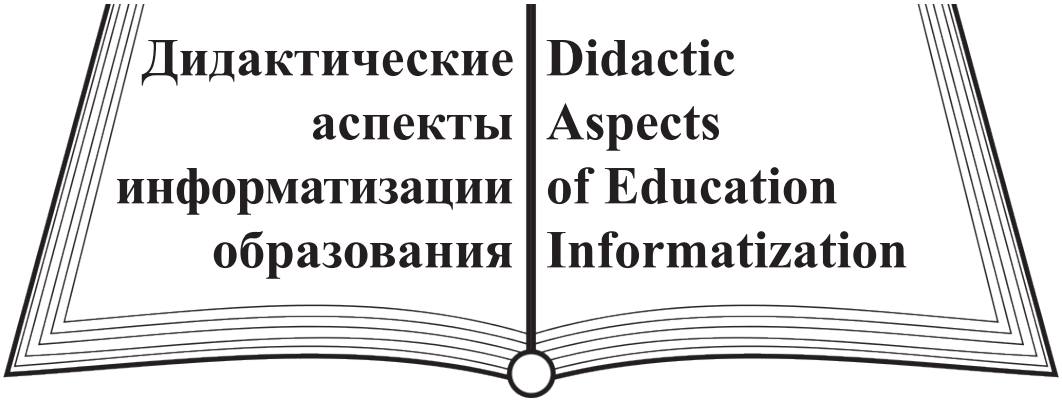

Научная статья
УДК 37
DOI: 10.24412/2072-9014-2025-373-50-60

РАЗРАБОТКА ЗАДАНИЙ К ДИДАКТИЧЕСКИМ ЕДИНИЦАМ ЦИФРОВОГО АДАПТИВНОГО УЧЕБНИКА С ПОМОЩЬЮ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Борис Борисович Ярмахов [1, a] ✉, ***Софья Ильинична Дрейцер*** [2, b]

[1, 2] Московский городской педагогический университет, Москва, Россия

[a] yarmakhovbb@mgpu.ru, https://orcid.org/0000-0001-6217-0871
[b] dreitsersi562@mgpu.ru, https://orcid.org/0000-0001-8549-1627

Аннотация. В данном исследовании поднимаются проблемные вопросы об изменении качества генерации текста с помощью больших языковых моделей на примере генерации вариантов ответа для вопросов с множественным выбором, в рамках разработки адаптивного учебника по биологии. Статья содержит обзор наиболее распространенных метрик для оценки качества решения задач с помощью больших языковых моделей и предлагает авторскую метрику, которая в большей степени подходит для решения обозначенных задач. В результате применения метрики для более чем 1 000 сгенерированных ответов в исследовании делается вывод об их качестве. Также обозначаются недостатки измерения — субъективность экспертного подхода к оценке вариантов ответа, и поднимаются проблемы автоматизации подобных измерений.

Ключевые слова: генеративный искусственный интеллект; вопросы с множественным выбором; метрики оценки качества сгенерированного текста; цифровой адаптивный учебник; цифровая дидактика.

Для цитирования: Ярмахов Б. Б. Разработка заданий к дидактическим единицам цифрового адаптивного учебника с помощью больших языковых моделей / Б. Б. Ярмахов, С. И. Дрейцер // Вестник МГПУ. Серия «Информатика и информатизация образования». 2025. № 3 (73). С. 50–60. https://doi.org/10.24412/2072-9014-2025-373-50-60

Original article
UDC 37
DOI: 10.24412/2072-9014-2025-373-50-60

DEVELOPING ASSIGNMENTS FOR DIDACTIC UNITS OF A DIGITAL ADAPTIVE TEXTBOOK USING LARGE LANGUAGE MODELS

Boris B. Yarmakhov [1, a] ✉, ***Sofya I. Dreitzer*** [2, b]

[1, 2] Moscow City University,
Moscow, Russia

[a] yarmakhovbb@mgpu.ru, https://orcid.org/0000-0001-6217-0871
[b] dreitsersi562@mgpu.ru, https://orcid.org/0000-0001-8549-1627

Abstract. This study raises problematic issues about changing the quality of text generation using large language models using the example of generating answer options for multiple choice questions as part of the development of an adaptive biology textbook. The material provides an overview of the most common metrics for assessing the quality of problem solving using large language models and suggests an author's metric that is more suitable for solving the identified tasks. After applying the metric to more than 1,000 generated responses, the study concludes on the quality of the generated responses. The disadvantages of measurement are also highlighted, namely the subjectivity of the expert approach to evaluating response options and the problems of automating such measurements are raised.

Keywords: generative artificial intelligence; multiple choice questions; metrics for evaluating the quality of generated text; digital adaptive textbook; digital didactics.

For citation: Yarmakhov B. B. Developing assignments for didactic units of a digital adaptive textbook using large language models / B. B. Yarmakhov, S. I. Dreitzer // MCU Journal of Informatics and Informatization of Education. 2025. № 3 (73). P. 50–60. https://doi.org/10.24412/2072-9014-2025-373-50-60

Введение

Наиболее популярным и проверенным инструментом для проверки знаний учащихся при онлайн-обучении является тестирование — способ оценки знаний, умений или поведения, представляющий собой вопросы с несколькими вариантами ответа, при этом правильными могут быть один или несколько вариантов. Основные критерии качества тестирования — надежность, валидность, трудность и дискриминативность тестов [1; 2].

Одним из критериев для тестовых вопросов является качество дистракторов, то есть неправильных ответов, которые предлагаются ученику для выбора [2]. В свою очередь, качество дистракторов оценивается по следующим критериям: они должны быть однородными, правдоподобными, не слишком длинными или переусложненными. Также они не должны содержать формулировки «все из перечисленного», «ничего из перечисленного» и субъективные оценочные слова, такие как «часто» и «обычно» [3]. Кроме этого, рекомендуется избегать слишком длинных правильных ответов, логических и грамматических подсказок, слова «кроме» или «не» в вводной части [4].

В качестве более общих критериев можно назвать следующие [5–7]:

– дистракторы должны быть очевидно неверными, в отличие от правильных ответов;

– дистракторы должны отличаться друг от друга по сути и содержанию;

– дистракторы должны быть правдоподобными, то есть в их основе должна лежать какая-либо логическая ошибка, которую мог бы допустить учащийся, размышляя над ответом на вопрос.

К числу наиболее распространенных способов оценки качества дистракторов относится их проверка в процессе обучения и оценивания учащихся. Например, те из них, которые выбирают меньше, чем в 5 % случаев, считаются нефункциональными и нуждаются в пересмотре в тестовом вопросе [6]. Эффективность тестового вопроса обозначается как высокая, средняя или низкая, где высокая означает, что нефункциональные дистракторы отсутствуют, средняя — что есть один или два нефункциональных дистрактора, а низкая — что есть три или четыре нефункциональных дистрактора из пяти вариантов ответа [7]. Такой способ оценки распространен при определении надежности и валидности тестового инструмента [3; 4].

Методы исследования

С распространением информационных технологий оценка эффективности дистракторов превратилась в задачу для автоматизации. В связи с этим для оценки качества дистракторов с помощью больших языковых моделей используются и другие критерии. В. Фо, А. Лигозат и Б. Грау оценивают семантическую однородность дистракторов: это значит, что они относятся к одной предметной области, но не идентичны друг другу (как, например, «Франция» и «Беларусь» по отношению друг к другу). Для этого исследователи разработали автоматический метод различения дистракторов и недистракторов на основании ряда специализированных метрик, которые авторы называют синтаксической гомогенностью и семантической гомогенностью [8]. П. Балдвин и другие исследователи рассматривают способы автоматической генерации дистракторов на основе их семантической схожести с банком вопросов и ответов, чтобы найти правдоподобные, но неправильные ответы на соответствующие вопросы [9].

Э. Тран и другие оценивают дистракторы путем сравнения их с правильным ответом [10]. В исследовании, где дистракторы генерировались с помощью моделей GPT-3, GPT-4, было установлено, что GPT-3 в 30,8 % случаев давал корректные ответы, то есть один правильный ответ и три правдоподобных, но неправильных, а еще в 10 % случаев — правильный ответ и правдоподобный ответ, который также мог считаться правильным. По отношению к GPT-4 статистика была значительно выше: 75,3 % для вопросов, где есть один правильный ответ и три правдоподобных, и 9,9 % случаев — для вопросов с правильным ответом и правдоподобным ответом, который тоже мог считаться правильным. В исследовании качество дистракторов оценивалось экспертно, производился подсчет количества верных ответов и правдоподобных дистракторов [10]. Отдельным вопросом является оценка качества вариантов ответа, сгенерированных с помощью технологий искусственного интеллекта. В этом контексте необходимо подчеркнуть, что каждая метрика призвана решить сразу две различные задачи: оценку сгенерированного текста и оценку содержания варианта ответа, поэтому необходимо в каждой из метрик рассматривать качество и того и другого.

Для оценки текста существуют как традиционные метрики, которые основаны на прямом подсчете совпадения слов и словосочетаний эталонного и сгенерированного текстов, так и нейросетевые метрики, которые используют векторные сравнения (embeddings) для определения степени схожести эталонного и сгенерированного текстов (BERT, Yisi) [11; 12]. Необходимо рассмотреть основные традиционные метрики для оценки качества сгенерированного текста, а также его содержания, то есть непротиворечивости сгенерированного текста содержанию исходного текста. Приведенные ниже метрики разработаны для оценки качества машинного перевода. Однако они используются и для оценки сгенерированного текста, так как уникальные метрики для оценки именно сгенерированного текста находятся еще в процессе разработки [13].

Метрика BLEU (BiLingual Evaluation Understudy) включает подсчет слов и словосочетаний из машинного перевода, которые встречаются также в эталонном переводе. Затем полученное число делится на общее количество слов и словосочетаний в машинном переводе, что дает значение «точность». Чтобы не исказить значение метрики из-за слишком коротких или неполных текстов, применяют «штраф за краткость», то есть понижающий коэффициент для «точности». Сначала вычисляются *n*-граммы, соответствующие предложениям. Затем *n*-граммы суммируются и делятся на количество *n*-грамм-кандидатов в тестовом корпусе, чтобы вычислить модифицированную оценку точности (*pn*) для всего тестового корпуса [14].

$$BLEU = BP \cdot \exp\left(\sum_{i=1}^{n} w_i \cdot \log p_i\right), BP = \min\left(1, \frac{c}{r}\right),$$

где w_i — положительные веса для каждого используемого параметра *n*-грамм, n — максимальная длина *n*-грамм, i — длина блока в пределах *n*-граммы,

p_i — модифицированная точность n-грамм, c — длина полученного машинного перевода, r — длина наилучшего совпадающего эталонного текста) [15].

Специфика метрики BLUE для анализа сгенерированных вариантов ответов заключается в следующем: данная метрика может быть использована как отношение слов, встречающихся в сгенерированных вариантах ответов, к тем словам, которые эксперт использовал при создании вариантов ответа. Однако применение данной метрики на практике нуждается в сложной автоматизации, так как сложности возникают на уровне алгоритма сравнения сгенерированных и экспертных вариантов ответа, выделения массива текста, предназначенного для сравнения, а также исключения или включения повторяющихся слов.

Метрика ROUGE, предложенная Ч. Лин, сходна с метрикой BLEU, так как тоже основана на подсчете совпадений слов и словосочетаний машинного и экспертного текстов. По сути ROUGE — это целый класс метрик, с помощью которых можно анализировать отдельные слова и словосочетания n-длины. Кроме непосредственной метрики подсчета «точность», в ней также используется «полнота» (recall) и мера F1. «Полнота» определяет, какую долю текста метрика не учитывает, то есть это обратная метрика к «точности». F1 позволяет определить соотношение метрик точности и полноты. Чем ближе к 1 каждая из метрик, тем более качественным является сгенерированный текст [16].

$$ROUGE(N) = \frac{\sum_{S \in \{\mathrm{Re}\, ferenceSummaries\}} \sum_{gram_n \in S} count_{match}(gram_n)}{\sum_{S \in \{\mathrm{Re}\, ferenceSummaries\}} \sum_{gram_n \in S} count(gram_n)}.$$

Специфика метрики ROUGE для анализа вариантов ответов такова, что при ее использовании также возникают проблемы автоматизации, в особенности в определении, какие именно конструкции необходимо сравнивать между собой и как их выделять в двух массивах вариантов ответов. Однако в отношении данной методики кажется логичным сравнивать сразу целый ответ, вернее, искать наиболее похожий на экспертный ответ в сгенерированных ответах. В случае отсутствия такового следует присваивать значение 0 или же значение, показывающее, в какой степени экспертный вариант соответствует найденному с точки зрения семантического анализа.

Метрика TER (Translation Edit Rate) основана на подсчете минимального количества правок, необходимых для приведения машинного перевода в полное соответствие с наиболее близким эталонным переводом. Подсчет числа правок осуществляется автоматизированно. Итоговый показатель TER вычисляется путем деления количества требуемых правок на усредненную длину предложения по всем эталонам [17].

$$TER = \frac{\text{Число редактирований}}{\text{Средняя длина эталонных переводов}}.$$

С точки зрения анализа сгенерированных вариантов ответа для тестовых вопросов с помощью TER метрика позволяет оценить скорее формальные характеристики текста и не говорит о его содержании, так как автоматическая оценка количества правок не уточняет их характер — идет ли речь о грамматике или содержании ответа.

В рамках данного исследования была разработана собственная метрика оценки, основанная на изучении методологии применения традиционных метрик и учитывающая как формальные характеристики текста, так и содержание сгенерированных вариантов ответа. В метрике мы использовали распространенные ответы для формализации оценки: соответствие языковым нормам (fluent/influent), правильное содержание кода или неправильное содержание дистрактора (correctness/incorrectness), правдоподобность дистрактора (plausibility), недостаточное разнообразие дистрактора (diversity).

При использовании данной метрики были проанализированы варианты ответов для тестовых вопросов к учебнику биологии 6-го класса в рамках разработки цифрового адаптивного учебника [18; 19], на основе школьного учебника биологии под редакцией В. В. Пасечника и С. В. Суматохина [20]. Всего было проанализировано 332 вопроса и 2 159 вариантов ответа.

Процесс анализа был выстроен следующим способом. Сначала были сгенерированы вопросы на основе текста учебника, по 6 или 7 вопросов для каждой части учебника. Генерировались как формулировка вопроса, так и ответы к нему. Поскольку учебник адаптивный, было сгенерировано большее количество ответов, чем необходимо для однократного тестирования ученика. В вопросах содержалось либо 2 правильных и 4 неправильных, либо 3 правильных и 4 неправильных ответа.

Затем эксперт оценивал вопросы и исправлял то, что считал нужным. После этого для подсчета метрики сравнивались варианты ответа на те вопросы, в которых эксперт сохранил формулировку. В случае, если была изменена формулировка ответа и внесены правки в окончания слов, а количество добавленных экспертом слов не превышало 35 %, такая правка оценивалась в 0,5 балла. Если формулировка ответа была изменена в значительной степени, количество добавленных слов превышало 50 %, или содержание ответа было полностью изменено, такая правка оценивалась в 1 балл, так как в случае правильного ответа от считался «неверным», а в случае неправильного ответа — «верным», «неправдоподобным» или «неразнообразным». Соответственно, каждому ответу был присвоен балл 0, 0,5 или 1, в зависимости от того, насколько он был ошибочным.

В результате подсчета 152 вопроса или 986 ответов не попали в список анализируемых, так как у них был изменен стем. Количество проанализированных вопросов составило 180, а количество ответов — 1 173.

Результаты и обсуждение

Соотношение баллов, начисленных в соответствии с правками экспертов, и количества сгенерированных вариантов ответа составило 0,584. Это означает, что почти все ответы, так или иначе, содержали незначительные ошибки формулировки, больше половины ответов — грубые ошибки, касающиеся содержания или правдоподобности ответа.

Таблица 1

Соотношение измененных экспертом и неизмененных ответов для подсчета метрики качества генерации вариантов ответа

	Ответы без замен	**Замена формулировки**	**Замена ответа**
Количество ответов	390 из 1 173	197 из 1 173	568 из 1 173
Процентное соотношение ответов к их общему количеству	33 %	17 %	50 %

Причины, по которым эксперты заменяли варианты ответа:

– сгенерированный правильный ответ на самом деле является неверным;

– сгенерированный неправильный ответ на самом деле является верным;

– сгенерированный неправильный ответ кажется слишком неправдоподобным;

– сгенерированный неправильный ответ кажется слишком однообразным;

– эстетические причины (например, ответ вызывает больше положительные ассоциации).

Необходимо отметить, что критерий экспертизы в оценке качества сгенерированных вопросов не является исчерпывающим, так как эксперты могут быть субъективны и иметь предвзятые суждения относительно тех или иных тем, формулировок или вопросов. Учет мнений экспертов и принятие мер для преодоления субъективности остается отдельным проблемным вопросом для исследований.

Заключение

В результате данного исследования можно сформулировать следующие выводы. В исследовании проанализированы и изучены наиболее популярные метрики для оценки машинного перевода текста, такие как BLEU, ROUGE и TER. В результате анализа была выявлена необходимость создания собственной метрики для оценки качества сгенерированных ответов для вопросов с множественным выбором.

Так как проанализированные метрики предназначены для оценки машинного перевода, они не позволяют с большой точностью оценить качество сгенерированных вариантов ответа, поэтому разработка такой метрики является

актуальной. В исследовании также обозначены вопросы по автоматизации метрики и возможности машинного подсчета соответствия между сгенерированными и экспертными ответами, что является заделом для дальнейших исследований. Еще одной темой для последующих исследований является преодоление субъективности экспертов в отношении оценки сгенерированных вариантов ответа.

В процессе исследования было проанализировано качество сгенерированных вопросов для учебника биологии 6-го класса, метрика составила 0,584. Это означает, что больше половины ответов были сгенерированы с ненадлежащим качеством.

Список источников

1. *Крокер Л.* Введение в классическую и современную теорию тестов: учебник / Л. Крокер, Дж. Алгина. М.: Логос, 2010. 667 с.

2. *Карпенко А. П.* Тестовый метод контроля качества обучения и критерии качества образовательных тестов. Обзор / А. П. Карпенко, А. С. Домников, В. В. Белоус // Наука и образование. 2011. № 4. С. 1–28.

3. *Kowash M.* Evaluating the Quality of Multiple Choice Question in Paediatric Dentistry Postgraduate Examinations / M. Kowash, I. Hussein, M. Al Halabi // Sultan Qaboos Univ Med J. 2019. No. 19 (2). P. e135–e141.

4. *Sajjad M.* Nonfunctional distractor analysis: An indicator for quality of Multiple choice questions / M. Sajjad, S. Iltaf, R. A. Khan // Pak J Med Sci. 2020. No. 36 (5). P. 982–986.

5. *Vatsal R.* Assessing Distractors in Multiple-Choice Tests / R. Vatsal, V. Raina, A. Liusie, M. Gales // Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems. 2023. P. 12–22.

6. *Ansari M.* Assessment of distractor efficiency of MCQS in item analysis / M. Ansari, R. Sadaf, A. Akbar // Professional Med. Journal. 2022. No. 29 (5). P. 730–734.

7. *Qiu Zh.* Automatic Distractor Generation for Multiple Choice Questions in Standard Tests / Zh. Qiu, X. Wu, W. Fan // Proceedings of the 28th International Conference on Computational Linguistics. 2020. P. 2096–2106.

8. *Pho V.-M.* Distractor quality evaluation in Multiple Choice Questions / V.-M. Pho, A.-L. Ligozat, B. Grau // Artificial Intelligence in Education. 2015. Vol. 9112. P. 377–386.

9. *Baldwin P.* A Natural-Language-Processing-Based Procedure for Generating Distractors for Multiple-Choice Questions / P. Baldwin, J. Mee, V. Yaneva // Evaluation & the Health Professions. 2022. No. 45 (4). P. 327–340.

10. *Tran A.* Generating Multiple Choice Questions for Computing Courses Using Large Language Models / A. Tran, K. Angelikas, E. Rama // IEEE Frontiers in Education Conference (FIE). 2023. P. 1–8.

11. *Соснин А. В.* Взаимосвязь экспертных категорий и автоматических метрик, используемых для оценки качества перевода / А. В. Соснин, Ю. В. Балакина, А. Н. Кащихин // Вестник Санкт-Петербургского университета. Серия: Язык и литература. 2022. Т. 19. № 1. С. 125–148.

12. *Ярмахов Б. Б.* Генерация вопросов к адаптивному учебнику по биологии на основе технологий искусственного интеллекта / Б. Б. Ярмахов, С. И. Дрейцер // Педагогическая инноватика и непрерывное образование в XXI веке: сб. науч. тр.

II Междунар. науч.-практ. конф. (Киров, 20 мая 2024 г.). Киров: Вятский ГАТУ, 2024. С. 161–165.

13. *Тихонова М. И.* Методы оценивания языковых моделей в задачах понимания естественного языка: дис. ... канд. физ.-мат. наук / М. И. Тихонова. Москва, 2023. 77 с.

14. *Snover M.* A study of translation edit rate with targeted human annotation / M. Snover, B. Dorr, R. Schwartz // Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers. Cambridge, 2006. P. 223–231.

15. *Нуриев В. А.* Методы оценки качества машинного перевода: современное состояние / В. А. Нуриев, А. Ю. Егорова // Информатика и ее применения. 2021. Т. 15. Вып. 2. С. 104–111.

16. *Тимохин И. В.* Автоматизация генерации заголовков новостных статей / И. В. Тимохин, Н. Б. Осипенко // Проблемы физики, математики и техники. 2020. Вып. 3 (44). С. 92–94.

17. *Митренина О. В.* Как и какой перевод (не) оценивают компьютеры / О. В. Митренина, А. Г. Мухамбеткалиева // Journal of applied linguistics and lexicography. 2021. Т. 3. № 2. С. 77–84.

18. *Ярмахов Б. Б.* Цифровой адаптивный учебник биологии: разработка и Апробация / Б. Б. Ярмахов, С. В. Суматохин, О. В. Кукушкина // Биология в школе. 2024. № 2. С. 23–31.

19. *Ярмахов Б. Б.* Трансформация диалога в эпоху больших языковых моделей / Б. Б. Ярмахов // Диалог культур — культура диалога в многонациональном городском пространстве: материалы Четвертой междунар. науч.-практ. конф. (Москва, 27 февраля 2024 г.). М.: Языки Народов Мира, 2024. С. 295–301.

20. *Пасечник В. В.* Биология. 6 класс: учебник / В. В Пасечник, С. В. Суматохин, З. Г. Гапонюк. М.: Просвещение, 2023. 160 с.

References

1. *Crocker L.* Introduction to classical and modern test theory: textbook / L. Crocker, D. Algina. Moscow: Logos, 2010. 668 p.

2. *Karpenko A. P.* Test method of quality control of education and criteria of quality of educational tests. Review / A. P. Karpenko, A. S. Domnikov, V. V. Belous // Mechanical engineering and computer technologies. 2011. No. 4. P. 1–28.

3. *Kowash M.* Evaluating the Quality of Multiple Choice Question in Paediatric Dentistry Postgraduate Examinations / M. Kowash, I. Hussein, M. Al Halabi // Sultan Qaboos Univ Med J. 2019. No. 19 (2). P. e135–e141.

4. *Sajjad M.* Nonfunctional distractor analysis: An indicator for quality of Multiple choice questions / M. Sajjad, S. Iltaf, R. A. Khan // Pak J Med Sci. 2020. No. 36 (5). P. 982–986.

5. *Vatsal R.* Assessing Distractors in Multiple-Choice Tests / R. Vatsal, V. Raina, A. Liusie, M. Gales // Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems. 2023. P. 12–22.

6. *Ansari M.* Assessment of distractor efficiency of MCQS in item analysis / M. Ansari, R. Sadaf, A. Akbar // Professional Med. Journal. 2022. No. 29 (5). P. 730–734.

7. *Qiu Zh.* Automatic Distractor Generation for Multiple Choice Questions in Standard Tests / Zh. Qiu, X. Wu, W. Fan // Proceedings of the 28th International Conference on Computational Linguistics. 2020. P. 2096–2106.

8. *Pho V.-M.* Distractor quality evaluation in Multiple Choice Questions / V.-M. Pho, A.-L. Ligozat, B. Grau // Artificial Intelligence in Education. 2015. Vol. 9112. P. 377–386.

9. *Baldwin P.* A Natural-Language-Processing-Based Procedure for Generating Distractors for Multiple-Choice Questions / P. Baldwin, J. Mee, V. Yaneva // Evaluation & the Health Professions. 2022. No. 45 (4). P. 327–340.

10. *Tran A.* Generating Multiple Choice Questions for Computing Courses Using Large Language Models / A. Tran, K. Angelikas, E. Rama // IEEE Frontiers in Education Conference (FIE). 2023. P. 1–8.

11. *Sosnin A. V.* The relationship of expert categories and automatic metrics used to assess the quality of translation / A. V. Sosnin, Yu. V. Balakina, A. N. Kashchikhin // Bulletin of the Saint Petersburg University. Language and literature. 2022. Vol. 19. No. 1. P. 125–148.

12. *Yarmakhov B. B.* Generation of questions for an adaptive biology textbook based on artificial intelligence technologies / B. B. Yarmakhov, S. I. Dreitzer // Pedagogical innovation and continuing education in the 21st century: proceedings of the II International Scientific and Practical Conference (Kirov, May 20, 2024). Kirov: Vyatka GATU, 2024. P. 161–165.

13. *Tikhonova M. I.* Methods of evaluating language models in the tasks of understanding natural language: diss. kand. a computer.Sciences / M. I. Tikhonova. Moscow. 2023. 77 p.

14. *Snover M.* A study of translation edit rate with targeted human annotation / M. Snover, B. Dorr, R. Schwartz // Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers. Cambridge, 2006. P. 223–231.

15. *Nuriev V. A.* Methods for assessing the quality of machine translation: the current state / V. A. Nuriev, A. Yu. Egorova // Inform. and its application. 2021. Vol. 15. No. 2. P. 104–111.

16. *Timokhin I. V.* Automation of headline generation for news articles / I. V. Timokhin, N. B. Osipenko // Problems of physics, mathematics and engineering. 2020. No. 3 (44). P. 92–94.

17. *Mitrenina O. V.* How and which translation is (not) evaluated by computers / O. V. Mitrenina, A. G. Mukhambetkalieva // Journal of applied linguistics and lexicography. 2021. Vol. 3. No. 2. P. 77–84.

18. *Yarmakhov B. B.* Digital adaptive biology textbook: development and testing / B. B. Yarmakhov, S. V. Sumatokhin, O. V. Kukushkina // Biology at school. 2024. No. 2. P. 23–31.

19. *Yarmakhov B. B.* The transformation of dialogue in the era of large language models / B. B. Yarmakhov // Dialogue of cultures — the culture of dialogue in a multinational urban space: Proceedings of the Fourth International Scientific and Practical Conference (Moscow, February 27, 2024). Moscow: Languages of the Peoples of the World, 2024. P. 295–301.

20. *Pasechnik V. V.* Biology. 6th grade: textbook / V. V. Pasechnik, S. V. Sumatokhin, Z. G. Gaponyuk. Moscow: Prosveshchenie, 2023. 160 p.

Статья поступила в редакцию: 02.06.2025;
одобрена после рецензирования: 04.08.2025;
принята к публикации: 11.08.2025.

The article was submitted: 02.06.2025;
approved after reviewing: 04.08.2025;
accepted for publication: 11.08.2025.

Информация об авторах / Information about the authors:

Борис Борисович Ярмахов — кандидат философских наук, доцент, доцент департамента информатизации образования, Институт цифрового образования, Московский городской педагогический университет, Москва, Россия.

Boris B. Yarmakhov — Candidate of Philosophical Sciences, Associate Professor, Associate Professor of the Department of Informatization of Education, Institute of Digital Education, Moscow City University, Moscow, Russia.

yarmakhovbb@mgpu.ru, https://orcid.org/0000-0001-6217-0871

Софья Ильинична Дрейцер — аспирант департамента информатизации образования, Институт цифрового образования, Московский городской педагогический университет, Москва, Россия.

Sofya I. Dreitzer — Postgraduate Student of the Department of Informatization of Education, Institute of Digital Education, Moscow City University, Moscow, Russia.

dreitsersi562@mgpu.ru, https://orcid.org/0000-0001-8549-1627

Вклад авторов: все авторы сделали эквивалентный вклад в подготовку публикации. Авторы заявляют об отсутствии конфликта интересов.

Contribution of the authors: the authors contributed equally to this article. The authors declare no conflicts of interest.