Научный и информационно-аналитический педагогический журнал

Отечественная и зарубежная педагогика

№ 5 (101) том 1
2024

СОДЕРЖАНИЕ

ИНФОРМАЦИОННАЯ ОБРАЗОВАТЕЛЬНАЯ СРЕДА

ИСТОРИЯ ПЕДАГОГИКИ

ДОПОЛНИТЕЛЬНОЕ ОБРАЗОВАНИЕ

ФИЛОСОФИЯ ОБРАЗОВАНИЯ

Научный и информационно-аналитический педагогический журнал
«ОТЕЧЕСТВЕННАЯ И ЗАРУБЕЖНАЯ ПЕДАГОГИКА»

Свидетельство о регистрации СМИ ПИ № ФС77-63015 от 10.09.2015.

Учредитель
Федеральное государственное бюджетное научное учреждение «Институт стратегии развития образования»

Журнал включен в Перечень российских рецензируемых научных журналов ВАК

Журнал размещен в каталоге научной периодики РИНЦ на платформе Научной электронной библиотеки eLibrary.ru

Журнал также индексируется в 10 российских и международных базах данных, в том числе: OCLC WorldCat, BASE, ROAR, RePEc, OpenAIRE, Соционет, EBSCO A-to-Z, EBSCO Discovery Service

Адрес редакции
101000, г. Москва, ул. Жуковского, д. 16
Тел.: 8 (495) 621-33-74
E-mail: redactor@instrao.ru
Сайт: ozp.instrao.ru

Периодичность: 6 номеров в год
Тираж 300 экз.
Свободная цена

Верстка: *В. В. Симонова*
Формат 60x90/16. Подписано в печать 25.10.2024.
Печать цифровая. Объем 13 п.л., 204 стр.
ООО «Паблит», г. Москва, ул. Полярная, 31В, стр. 1. Заказ

Уважаемые авторы!
Редакция и учредитель журнала просят присылать предложения о публикации своих статей на адрес редакции.

Индекс для подписчиков по каталогам «Почта России» и «Урал-Пресс»: 83284

12+

Журнал «Отечественная и зарубежная педагогика» включен в Перечень российских рецензируемых научных журналов ВАК

Отечественная и зарубежная педагогика. 2024. Т. 1, № 5 (101). С. 123–137.
Domestic and foreign pedagogy. 2024. Vol. 1, no. 5 (101). P. 123–137.

Научная статья
УДК 372.881.111.1
doi: 10.24412/2224–0772–2024–101–123–137

ИССЛЕДОВАНИЕ ПОТЕНЦИАЛА ГЕНЕРАТИВНЫХ МОДЕЛЕЙ ДЛЯ ОЦЕНИВАНИЯ ЭССЕ И ОБЕСПЕЧЕНИЯ ОБРАТНОЙ СВЯЗИ

С. В. Боголепова

М. Г. Жаркова

Светлана Викторовна Боголепова[1], Марина Геннадьевна Жаркова[2]
[1, 2] Национальный исследовательский университет «Высшая школа экономики», Москва, Россия
[1] sbogolepova@hse.ru, SPIN РИНЦ: 1715-1368, ORCID: 0000-0003-1050-9110
[2] marina.zharkova.2001@mail.ru

Аннотация. В эпоху интенсивного развития генеративных языковых моделей эти инструменты все больше используются преподавателями и студентами. Данная работа посвящена исследованию потенциала использования генеративных моделей, взаимодействующих с пользователем посредством чат-ботов ChatGPT и PerplexityAI, для оценки студенческих эссе, написанных в формате стандартизированного экзамена по английскому языку, и формулировки обратной связи по качеству студенческих работ. С учетом специфики каждого чат-бота и стандартизированных критериев оценивания были сформулированы запросы, на основании которых чат-боты выставили баллы 19 эссе как в целом, так и по отдельным аспектам, а также дали обратную связь. Выставленные баллы были сопоставлены с оценкой преподавателя и друг с другом путем вычисления коэффициентов согласованности (альфа Кронбаха) и межэкспертного согласия (каппы Коэна и Флейса). Хотя согласованность была определена как достаточная или высо-

кая, то есть чат-боты и преподаватель интерпретировали критерии сходным образом, межэкспертное согласие было незначительным. В результате качественного анализа выявлены особенности обратной связи от чат-ботов, такие как периодическое игнорирование инструкций в запросе, тенденция к нахождению несуществующих ошибок, выставление разных баллов одной и той же работе при последовательных запросах. Сделан вывод о том, что чат-боты могут использоваться для приблизительной оценки работ и формулировки обратной связи, но их выдача не может считаться полностью надежной и нуждается в экспертной корректировке.

Ключевые слова: искусственный интеллект, генеративные модели, автоматическая проверка текста, эссе, оценивание, обратная связь

Для цитирования: Боголепова С. В., Жаркова М. Г. Исследование потенциала генеративных моделей для оценивания эссе и обеспечения обратной связи // Отечественная и зарубежная педагогика. 2024. Т. 1, № 5 (101). С. 123–137. doi: 10.24412/2224–0772–2024–101–123–137

Original article

RESEARCHING THE POTENTIAL OF GENERATIVE LANGUAGE MODELS FOR ESSAY EVALUATION AND FEEDBACK PROVISION

Svetlana V. Bogolepova[1], Marina G. Zharkova[2]
[1, 2] National Research University Higher School of Economics, Moscow, Russia
[1] sbogolepova@hse.ru, SPIN РИНЦ: 1715-1368, ORCID: 0000-0003-1050-9110
[2] marina.zharkova.2001@mail.ru

Abstract. In the era of rapid development of generative language models these tools are increasingly being used by both students and instructors. This paper aims to investigate the potential of generative models interacting with users via chatbots ChatGPT и PerplexityAI for the evaluation of standardised essays in English and the provision of feedback on their quality. Accounting for the specific features of each chatbot and standardised assessment criteria, we developed prompts which were consequently fed to the chatbots together with 19 students' essays. The chatbots both awarded overall grades and gave points and feedback on specific aspects. The chatbots' grades were compared to the ones provided by the instructor, and to each other. Cronbach's alpha was used to measure the consistency of grading, whereas Koen's and Fleiss's kappas helped to evaluate inter-rater agreement. Though the consistency of grading among the raters was shown to be from acceptable to excellent on different aspects, which indicates similar interpretations of assessment criteria by the instructor and the chatbots, inter-rater agreement was slight. Qualitative analysis revealed such features of feedback from chatbots as ignoring

instructions in the prompt, finding non-existent errors, or awarding different grades in consecutive inquiries. We conclude that chatbots can be used for rough evaluation of standardised essays; however, their output cannot be considered reliable and needs expert editing.

Keywords: artificial intelligence, generative models, automatic text evaluation, essay, assessment, feedback

For citation: Bogolepova S. V., Zharkova M. G. Researching the potential of generative language models for essay evaluation and feedback provision. *Domestic and Foreign Pedagogy.* 2024;1(5):123–137. (In Russ.). doi: 10.24412/2224–0772–2024–101–123–137

Введение

Сейчас инструменты на основе искусственного интеллекта используются в различных областях, и образование не является исключением. Преподаватели используют подобные инструменты для решения различных профессиональных задач, не преминут ими воспользоваться и студенты [1; 11; 15]. При этом научных работ, посвященных изучению опыта использования таких инструментов участниками образовательного процесса, в силу новизны феномена мало.

Целью данной работы является исследование потенциала генеративных моделей в формате чат-ботов ChatGPT и PerplexityAI для оценки эссе в формате экзамена IELTS и формулировки обратной связи для студента. Данные модели были выбраны благодаря их доступности и обученности на больших объемах данных. Для достижения цели были решены следующие задачи: 1) был сформулирован запрос, побуждающий чат-бот проверить эссе по критериям стандартизированного экзамена, выставить баллы по оцениваемым аспектам и дать детализированную обратную связь по шаблону; 2) на основании запроса чат-ботами были оценены студенческие эссе, и выставленные баллы были статистически сопоставлены с результатами проверки эссе преподавателем; 3) методом сплошной выборки были выявлены черты, присущие обратной связи, сформулированной чат-ботами.

Обзор литературы

В настоящее время потенциал инструментов на основе искусственного интеллекта используется повсеместно [5]. Генеративные языковые модели — это искусственный интеллект, способный анализировать и синтезировать текстовый материал. Модели предварительно обучаются на обширном корпусе текстовых данных, что по-

зволяет им понимать человеческий язык и генерировать логичные и последовательные тексты. Генеративные модели часто взаимодействуют с пользователями в формате чат-бота.

Такие модели становятся незаменимыми помощниками преподавателя иностранного языка, позволяющего ей/ему создавать и адаптировать текстовый материал для занятий, разрабатывать задания разных типов под нужды и запросы учебной аудитории, планировать обучение, анализировать работы студентов по заданным критериям. Все перечисленное способствует персонализации и индивидуализации обучения, а также экономии ресурсов преподавателя [4; 17].

Генерация задается с помощью промпта — текстового запроса, подробно описывающего желаемый результат. Промпт-инжиниринг — относительно новая дисциплина, которая занимается процессом создания текстовых запросов путем поиска подходящих комбинаций входных данных, которые могут обеспечить получение высококачественных ответов от больших языковых моделей [19].

При формулировке текстового запроса пользователи должны иметь четкое представление о том, что они ожидают получить от языковой модели в ответ. Необходимо придерживаться минимального ряда правил для создания рабочего, «насыщенного» запроса: 1) прописать четкие инструкции, используя при этом активные глаголы в императиве; 2) подчеркнуть ключевые элементы; 3) детализировать контекст запроса, указать на формат желаемого ответа, ход работы, роль и др. Необходимо снабдить языковую модель контекстом, то есть достаточным количеством дополнительной информации по теме запроса, которая поможет модели сгенерировать наиболее подходящий и точный ответ [7]. Создание «насыщенного» запроса — это процесс, который требует анализа выдачи и последовательной доработки [23]. Именно преподаватель как эксперт в своей области определяет, что именно необходимо доработать в выдаче перед тем, как использовать ее в учебной аудитории [10].

Хотя системы автоматической обработки и оценки текста на основе искусственного интеллекта используются в узкопрофессиональных кругах уже довольно давно [20], доступность и простота использования генеративных моделей сейчас позволяет рядовым преподавателям иностранного языка применять их потенциал для оценивания студенческих работ и формулировки обратной связи. Предыдущие

исследования показали эффективность автоматических систем и высокий уровень соответствия оценок, данных системой и экспертом [8; 18]. Другими преимуществами являются детальность предоставляемой обратной связи и ее объективность [4]. Беспристрастность автоматической оценки позволяет избежать эффекта гало, то есть влияния предыдущего опыта взаимодействия с автором текста на оценку [16].

Помимо преимуществ, современные исследования выделяют и некоторые недостатки автоматической проверки текста, в том числе с помощью генеративных языковых моделей. Хотя модель успешно анализирует текст с точки зрения прагматики, семантики, связности, формата, синтаксиса, она не способна уловить авторский стиль, передаваемые эмоции, использование риторических приемов [3]. Обратная связь от генеративных моделей на содержание и структуру характеризуется обобщенными формулировками и уступает по качеству обратной связи от преподавателя [14]. Тем не менее исследований, в том числе сравнительных, качества оценивания письменных работ генеративными языковыми моделями еще мало, и данная работа вносит вклад в понимание их потенциала для использования преподавателями иностранного языка.

Методы

Материал исследования

В качестве материала исследования были использованы 19 эссе студентов второго курса НИУ ВШЭ, проходящих курс подготовки к стандартизированному экзамену по английскому языку. Эссе были проверены преподавателем, были выставлены баллы согласно шкале для оценивания письменных работ IELTS как по отдельным аспектам, так и за работу в целом. Оценивались такие аспекты, как достижение коммуникативной задачи (Task Response, далее — TR), связанность (Coherence and Cohesion, далее — C&C), лексический ресурс (Lexical Resource, далее — LR), грамматические навыки (Grammatical Range and Accuracy, далее — GR&A). Баллы могли быть целочисленными или кратными 0,5.

Генеративные модели

В исследовании использовались два чат-бота: ChatGPT от OpenAI и PerplexityAI. Они были выбраны исходя из: а) общедоступности (оба имеют бесплатную версию), б) простоты использования и в) способности обрабатывать и генерировать тексты разных жанров.

Оба чат-бота работают на базе модели OpenAI GPT-3.5 в бесплатной версии, используют алгоритмы машинного обучения для обработки естественного языка и генерации текста. Несмотря на общую языковую модель, чат-боты имеют свои отличия. Особенностью чат-бота PerplexityAI является подключение к сети Интернет, что обеспечивает доступ к наиболее актуальным данным, на которых модель постоянно обучается [9]. При генерации ответа чат-бот также выдает ссылки на источники, подходящие по теме запроса, среди которых есть и сайты о критериях оценивания эссе в экзамене IELTS.

Формулировка запроса

Существует целый ряд методов и стратегий промпт-инжиниринга, и чаще всего пользователи прибегают к смешанной методике, которая включает в себя несколько стратегий и составных элементов. В рамках данной работы после проработки формулировки запроса совместно с выбранными генеративными моделями были выведены текстовые запросы. Такой запрос вместе с шаблоном сообщения студенту представлен в таблице 1.

Таблица 1

Запрос для генеративной модели и шаблон обратной связи

Основная инструкция	Act as an English teacher of “International exam preparation course” with the focus on the IELTS exam. Evaluate an IELTS Academic writing task 2 essay, providing feedback on strengths, weaknesses, and suggest improvements. Assign band scores according to IELTS criteria and band descriptors, determining the overall band score for the essay. Provide band scores for Task Achievement, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy, explaining each score. Highlight positive aspects and drawbacks, listing errors with explanations and corrections. Offer strategies for improvement. Formulate feedback as a text message to the student.

Шаблон оформления обратной связи	Dear Student, I have carefully evaluated your essay and would like to provide you with detailed feedback on your performance. (write about each criteria — task achievement, coherence and cohesion, lexical resource, grammatical range and accuracy — in more detail, including all errors and their corrections, provide the band score for each criteria and consider the band descriptors according to which the score was assigned) Task Achievement: Coherence and Cohesion: Lexical Resource: Grammatical Range and Accuracy: Total Band: Based on the above criteria, the total band for this written work is (provide the overall band score). (suggest areas and strategies for improvement) To improve your performance, I recommend focusing on the following areas and strategies: Overall, you have shown (summarise the strengths and weaknesses). Keep up the good work and never hesitate to seek further guidance or assistance.

Анализ данных

Нами были проанализированы оценки, выставленные преподавателем и чат-ботами по 9-балльной шкале IELTS для всего эссе и отдельных аспектов. Был подсчитан процент оценок, отличающихся не более чем на 1 балл, так как подобный люфт позволяется при оценке стандартизированных экзаменов, а также процент полностью совпадающих баллов. Для оценки согласованности оценок преподавателя и чат-ботов были вычислены α Кронбаха [6], для чего использовался онлайн-инструмент http://www.wessa.net/rwasp_cronbach.wasp.

Коэффициент каппа Коэна использовался для оценки согласия между двумя парами чат-бот — преподаватель и ChatGPT — PerplexityAI, а каппа Флейса применялся для проверки надежности оценки трех проверяющих между собой. Каппа Флейса измеряет общее согласие и по отдельным категориям, которые в данном исследовании представлены оценками по 9-балльной шкале IELTS. Для статистического анализа была использована программа JASP (https://jasp-stats.org/).

Первые два коэффициента находятся в диапазоне от 0 до 1, где 0 обозначает отсутствие согласия, а 1 указывает на идеальное согласие между экспертами. Значения каппы Флейса меньше нуля указыва-

ют на отсутствие согласия между экспертами. Значение между 0,01 и 0,20 показывает незначительное согласие, между 0,21 и 0,40 — справедливое, 0,41–0,60 — умеренное, между 0,61 и 0,80 — существенное и диапазон 0,81–1 указывает на почти идеальное согласие [12]. Интерпретация каппы Коэна также происходит в соответствии с указанными диапазонами.

Методом сплошной выборки в выдаче чат-ботов были обнаружены отличительные черты обратной связи, также были проанализирован опыт работы с генеративными моделями.

Результаты

Статистический анализ

Результаты подсчета процентного соотношения оценок и α Кронбаха представлены в таблице 2. При возможной погрешности в 1 балл во всех трех случаях процент согласия довольно высокий и варьируется от 84 до 100%; при этом наибольшие средние значения наблюдаются в парах преподаватель — PerplexityAI (94%) и ChatGPT — PerplexityAI (97%). Результаты по α Кронбаха показывают, что преподаватель и чат-боты интерпретировали критерии оценивания почти единообразно, коэффициент показывает высокую согласованность для общей оценки, а также оценок по лексическому и грамматическому аспектам (более 0,9). Самая низкая, но тем не менее достаточная согласованность наблюдается по аспекту достижения коммуникативной задачи (0,7). При этом доля оценок, совпадающих полностью, значительно ниже. Для всех трех пар проверяющих самые высокий процент совпадений наблюдается по аспекту связности и логики изложения.

Таблица 2

Согласованность между оценками преподавателя и чат-ботов

Аспект	Преп. — GPT %*	Преп. — GPT % полн.**	Преп. — Perplex. %	Преп. — Perplex. % полн.**	GPT — Perplex. %	GPT — Perplex. % полн.**	α Кронбаха
Общая оценка	95%	21%	95%	10,5%	95%	31,6%	0,918

TA	79%	31,6%	89%	26,3%	100%	42%	0,7
C&C	95%	47,4%	100%	52,6%	95%	63%	0,867
LR	84%	21%	100%	42,1%	95%	15,8%	0,907
GR&A	89%	26,3%	84%	5,3%	100%	31,6%	0,934

%* — процент оценок, отличавшихся не более чем на балл
% полн.** — процент оценок, совпадавших полностью

Проанализируем статистику по каппам Флейса и Коэна, а также значения стандартных ошибок (SE) и доверительного интервала. Стандартная ошибка измеряет точность расчетов, вариабельность или неопределенность значения каппы [2]. Доверительный интервал — это диапазон значений, содержащий истинный параметр, при этом измеренные значения соответствуют доверительной надежности в 95%.

Коэффициент корреляции между оценкой преподавателя и ChatGPT ($\varkappa$ = 0,059 при SE = 0,095), между оценкой преподавателя и PerplexityAI ($\varkappa$ = 0,041 при SE = 0,094), а также доверительный интервал (колеблющийся от –0,127 до 0,246 в первом случае и от –0,142 до 0,224 — во втором) указывают на крайне малое совпадение между двумя парами проверяющих, и эти значения не является статистически значимыми.

Средний показатель каппа по трем парам составляет 0,105, что находится в пределах диапазона незначительного совпадения. Эти результаты свидетельствуют о различном уровне согласия между преподавателем и двумя чат-ботами, при этом наиболее сильное согласие наблюдается между ChatGPT и PerplexityAI. В данном случае только данные корреляции между чат-ботами показывают статистически значимое совпадение, что вызвано большой долей целых чисел в оценках у обеих программ.

Анализируя оценки по баллам (табл. 3), видим, что чат-боты показали отрицательные значения каппы для оценок 6,5 и 8, что означает низкий уровень согласия с преподавателем при оценке эссе по этим баллам. Чат-боты были умеренно/справедливо согласны с преподавателем в отношении оценок в 5/6 баллов.

Таблица 3

Расчет каппы Флейса

			95% CI	
Баллы	$\varkappa$	SE	Lower	Upper
Общ.	0,077	0,057	–0,035	0,190
5	0,482	0,132	0,222	0,742
5,5	0,123	0,132	–0,137	0,383
6	0,261	0,132	0,001	0,521
6,5	–0,014	0,132	–0,274	0,246
7	0,030	0,132	–0,230	0,290
7,5	0,030	0,132	–0,230	0,290
8	–0,118	0,132	–0,378	0,142

Качественный анализ выдачи чат-ботов

Проводя более детальное сравнение обратной связи, полученной от двух чат-ботов и преподавателя, можно увидеть следующие закономерности и отличия.

В целом ответы двух чат-ботов довольно похожи и демонстрируют лишь незначительные различия в качестве и детальности обратной связи. При этом стоит отметить нестабильность работы и неполное соответствие инструкциям в запросе. Во-первых, один и тот же запрос, введенный в разные дни, может приводить к изменениям в структуре обратной связи. Во-вторых, несмотря на четкое указание приводить примеры ошибок, оба чат-бота могут время от времени игнорировать эту деталь в запросе. При повторном запросе чат-боты могут присвоить другие оценки по каждому критерию, при этом общая оценка может остаться прежней.

Длина текста обратной связи нестабильна, что вполне обосновано отличиями в эссе и количеством ошибок, а также настройками самого чат-бота. Самая короткая обратная связь обычно дается преподавателем. PerplexityAI детальнее аргументирует оценку, чем ChatGPT (табл. 4).

Таблица 4
Примеры формулировки обратной связи чат-ботами по разным аспектам

	ChatGPT	PerplexityAI
Обратная связь по критерию C&C	Your essay is well-organized and coherent, with clear introduction, body paragraphs discussing advantages and disadvantages, and a conclusion summarizing your opinion. You have used cohesive devices effectively to link ideas and maintain the flow of your arguments.	Your essay is well-organized and coherent, with a clear introduction, body paragraphs, and conclusion. You have used appropriate cohe-sive devices and transitional phrases effectively, such as "Re-garding positive effects", "On the downside", and "To conclude". The logical flow of ideas is generally smooth, but there are a few in-stances where the connections be-tween ideas could be more explicit.

Подключение к интернету, имеющееся у PerplexityAI, однозначно имеет преимущество в том плане, что обратная связь по каждому критерию формулируется именно в рамках описаний, представленных официальными источниками экзамена IELTS.

Чат-боты неплохо справляются с нахождением грамматических и лексических ошибок, довольно часто «соглашаются» относительно самой ошибки, однако варианты ее исправления могут разниться. Выделяются несуществующие ошибки нарушения согласования подлежащего и сказуемого или сочетаемости слов. Словосочетание *dwindling interest* чат-бот предлагает заменить более частотным *declining interest*, хотя они являются синонимичными. При этом оба чата могут неверно идентифицировать ошибку по аспекту языка (отнести ошибку в выборе слова к грамматике).

Обсуждение и выводы

Статистический анализ согласия между тремя проверяющими продемонстрировал неоднозначные результаты. Хотя доля оценок, отличавшихся не более чем на балл, была высока и варьировалась от 79 до 100%, доля полностью идентичных оценок была сравнитель-

но мала, и наибольшее согласие прослеживалось у двух чат-ботов. Причина, вероятно, заключается в том, что чат-боты отличаются от преподавателя тенденцией к выставлению целочисленных оценок. Для преподавателя отдельный критерий может иметь характеристики, в равной принадлежащие сразу двум оценкам, и он(а) выставляет промежуточный балл, тогда как чат-боты, как видится, стремятся к однозначности.

Исходя из полученных статистических данных, межэкспертное соответствие между общей оценкой преподавателя и двух чат-ботов очень низкое. Коэффициенты каппы Коэна находятся в диапазоне от 0,041 до 0,216, при среднем значении 0,105, что указывает на плохое соответствие между оценками, присвоенными тремя проверяющими. Доверительные интервалы относительно велики, что предполагает отсутствие статистической значимости результатов. Последующие исследования на большей выборке смогут подтвердить или опровергнуть полученные результаты.

Статистический анализ каппы Флейса также демонстрирует, что преподаватель и два чат-бота не достигли согласия в оценке эссе, наибольшие совпадения наблюдаются относительно баллов 5 и 6.

Умеренное совпадение некоторых оценок и верная интерпретация критериев оценивания чат-ботами говорит о том, что они могут справляться с задачами по оценке эссе и использоваться студентами для грубой самостоятельной оценки своих работ. Чат-бот дает обратную связь не только на язык, но и на аргументацию, что способствует совершенствованию соответствующих умений [25]. Однако необходимо доучивать и совершенствовать чат-боты, прежде чем применять для оценки письменных работ студентов, особенно в условиях тестирования с высокими ставками. Возможность тонкой настройки чат-ботов является их несомненным преимуществом, позволяющим пользователям адаптировать модель для выполнения конкретных задач [22].

Несомненным плюсом является способность чат-бота формулировать обратную связь по заданному шаблону. Однако один и тот же текст запроса не гарантирует унификации результата и четкого следования инструкциям. Может требоваться ведение дальнейшего диалога с чат-ботом для улучшения результата и получения недостающих или дополнительных деталей. Будет необходима доработка полученной выдачи, что является стандартной практикой при использовании генеративных моделей [1].

Хотя языковые модели обладают рядом преимуществ, у них также есть ограничения, касающиеся качества и разнообразия используемых тренировочных данных [13]. В сравнении со специализированными средствами автоматической проверки эссе, обученными для работы с конкретным жанром текстов, языковые модели основываются на большом корпусе данных, представленных текстами разного формата из открытых источников сети Интернет, поэтому могут отличным от эксперта образом интерпретировать качество работы по разным аспектам.

В отличие от систем автоматической проверки или генеративных языковых моделей преподаватели могут учитывать индивидуальные особенности учащихся, такие как уровень владения языком, родной язык, наличие фоновых знаний, пробелы в знаниях и умениях, и подстраивать обратную связь под конкретного студента. Например, преподаватели могут выбрать между прямым или косвенным типом обратной связи в зависимости от целей обучения, приоритетов студентов и самого преподавателя [21]. Помимо этого, преподавателям свойственна гибкость, умение подстраиваться под изменения в жизни или в образовании. Так, при современном темпе жизни тенденции ускорения всех процессов и наличии привычки получения важной информации в максимально сжатом виде и за короткий промежуток времени важно предоставлять студентам обратную связь в читабельном объеме, при этом указывая на важные детали.

Таким образом, чат-боты могут оказать содействие преподавателю иностранного языка в выявлении недочетов в письменных работах студентов и формулировке обратной связи, однако оценка работы в соответствии с критериями и выбор информации, используемой в обратной связи, должен оставаться за специалистом [24]. Студенты могут воспользоваться данными технологиями для получения приблизительной оценки своих работ при самостоятельной подготовке к стандартизированным экзаменам, корректным образом формулируя запрос и осознавая ограничения со стороны чат-бота.

Список источников / References

1. *Боголепова С. В., Бабасян Е. Р.* Возможности искусственного интеллекта для разработки учебных и оценочных заданий по иностранным языкам // Преподаватель XXI век. 2024. № 1. С. 137–154.

2. *Albakkosh I.* Using Fleiss' kappa coefficient to measure the intra and inter- rater reliability of three AI software programs in the assessment of EFL learners' story writing // International Journal of Educational Sciences and Arts. 2024. Vol. 3, no. 1. P. 69–96. doi: 10.59992/ijesa.2024.v3n1p4.

3. *Barrot J. S.* Using ChatGPT for second language writing: Pitfalls and potentials // Assessing Writing. 2023. Vol. 57. doi: 10.1016/j.asw.2023.100745.

4. *Benali A.* The impact of using automated writing feedback in ESL/EFL classroom contexts // English Language Teaching. 2021. Vol. 14, no. 12. P. 189–195. doi: 10.5539/elt.v14n12p189.

5. *Doroudi S.* The intertwined histories of artificial intelligence and education // International Journal of Artificial Intelligence in Education. 2022. Vol. 33, no. 4. P. 885–928. doi: 10.1007/s40593-022-00313-2.

6. *Fraga L.* Testing the reliability of two rubrics used in official English certificates for the assessment of writing // Revista Alicantina de Estudios Ingleses. 2022. No. 36. P. 85–109. doi: 10.14198/raei.2022.36.05.

7. *Giray L.* Prompt Engineering with ChatGPT: A Guide for Academic Writers // Annals of Biomedical Engineering. 2023. Vol. 51, no. 12. P. 2629–2633. doi: 10.1007/s10439-023-03272-4.

8. *González-Carrillo C. D.* Automatic Grading Tool for Jupyter Notebooks in Artificial Intelligence Courses // Sustainability. 2021. Vol. 13, no. 21. doi: 10.3390/su132112050.

9. *Guinness H.* What is Perplexity AI? // Zapier. 2024 [Электронный ресурс]. URL: https://zapier.com/blog/perplexity-ai/ (дата обращения: 15.05.2024).

10. *Kasneci E.* ChatGPT for good? On opportunities and challenges of large language models for education // Learning and Individual Differences. 2023. Vol. 103. doi: 10.1016/j.lindif.2023.102274.

11. *Kohnke L., Moorhouse B. L., Zou D.* ChatGPT for language teaching and learning // RELC Journal. 2023. Vol. 54, no. 2. P. 537–550. doi: 10.1177/00336882231162868.

12. *Landis J. R., Koch G. G.* The measurement of observer agreement for categorical data // Biometrics. 1977. Vol. 33, no. 1. P. 159–174. doi: 10.2307/2529310.

13. *Lee A., Miranda B., Sundar S., et al.* Beyond Scale: the Diversity Coefficient as a Data Quality Metric Demonstrates LLMs are Pre-trained on Formally Diverse Data // ArXiv. 2023 [Электронный ресурс]. URL: https://arxiv.org/abs/2306.13840 (дата обращения: 12.05.2024).

14. *Liu M.* Exploring the Application of Artificial intelligence in Foreign Language Teaching: Challenges and future development // SHS Web of Conferences. 2023. Vol. 168. doi: 10.1051/shsconf/202316803025.

15. *Meniado J. C.* The Impact of ChatGPT on English Language Teaching, Learning, and Assessment: A Rapid Review of Literature // Arab World English Journals (AWEJ). 2023. Vol. 14, no. 4. P. 3–18. doi: 10.24093/awej/vol14no4.1.

16. *Mizumoto A., Eguchi M.* Exploring the potential of using an AI language model for automated essay scoring // Research Methods in Applied Linguistics. 2023. Vol. 2, no. 2. doi: 10.1016/j.rmal.2023.100050.

17. *Mollick E., Mollick L.* Using AI to implement effective teaching strategies in classrooms: Five strategies, including prompts // SSRN Electronic Journal. 2023. doi: 10.2139/ssrn.4391243.

18. *Ndukwe I. G., Daniel B. K., Amadi C. E.* A machine learning grading system using chatbots // Artificial Intelligence in Education. Lecture Notes in Computer Science. 20th International Conference, AIED 2019, Chicago, IL, USA, June 25–29, 2019, Proceedings, Part II. P. 365–368. doi: 10.1007/978-3-030-23207-8_67.

19. *Park D.* A Study on Performance Improvement of Prompt Engineering for Generative AI with a Large Language Model // Journal of Web Engineering. 2024. Vol. 22, no. P. 1187–1206. doi: 10.13052/jwe1540-9589.2285.

20. *Ramesh D., Sanampudi S. K.* Automated essay scoring systems: a systematic literature review // Artificial Intelligence Review. 2021. Vol. 55, no. 3. P. 2495–2527. doi: 10.1007/s10462-021-10068-2.

21. *Ranalli J.* Automated written corrective feedback: how well can students make use of it? // Computer Assisted Language Learning. 2018. Vol. 31, no. 7. P. 653–674. doi: 10.1080/09588221.2018.1428994.

22. *Ray P. P.* ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope // Internet of Things and Cyber-physical Systems. 2023. Vol. 3. P. 121–154. doi: 10.1016/j.iotcps.2023.04.003.

23. *Sidorkin A. M.* Embracing Chatbots in Higher Education: The Use of Artificial Intelligence in Teaching, Administration, and Scholarship. New York: Routledge, 2024. 114 p. doi: 10.4324/9781032686028.

24. *Tekin S., Aydoğdu Ş.* Automated Assessment of Students' Critical Writing Skills with Chatgpt // SSRN. 2024 [Электронный ресурс]. URL: https://ssrn.com/abstract=4826249 (дата обращения: 15.05.2024).

25. *Wang Z.* Computer-assisted EFL writing and evaluations based on artificial intelligence: a case from a college reading and writing course // Library Hi Tech. 2020. Vol. 40, no. 1. P. 80–97. doi: 10.1108/lht-05-2020-0113.

Информация об авторах
С. В. Боголепова — кандидат филологических наук, доцент школы иностранных языков
М. Г. Жаркова — преподаватель школы иностранных языков

Information about the authors
S. V. Bogolepova — PhD in Philology, Associate Professor at School of Foreign Languages
M. G. Zharkova — Master's student at School of Foreign Languages

Статья поступила в редакцию 07.06.2024; одобрена после рецензирования 27.06.2024; принята к публикации 17.09.2024.
The article was submitted 07.06.2024; approved after reviewing 27.06.2024; accepted for publication 17.09.2024.