

ПРОБЛЕМЫ ПРИКЛАДНОЙ ЛИНГВИСТИКИ

УДК 81

Гутовская Марина Степановна

доктор филологических наук, профессор,
заведующий кафедрой
германского языкознания
Белорусский государственный
университет
г. Минск, Беларусь

Marina Gutovskaya

Doctor of Philology, Full Professor,
Head of the Department
of Germanic Linguistics
Belarusian State University
Minsk, Belarus
marina-gutovskaya@yandex.by

Шайбакова Дарья Вадимовна

аспирант кафедры
германского языкознания
Белорусский государственный
университет
г. Минск, Беларусь

Darya Shaibakova

PhD Student at the Department
of Germanic Linguistics
Belarusian State University
Minsk, Belarus
ShaibakovaDV@yandex.by

ОЦЕНКА КАЧЕСТВА СГЕНЕРИРОВАННЫХ НАУЧНЫХ СТАТЕЙ ЭКСПЕРТАМИ И БОЛЬШИМИ ЯЗЫКОВЫМИ МОДЕЛЯМИ¹

Статья посвящена оценке качества научных статей, сгенерированных большими языковыми моделями (БЯМ), в сопоставлении с антропогенными статьями двух жанров (общетеоретического и научно-исследовательского) посредством интегративной лингводискурсивной методики исследования продуктов искусственного текстопорождения. Материалом послужили двенадцать текстов: два антропогенных текста научных статей (общетеоретическая и научно-исследовательская) и тексты, сгенерированные пятью БЯМ (DeepSeek-V4, Gemini 3 Flash, GigaChat 2 Ultra, ChatGPT-5.3 Instant, Alice AI LLM) на базе представленных в антропогенных статьях данных. Оценка проводилась экспертами-лингвистами и БЯМ-оценщиками по двадцати пяти критериям (общезыковым и дискурсно-специфичным) с последующим сопоставлением полученных показателей. Выявлена асимметрия экспертной и машинной оценок, зависимость качества генераций от жанра и архитектуры модели, установлена иерархия изучаемых БЯМ по качеству порождаемых текстов. Показано, что общезыковая правильность является необходимой, но недостаточной основой для создания полноценного академического текста; наиболее уязвимы параметры, требующие эпистемической позиции, – глубина аргументации, доказательность и личный вклад автора. Современные БЯМ успешно имитируют внешние жанровые признаки, однако не достигают содержательной глубины, что ставит вопрос о разработке дискурсно-ориентированных подходов к оценке академических текстов.

¹ Исследование выполнено в рамках научно-исследовательской работы «Возможности больших языковых моделей в современной лингвистике: теоретико-методологические и прикладные решения» (номер государственной регистрации 20262323) в задании «Современные информационные технологии в лингвистических исследованиях» подпрограммы «Языкознание» государственной программы научных исследований «Социально-экономическая, политическая и национально-культурная безопасность белорусской государственности» (2026–2030 гг.).

Ключевые слова: *искусственное текстопорождение; большая языковая модель; экспертная оценка; автоматическая оценка; критерии качества сгенерированного контента.*

QUALITY ASSESSMENT OF GENERATED SCIENTIFIC ARTICLES BY EXPERTS AND LARGE LANGUAGE MODELS

This paper presents a comparative assessment of the quality of scientific articles generated by large language models (LLMs) against human-authored texts representing two genres – general theoretical and research-oriented – using an integrative linguo-discursive methodology for studying artificial text generation products. The research material comprised twelve texts: two human-written scholarly articles (one theoretical and one empirical) and texts generated by five LLMs (DeepSeek-V4, Gemini 3 Flash, GigaChat 2 Ultra, ChatGPT-5.3 Instant, and Alice AI LLM) based on the data provided in the human-authored sources. Evaluation was carried out by expert linguists and by LLM-as-a-judge systems across twenty-five criteria encompassing both general linguistic and discourse-specific parameters, followed by a comparative analysis of the obtained scores. The study reveals a significant asymmetry between human and machine assessments, demonstrates the dependence of generation quality on genre and model architecture, and establishes a hierarchy of the examined LLMs in terms of the quality of the texts they produce. It is shown that general linguistic correctness is a necessary but insufficient foundation for producing a fully-fledged academic text; the most vulnerable parameters are those requiring an epistemic stance – depth of argumentation, evidential substantiation, and authorial contribution. Modern LLMs successfully emulate external genre markers, yet fall short of substantive depth, thereby raising the question of developing discourse-oriented assessment approaches for academic texts.

Key words: *artificial text generation; large language model; expert evaluation; automated evaluation; quality criteria for generated content.*

Расширение возможностей больших языковых моделей (БЯМ) в задаче генерации связного, стилистически выдержанного текста ставит перед лингвистикой вопрос, выходящий за пределы технической оценки производительности систем искусственного интеллекта (ИИ): насколько искусственно созданный текст (ИСТ) соответствует требованиям, предъявляемым к тексту определенного жанра, и какова мера его качественного тождества или расхождения с естественно созданным текстом (ЕСТ). По данным исследования, охватившего более 1,1 миллиона рецензируемых публикаций на платформах arXiv, bioRxiv и в журналах Nature за 2020–2024 годы, доля текстов, демонстрирующих признаки участия БЯМ в их подготовке, достигает 22 % в области компьютерных наук и 9 % в работах по математике, при этом зафиксирована устойчивая положительная динамика данного показателя [1]. На этом фоне особую остроту приобретает проблема качества ИСТ именно в научном дискурсе, который требует содержательной глубины, аргументированности и авторской ответственности за истинность утверждений.

Существенное затруднение при оценке качества ИСТ создает расхождение между автоматическими метриками и лингвистической экспертизой. Традиционные количественные показатели (BLEU, ROUGE, BERTScore) фиксируют

поверхностные статистические закономерности и слабо коррелируют с оценками квалифицированных читателей [2]. Экспертная же оценка, как правило, носит интуитивный характер, а согласованность работы рецензентов при анализе дискурсно-стилистических характеристик ИСТ остается низкой [3]. Стремление совместить автоматические решения с гибкостью человеческого суждения привело к распространению метода БЯМ-как-судья (LLM-as-a-Judge), в котором языковая модель выступает оценщиком качества ИСТ [4; 5]. Однако достоверность подобной практики применительно к текстам высокой жанровой сложности, к которым относится научная статья, остается предметом дискуссии и требует эмпирической проверки, сопоставимой по строгости с экспертной лингвистической оценкой.

Решению этой задачи отвечает интегративная лингводискурсивная методика исследования продуктов искусственного текстопорождения, предполагающая на одном из этапов реализации многокритериальную оценку ИСТ по двум классам параметров – общезыковым (лексико-семантическим и грамматическим) и дискурсно-специфичным (ориентированным на жанровую специфику научного текста) [6]. К оценке представлены 12 текстов: 2 ЕСТ и 10 ИСТ. Антропогенную основу корпуса составили общетеоретическая статья Т. Ф. Плехановой «От лингвистики текста к лингвистике дискурса» и научно-исследовательская статья М. С. Гутовской и Д. В. Шайбаковой «Организация образных представлений о знании в русской и английской фразеологии: концептуальные метафоры и метафорические модели». На базе двух ЕСТ с помощью специализированных промптов каждая из пяти исследуемых БЯМ – DeepSeek-V4, Gemini 3 Flash, GigaChat 2 Ultra, ChatGPT-5.3 Instant, Alice AI LLM – сгенерировала по одной имитации.

К оценке были привлечены квалифицированные эксперты-лингвисты (Е. С. Астапкина – кандидат филологических наук, доцент кафедры германского языкознания Белорусского государственного университета; Л. М. Длюбарская – преподаватель кафедры английского языка Лицея им. Ф. Э. Дзержинского Белорусского государственного университета; Т. В. Никитенко – кандидат филологических наук, заведующий кафедрой теории и практики перевода Белорусского государственного университета) и эффективные БЯМ (Grok 4.1, Qwen2.5-Max, Claude Sonnet 4.6), выступающие в роли автоматизированных судей. В целях организации слепого эксперимента тексты исследовательского корпуса представлялись на оценку с кодовой нумерацией: 07 и 11 – антропогенные статьи; 01, 03, 04, 06, 12 – генерации на основе общетеоретической статьи; 02, 05, 08, 09, 10 – генерации на основе научно-исследовательской статьи. Такой алгоритм позволил получить количественно сопоставимые показатели качества генерации, выявить характер расхождений между человеческим и машинным восприятием текста и установить, по каким параметрам академического письма современные БЯМ достигают нормы, а по каким – систематически от нее отклоняются.

Обсуждение

Прежде чем обратиться к покритериальному анализу, рассмотрим обобщенные показатели, характеризующие каждый текст в совокупности. Для каждого текста вычислены: средняя экспертная оценка ($E_{average}$), средняя оценка БЯМ-оценщиков ($L_{average}$), общая средняя оценка текста ($O_{average}$), абсолютная разница между двумя видами оценки ($D_{average}$) и относительная разница в процентном выражении ($D_{\%}$). Каждый из 25 критериев оценивался экспертами и БЯМ-оценщиками по пятибалльной шкале, максимально возможный итоговый балл составил 125 (25 критериев $\times$ 5 баллов). Показатель $E_{average}$ вычислялся как среднее арифметическое суммарных баллов трех экспертов-лингвистов; $L_{average}$ – как среднее арифметическое суммарных баллов трех БЯМ-оценщиков; $O_{average}$ определялся как среднее арифметическое между средней экспертной оценкой и средней оценкой БЯМ-оценщиков; $D_{average}$ – как разность между средней экспертной оценкой и средней оценкой БЯМ-оценщиков; $D_{\%}$ – как отношение абсолютной разницы к общей средней оценке текста, выраженное в процентах. Высококачественными текстами считались те, показатель $O_{average}$ которых составил ≥ 100 баллов; удовлетворительными – $O_{average}$ от 75 до 99 баллов; низкокачественными – $O_{average} < 75$ баллов. Границы уровней качества определены путем перевода балла $O_{average}$ в процентную шкалу с шагом в 20 %: значению 125 баллов соответствуют максимальные 100 %, 100 баллов – 80 % от максимума, 75 баллов – 60 %. Такое деление (≥ 80 % – высококачественный, 60–79 % – удовлетворительный, < 60 % – низкокачественный текст) соответствует широко используемой логике построения уровней шкал оценивания [7]. Показатели текстов приводятся с указанием категории качества: В – высококачественный, У – удовлетворительный, Н – низкокачественный (табл. 1).

Таблица 1

Итоговые показатели оценки текстов исследовательского подкорпуса

Код	Текст / источник	$E_{average}$	$L_{average}$	$O_{average}$	$D_{average}$	$D_{\%}$	Категория
01	ChatGPT-5.3 Instant / статья Т. Ф. Плехановой	93,0	93,3	93,2	-0,3	0,32	У
02	DeepSeek-V4 / статья М. С. Гутовской, Д. В. Шайбаковой	122,7	105,7	114,2	+17,0	14,9	В
03	Gemini 3 Flash / статья Т. Ф. Плехановой	119,3	99,0	109,2	+20,3	18,6	В
04	Alice AI LLM / статья Т. Ф. Плехановой	86,3	91,0	88,6	-4,7	5,3	У

Окончание таблицы 1

05	Alice AI LLM / статья М. С. Гутовской, Д. В. Шайбаковой	85,3	82,0	83,7	+3,3	3,9	У
06	DeepSeek-V4 / статья Т. Ф. Плехановой	111,3	108,3	109,8	+3,0	2,7	В
07	ECT – Т. Ф. Плеханова	101,3	108,0	104,7	–6,7	6,4	В
08	ChatGPT-5.3 Instant / статья М. С. Гутовской, Д. В. Шайбаковой	86,3	91,0	88,7	–4,7	5,3	У
09	GigaChat 2 Ultra / статья М. С. Гутовской, Д. В. Шайбаковой	113,3	89,0	101,2	+24,3	24,0	В
10	Gemini 3 Flash / статья М. С. Гутовской, Д. В. Шайбаковой	121,7	101,7	111,7	+20,0	17,9	В
11	ECT – М. С. Гутовская, Д. В. Шайбакова	125,0	113,3	119,2	+11,7	9,8	В
12	GigaChat 2 Ultra / статья Т. Ф. Плехановой	104,0	90,0	97,0	+14,0	14,4	У

Ни один из двенадцати текстов не получил квалификации «низкокачественный». Наименьший показатель $O_{average}$ зафиксирован для текста с кодом 05 (Alice AI LLM на научно-исследовательскую статью М. С. Гутовской, Д. В. Шайбаковой) – 83,7 балла, – что соответствует нижней половине удовлетворительного диапазона. То есть рассмотренные БЯМ в той или иной мере освоили структурно-стилистическую специфику научного дискурса и способны достаточно правдоподобно воспроизводить тексты теоретической и исследовательской направленности. Из двенадцати текстов семь отнесены к категории высококачественных. В порядке убывания $O_{average}$:

1. Статья М. С. Гутовской, Д. В. Шайбаковой (119,2 балла);
 2. DeepSeek-V4 / статья М. С. Гутовской, Д. В. Шайбаковой (114,2 балла);
 3. Gemini 3 Flash / статья М. С. Гутовской, Д. В. Шайбаковой (111,7 балла);
 4. DeepSeek-V4 / статья Т. Ф. Плехановой (109,8 балла);
 5. Gemini 3 Flash / статья Т. Ф. Плехановой (109,2 балла);
 6. ECT – статья Т. Ф. Плехановой (104,7 балла);
 7. GigaChat 2 Ultra / статья М. С. Гутовской, Д. В. Шайбаковой (101,2 балла).
- Пять текстов охарактеризованы как удовлетворительные:
8. GigaChat 2 Ultra / статья Т. Ф. Плехановой (97,0 баллов);

9. ChatGPT-5.3 Instant / статья Т. Ф. Плехановой (93,2 балла);
10. ChatGPT-5.3 Instant / статья М. С. Гутовской, Д. В. Шайбаковой (88,7 баллов);
11. Alice AI LLM / статья Т. Ф. Плехановой (88,6 баллов);
12. Alice AI LLM / статья М. С. Гутовской, Д. В. Шайбаковой (83,7 балла).

В числе высококачественных текстов оказались пять ИСТ. Два ИСТ – DeepSeek-V4 (код 06, $O_{average} = 109,8$) и Gemini 3 Flash (код 03, $O_{average} = 109,2$) на материале статьи Т. Ф. Плехановой – по общей оценке превосходили соответствующий ЕСТ-эталон (код 07, $O_{average} = 104,7$). Данное наблюдение, вероятнее всего, объясняется ориентированностью ИСТ на полное соответствие стандартам академического письма, тогда как ЕСТ Т. Ф. Плехановой – произведение с выраженной авторской позицией, которая не всегда совпадает со строгими конвенциями жанра. ЕСТ научно-исследовательской статьи не превосшла ни одна из БЯМ: разрыв между лучшим ИСТ (код 02, DeepSeek-V4, $O_{average} = 114,2$) и ЕСТ (код 11, $O_{average} = 119,2$) составляет 5 баллов.

Среднее $O_{average}$ пяти ИСТ на базе научно-исследовательской статьи равно 99,9 балла – на 19,3 балла ниже ЕСТ. Для общетеоретической статьи разрыв между средним $O_{average}$ ИСТ (99,56) и ЕСТ (104,7) существенно меньше – 5,14 балла. Превосходство антропогенного авторства сохраняется в наибольшей степени применительно к исследовательским текстам, а при создании теоретических статей БЯМ демонстрируют результаты, более сопоставимые с уровнем профессионального ученого. Данная закономерность коррелирует с природой знания, воплощенного в каждом из двух рассматриваемых жанров. Обзорно-теоретическая статья в значительной мере опирается на обобщение и анализ существующих научных публикаций, с чем ИИ-системы справляются весьма успешно. Исследовательская же статья требует аналитической интерпретации оригинальных эмпирических данных и демонстрации авторского вклада – того, что современные БЯМ воспроизводят менее качественно.

Обратимся к показателям $E_{average}$, $L_{average}$, $D_{average}$ и $D\%$, чтобы выявить характер расхождений между двумя типами оценщиков. Согласно принятому в методике критерию ($|D\%| > 10\%$), значимые расхождения между оценками экспертов и БЯМ выявлены для пяти текстов (в порядке убывания $|D\%|$):

1. GigaChat 2 Ultra / статья М. С. Гутовской, Д. В. Шайбаковой, $|D\%| = 24,0\%$;
2. Gemini 3 Flash / статья Т. Ф. Плехановой, $|D\%| = 18,6\%$;
3. Gemini 3 Flash / статья М. С. Гутовской, Д. В. Шайбаковой, $|D\%| = 17,9\%$;
4. DeepSeek-V4 / статья М. С. Гутовской, Д. В. Шайбаковой, $|D\%| = 14,9\%$;
5. GigaChat 2 Ultra / статья Т. Ф. Плехановой, $|D\%| = 14,4\%$.

Во всех пяти случаях $E_{average} > L_{average}$. Это обнаруживает закономерность: БЯМ более взыскательны, нежели эксперты, при оценке наиболее успешных ИСТ (4 из 5 обозначенных текстов ранее отнесены к категории высококачественных).

В список дважды попали генерации GigaChat 2 Ultra и Gemini 3 Flash, т. е. данные модели способны создавать тексты, которые эксперты-лингвисты склонны оценивать существенно выше, чем БЯМ. Это позволяет предположить, что данные модели создают тексты высокого качества, нюансы которых лучше улавливаются при автоматическом анализе, но менее заметны лингвистам. DeepSeek-V4 попала в перечень однократно и с небольшим расхождением ($|D\%| = 14,9\%$), поэтому аналогичный вывод для нее неправомерен. Генерации Alice AI LLM и ChatGPT-5.3 Instant значимых расхождений не показали. Вероятно, они содержат достаточное количество артефактов, которые одинаково заметны обеим категориям оценщиков. Умеренное (среднее) качество текста порождает согласие, а высокое качество, достигнутое некоторыми БЯМ, порождает разногласие между типами оценщиков.

Для ЕСТ М. С. Гутовской и Д. В. Шайбаковой $D\%$ составляет 9,8 %, приближаясь к десятипроцентному порогу значимости расхождений. Эксперты присвоили тексту максимальный балл ($E_{average} = 125$), тогда как средняя оценка БЯМ составила 113,3. Данное соотношение указывает на предел восприятия БЯМ-оценщиков. Они избегают слишком высоких оценок, удерживая диапазон 82–108 баллов (табл. 1), что соответствует оценкам «хорошо» и «очень хорошо», но никогда – «превосходно». Показатель $L_{average} = 113,3$ – наиболее высокий из всех оцененных текстов. Ни один научный текст не оценивается БЯМ на высший балл в целях избегания риска ошибочного завышения, которое могло быть истолковано как некомпетентность.

Среди пяти БЯМ выявлена иерархия по показателю $O_{average}$, воспроизводимая применительно к генерациям на оба ЕСТ (табл. 2).

Т а б л и ц а 2

Ранжирование БЯМ по показателю $O_{average}$

Модель	O_{avg} (статья Т. Ф. Плехановой)	O_{avg} (статья М. С. Гутовской, Д. В. Шайбаковой)	Среднее O_{avg}	% от макс.	Ранг
DeepSeek-V4	109,8	114,2	112,0	89,6 %	1
Gemini 3 Flash	109,2	111,7	110,45	88,4 %	2
GigaChat 2 Ultra	97,0	101,2	99,1	79,3 %	3
ChatGPT-5.3 Instant	93,2	88,7	90,95	72,8 %	4
Alice AI LLM	88,6	83,7	86,15	68,9 %	5

DeepSeek-V4 занимает первое место со средним $O_{average} = 112$ (89,6 %), минимально опережая Gemini 3 Flash – 110,45 (88,4 %). Данные БЯМ практически эквиваленты по эффективности при решении задачи порождения академических текстов, стабильно порождают высококачественные тексты

независимо от типа исходной статьи. Обе БЯМ применяют расширенные режимы обработки информации – «Глубокое мышление» (DeepSeek-V4) и «Deep research» в сочетании с «Думающей» моделью (Gemini 3 Flash), – что может свидетельствовать о значимой роли многошагового рассуждения БЯМ при создании академического текста.

GigaChat 2 Ultra занимает третью позицию ($O_{average} = 99,1; 79,3 \%$). Разрыв между GigaChat 2 Ultra и лидирующей парой составляет около 12–13 баллов, что существенно отделяет качество генераций DeepSeek-V4 и Gemini 3 Flash от следующих в иерархии БЯМ. Четвертую и пятую позиции занимают ChatGPT-5.3 Instant ($O_{average} = 90,95; 72,8 \%$) и Alice AI LLM ($O_{average} = 86,15; 68,9 \%$). Оба этих показателя соответствуют удовлетворительному качеству. Обе БЯМ используют режим «Рассуждать», что указывает на зависимость между глубиной механизмов рассуждения при генерации и итоговым качеством текста. Вероятной причиной отставания двух БЯМ может быть относительно небольшая величина их контекстного окна и ориентация на «живое» общение с пользователем в форме чата, коротких вопросов и ответов. Перечисленные факторы могли значительно повлиять на длину и качество выдач данных БЯМ в жанре научной статьи.

При сопоставлении средних значений межмодельная вариативность существенно различается в зависимости от типа исходного текста. Для общетеоретической статьи диапазон $O_{average}$ составляет $109,8 - 88,6 = 21,2$ балла; для научно-исследовательской статьи – $114,2 - 83,7 = 30,5$ балла. Ведущие системы (DeepSeek-V4, Gemini 3 Flash) при работе с исследовательской статьей демонстрируют высокие результаты, тогда как системы следующих уровней (Alice AI LLM, ChatGPT-5.3 Instant) проявляют более выраженное отставание, нежели при работе с теоретическим текстом. Тем не менее, ни в одном случае не зафиксировано изменения позиций при переходе от одного объекта генерации (общетеоретической статьи) к другому (научно-исследовательской статье). DeepSeek-V4 и Gemini 3 Flash неизменно превосходят GigaChat 2 Ultra, который, в свою очередь, неизменно превосходит ChatGPT-5.3 Instant и Alice AI LLM.

Для обнаружения особенностей ИСТ, которые являются преимуществами или, наоборот, недостатками генераций БЯМ, вычислены показатели E_{kAv} (средняя оценка всех экспертов по одному критерию) и L_{kAv} (средняя оценка всех БЯМ по одному критерию) для десяти ИСТ корпуса. Критерии, по которым осуществлялась оценка, ранжировано представлены в табл. 3 и зашифрованы с помощью обозначений O (общезыковые) и D (дискурсно-специфичные), порядковых номеров самих критериев и краткого описания содержания критерия. Критерии ранжированы по показателю E_{kAv} , определенному методикой как диагностический показатель характерных слабостей искусственного дискурса, доступных человеческому восприятию (табл. 3).

Т а б л и ц а 3

Ранжированный список критериев по среднему значению E_{kAv}
(по совокупности 10 ИСТ)

Ранг	Критерий	E_{kAv}	L_{kAv}	Δk
1	Д10 – корректность научных цитат	2,96	3,24	-0,28
2	Д13 – личный вклад автора	3,38	2,47	+0,91
3	Д3 – оформление (таблицы, ссылки)	3,54	3,30	+0,24
4	Д15 – глубина раскрытия темы	3,63	3,18	+0,45
5	Д2 – информативность, доказательность	3,73	2,97	+0,76
6	Д11 – фактическая точность	3,93	3,30	+0,63
7	О5 – отсутствие однотипных идей	3,96	3,04	+0,92
8	О3 – лексическое разнообразие	3,97	2,52	+1,45
9	О6 – синтаксическое разнообразие	4,00	2,85	+1,15
10	Д7 – коммуникативно-стилистическое соответствие	4,06	3,91	+0,15
11	О4 – терминологическая точность	4,23	4,32	-0,09
12	Д9 – отсутствие логических ошибок	4,27	3,54	+0,73
13	О7 – семантическая точность	4,30	3,90	+0,40
14	Д14 – отсутствие второстепенной информации	4,33	4,16	+0,43
15	О1 – отсутствие лексических повторов	4,34	3,35	+0,99
16	О2 – отсутствие тавтологии	4,34	3,68	+0,66
17	Д5 – соответствие научному стилю	4,34	4,19	+0,15
18	Д1 – научность темы	4,36	4,38	-0,02
19	Д6 – стилистическая чистота	4,37	4,33	+0,04
20	Д4 – композиционная структура	4,38	4,65	-0,27
21	Д12 – членение на абзацы	4,43	4,35	+0,08
22	О8 – грамматическая правильность	4,49	4,65	-0,16
23	Д8 – объективность изложения	4,64	4,51	+0,13
24	О9 – орфографическая грамотность	4,71	4,68	+0,03
25	О10 – пунктуационная грамотность	4,77	4,29	+0,48

Полученное распределение двадцати пяти позиций допускает выделение четырех качественно различных зон.

Зона критического дефицита ($E_{kAv} < 3,5$) включает два критерия: корректность научного цитирования (Д10) и демонстрация личного вклада автора (Д13). Это те аспекты, которые наиболее последовательно отмечаются лингвистами-оценщиками как недостатки представленных к анализу текстов. Зона повышенной уязвимости ($3,5 \leq E_{kAv} < 4,0$) включает шесть критериев – внешнее оформление научного аппарата (Д3), глубину раскрытия темы (Д15), информативность и доказательность (Д2), фактическую точность (Д11), отсутствие однотипных идей (О5) и лексическое разнообразие (О3). Эти аспекты отмечаются среди недостатков ИИ менее закономерно, и их появление в конкретной генерации зависит скорее от архитектурных особенностей БЯМ и сложности задания. Из упомянутых восьми позиций шесть принадлежат дискурсно-специфичной группе критериев и две – общеязыковой. ИСТ в большей мере страдают от неспособности соответствовать специфике научного дискурса, а не от некомпетентности в области владения языком.

Зона функционального освоения ($4,0 \leq E_{kAv} < 4,5$) охватывает тринадцать критериев. Восемь из тринадцати попавших сюда параметров относятся к дискурсно-специфичной группе: коммуникативно-стилистическое соответствие (Д7), отсутствие логических ошибок (Д9) и второстепенной информации (Д14), соответствие научному стилю (Д5), научность темы (Д1), стилистическая чистота (Д6), композиционная структура (Д4) и логичное членение на абзацы (Д12). Оставшиеся пять критериев зоны относятся к общеязыковой группе: синтаксическое разнообразие (О6), терминологическая точность (О4), семантическая точность (О7), отсутствие лексических повторов (О1) и тавтологии (О2). Таким образом, БЯМ способны удерживать научную тематику, выдерживать стилистику, композиционно организовывать текст и избегать большинства грубых логических сбоев. Однако их лексическая и синтаксическая вариативность оценивается человеком умеренно и при этом вызывает повышенную критику со стороны БЯМ-оценщиков (Δk по О3, О6, О1, О5).

Зона уверенного освоения со значением $E_{kAv} \geq 4,5$ включает четыре критерия – отсутствие грамматических ошибок (О8), объективность и абстрагированность изложения (Д8), орфографическую (О9) и пунктуационную (О10) грамотность. Эти параметры, в большинстве своем общеязыковые, освоены БЯМ практически безупречно с точки зрения человека-эксперта, и именно по ним между человеческой и машинной оценками достигается максимальное согласие. По орфографии (О9) расхождение Δk составляет всего +0,03, по объективности (Д8) +0,13, по грамматике (О8) оно и вовсе отрицательно (-0,16), а значит, машинная оценка здесь даже несколько выше экспертной. Более заметный разрыв по пунктуации (+0,48) все равно остается далеким от значений, характерных для лексико-синтаксических критериев. Такая согласованность свидетельствует о том, что языковая и стилевая пра-

вильность – те качества, где способности БЯМ и ожидания академического сообщества практически совпадают. Однако это не следует интерпретировать как полноценное владение академическим дискурсом. Это прежде всего способность БЯМ воспроизводить поверхностные особенности текста, сформированные в результате обучения на массивах выверенных письменных источников.

Столбец Δk фиксирует в количественных показателях разницу между экспертной и машинной строгостью оценки. Преобладание положительных значений Δk (18 из 25) подтверждает большую строгость БЯМ-оценщиков и их особую чувствительность к тем параметрам текста, которые обусловлены статистической природой порождения. Максимальное положительное расхождение зафиксировано для критерия ОЗ «лексическое разнообразие» ($\Delta k = +1,45$), за ним следует Об «синтаксическое разнообразие» ($\Delta k = +1,15$) и О5 «отсутствие однотипных идей» ($\Delta k = +0,92$). Все три показателя принадлежат к группе общеязыковых и отражают вариативность речевого оформления. Тогда как человек-эксперт помещает эти же критерии в нижнюю часть зоны функционального освоения (E_{kAv} от 3,96 до 4,00), БЯМ-оценщик проявляет здесь экстремальную требовательность. БЯМ обладает большей восприимчивостью к повторяемости n-грамм, однообразию синтаксических конструкций и повторяемости одной и той же информации, присущих ее собственным генерациям. Эксперт может не замечать или не воспринимать данные аспекты как критический недостаток.

Отрицательные значения Δk , указывающие на большую снисходительность БЯМ по сравнению с человеком, образуют небольшую группу критериев, связанных с формальной нормативностью и фундаментальными параметрами научного текста. Наиболее показателен здесь критерий Д10 «корректность научных цитат» ($\Delta k = -0,28$), который идет первым в зоне критического дефицита по вердикту экспертов ($E_{kAv} = 2,96$), но получает от БЯМ более высокую оценку ($L_{kAv} = 3,24$). Данный феномен указывает на известную уязвимость БЯМ – неспособность верифицировать фактическое существование источника и релевантность цитируемого фрагмента. При проверке БЯМ оценивают скорее формальное правдоподобие библиографической записи и ее интеграцию в текст, но не ее истинность, тогда как эксперт фокусируется на обратном.

Помимо анализа двух рядов оценок (экспертной и машинной) методологически оправдан и анализ их объединения. Для каждого текста по двадцати пяти критериям вычислен сводный покритеиальный показатель как среднее арифметическое между E_{kAv} и L_{kAv} . Табл. 4 демонстрирует усредненные покритеиальные показатели 10 ИСТ исследовательского корпуса по общеязыковым и дискурсно-специфичным критериям, расположенным в порядке возрастания сводного балла с принятыми обозначениями – О (общеязыковые критерии) и Д (дискурсно-специфичные).

Ранжированный список 25 критериев по сводному показателю

Ранг	Критерий	Сводный балл
1	Д13 – личный вклад автора	2,94
2	Д10 – корректность научных цитат	3,13
3	О3 – лексическое разнообразие	3,26
4	Д2 – информативность, доказательность	3,38
5	Д15 – глубина раскрытия темы	3,43
6	О6 – синтаксическое разнообразие	3,45
7	Д3 – оформление научного аппарата	3,46
8	О5 – отсутствие однотипных идей	3,52
9	Д11 – фактическая точность	3,64
10	О1 – отсутствие лексических повторов	3,86
11	Д9 – отсутствие логических ошибок	3,93
12	Д7 – коммуникативно-стилистическое соответствие	4,00
13	О2 – отсутствие тавтологии	4,04
14	О7 – семантическая точность	4,12
15	Д14 – отсутствие второстепенной информации	4,26
16	Д5 – соответствие научному стилю	4,29
17	О4 – терминологическая точность	4,30
18	Д6 – стилистическая чистота	4,38
19	Д1 – научность темы	4,40
20	Д12 – членение на абзацы	4,42
21	Д4 – композиционная структура	4,53
22	О10 – пунктуационная грамотность	4,57
23	О8 – грамматическая правильность	4,59
24	Д8 – объективность изложения	4,60
25	О9 – орфографическая грамотность	4,72

Наименее успешным для соблюдения критерием становится личный вклад автора (Д13), который даже с учетом более высокой экспертной оценки (3,38) под влиянием оценки БЯМ ($\Delta k = +0,91$) становится самым проблемным аспектом генераций. Следом идет корректность научных цитат (Д10) – единственный критерий, для которого отрицательное расхождение не смогло вывести его из зоны дефицита, а наоборот, подтвердило неспособность модели распознавать фиктивные ссылки. В этой же группе оказываются информативность и доказательность (Д2), глубина раскрытия темы (Д15)

и оформление научного аппарата (Д3), т. е. практически вся архитектура научной аргументации. На противоположном полюсе с максимальным сводным баллом (4,72) находится орфографическая грамотность (О9), объективность изложения (Д8), грамматическая правильность (О8) и пунктуационная грамотность (О10). Этим подтверждается наблюдение о том, что ИСТ относительно успешно имитирует внешние признаки научного стиля, но не совсем успешно справляется с аспектами, которые требуют субъектности и содержательной глубины.

Нижняя зона критического дефицита ($< 3,5$ балла) сильно расширяется – с двух критериев по E_{kAv} до семи по сводному показателю (к Д10 и Д13 присоединяются О3, Д2, Д15, О6, Д3), т. е. равновзвешенное участие машинной оценки более чем втрое увеличивает число дефицитарных параметров. Однако верхняя зона уверенного освоения ($\geq 4,5$ балла), вопреки логике общего понижения значений при подключении более строгого оценщика, не сокращается, а, напротив, расширяется с трех позиций (Д8, О9, О10) до пяти: к ним присоединяются композиционная структура (Д4) и отсутствие грамматических ошибок (О8) – два критерия, для которых Δk отрицателен и для которых исходная экспертная оценка была изначально близка к порогу (4,38 и 4,49 соответственно).

Промежуточное положение зоны функционального освоения (сводные баллы от 4,00 до 4,49) также претерпевает качественную трансформацию. Если в экспертной оценке эта зона включала 13 критериев, в том числе многие общеязыковые показатели (О1, О2, О3, О5, О6), то теперь она сокращается до девяти. В функциональной зоне остаются преимущественно дискурсно-специфичные критерии: коммуникативно-стилистическое соответствие (Д7), отсутствие второстепенной информации (Д14), соответствие научному стилю (Д5), стилистическая чистота (Д6), научность темы (Д1) и членение на абзацы (Д12). Иными словами, жанровая адекватность на среднем уровне признается удовлетворительной обеими сторонами, но не до конца освоенной. Из общеязыковых параметров здесь удерживаются лишь те, по которым наблюдалось либо минимальное расхождение (терминологическая точность, семантическая точность), либо по которым машинная строгость не была экстремальной (отсутствие тавтологии).

Средний сводный балл по всей совокупности критериев составляет около 3,97, что ощутимо ниже среднего экспертного ($\approx 4,14$) и выше среднего машинного ($\approx 3,80$). Там, где качество текста определяется алгоритмически верифицируемыми правилами (орфография, грамматика, композиция), сводная оценка максимальна. В измерениях, требующих интенциональности, ответственности за истинность высказывания, содержательной новизны и индивидуального авторского голоса, она опускается до критических значений.

Покритериальный анализ показал, что наиболее резкие ранговые перестановки при переходе от экспертной оценки к сводному показателю кон-

центрируются в общеязыковой группе критериев. Для выявления закономерностей, определяющих различие в восприятии языковой и дискурсной нормы, необходимым становится обобщение данных на уровне критериальных групп. Сопоставление средних значений по совокупности десяти анализируемых в исследовании ИСТ для двух классов критериев – общеязыковых (О1–О10) и дискурсно-специфичных (Д1–Д15) – представлено в табл. 5.

Т а б л и ц а 5

Сравнение групп критериев по совокупности 10 ИСТ

Группа	E_{kAv}	L_{kAv}	Δk
Общеязыковые (О1–О10)	4,311	3,728	+0,583
Дискурсно-специфичные (Д1–Д15)	4,023	3,765	+0,258

С точки зрения экспертов, общеязыковая компетентность ИСТ оценивается значительно выше, нежели дискурсно-специфичная: средняя экспертная оценка для группы общеязыковых критериев составляет 4,311 против 4,023 для дискурсно-специфичных, т. е. разрыв достигает 0,288 балла. В восприятии человека-эксперта основные проблемы академического ИСТ сосредоточены прежде всего в области следования жанрово-дискурсивной специфике, а владение языковой нормой мыслится как более освоенная область. Иначе распределяются приоритеты среди БЯМ-оценщиков. Для них средняя оценка обеих групп практически равна: L_{kAv} для общеязыковых критериев (3,728) оказывается немного ниже, чем для дискурсно-специфичных (3,765). Ключом к пониманию этого расхождения служит средняя разность оценок Δk . Для общеязыковой группы она составляет +0,583, что более чем вдвое превышает аналогичный показатель для дискурсно-специфичной группы (+0,258). Это следствие того, что среднее значение Δk для общеязыковой группы (+0,583) более чем вдвое превышает среднее значение для дискурсно-специфичной группы (+0,258). БЯМ-оценщики требовательны к лексико-синтаксической текстуре текста, тогда как в оценке жанрово-дискурсивной адекватности их позиция значительно ближе к экспертной.

Данная закономерность имеет когнитивно-технологическую интерпретацию. Большая языковая модель, будучи по своей природе статистическим генератором, обладает чувствительностью к закономерностям лексической сочетаемости, синтаксического варьирования и распределения n-грамм. Оценивая чужую генерацию, БЯМ в значительной степени проецирует на нее те критерии, которые усвоены ею в ходе обучения на текстовых массивах, и потому фиксирует малейшие отклонения от ожидаемого языкового разнообразия как существенный дефект. Напротив, распознавание дискурсно-жанровой адекватности требует не столько статистической, сколько содер-

жательно-нормативной компетенции – способности соотнести текст с эталонными текстами жанра. В этой области способности БЯМ-оценщика и человека-эксперта, как показывают средние значения, сближаются, однако природа этой близости амбивалентна: модель не столько глубже понимает дискурс, сколько проявляет меньшую чувствительность к его тонким нарушениям, которые эксперт распознает точнее.

Усреднение оценок по всей совокупности текстов несколько нивелирует индивидуальные различия между БЯМ, каждая из которых обладает набором сильных и слабых сторон. В связи с этим выполнено усреднение экспертных оценок (E_{kAv}) по двум генерациям для каждой из пяти моделей. Результаты этого усреднения сведены в табл. 6, где для каждой модели указаны средний балл по общеязыковой группе критериев (О-группа), средний балл по дискурсно-специфичной группе (Д-группа) и разрыв между ними (О–Д).

Т а б л и ц а 6

Профили БЯМ по группам критериев (E_{kAv})

Модель	О-группа	Д-группа	Разрыв О–Д
Gemini 3 Flash	4,82	4,81	0,01
DeepSeek-V4	4,73	4,65	0,08
GigaChat 2 Ultra	4,64	4,17	0,47
ChatGPT-5.3 Instant	3,52	3,36	0,16
Alice AI LLM	3,85	3,13	0,72

Оценки Gemini 3 Flash и DeepSeek-V4 находятся в зоне уверенного и высокого освоения по обеим группам критериев. Для Gemini 3 Flash разрыв О–Д составляет всего 0,01 балла, что свидетельствует о практически идеальной сбалансированности языковой и дискурсно-жанровой компетенций в восприятии экспертов. DeepSeek-V4 демонстрирует высокий уровень (4,73 и 4,65 соответственно) с более заметным, но все еще минимальным разрывом в 0,08 балла. Эти БЯМ проявили себя как системы, владеющие лексико-синтаксическими ресурсами языка и способные соблюдать правила научного дискурса.

Оценки ChatGPT-5.3 Instant (3,52 по О-группе и 3,36 по Д-группе) располагаются на границе зоны повышенной уязвимости, а разрыв О–Д в 0,16 балла фиксирует незначительную склонность к более низкой дискурсной компетентности. ChatGPT-5.3 Instant в данном случае не достигает порога зрелого академического текstopорождения ни в языковом, ни в жанровом отношении.

GigaChat 2 Ultra получает высокую оценку по общеязыковым критериям (4,64), однако ее дискурсно-специфичная оценка падает до 4,17, что фор-

мирует разрыв в 0,47 балла. БЯМ владеет нормами языка, но уступает в глубине раскрытия темы, информативности, корректности оформления научного аппарата и параметрах, связанных с прагматикой академического жанра. Такой набор черт указывает на то, что сильная БЯМ сама по себе не гарантирует адекватности порождаемого научного текста.

Еще более отчетливо эта закономерность проявлена у Alice AI LLM. При приемлемой общеязыковой оценке 3,85 дискурсно-специфичная группа опускается до 3,13, давая максимальный среди всех моделей разрыв в 0,72 балла. Alice AI LLM – случай удовлетворительной языковой компетенции без дискурсной состоятельности. Текст порождается грамотно и разнообразно с точки зрения языка, но мало соответствует требованиям, предъявляемым к научной публикации. Это пример того, как владение системой языка и владение знаниями об особенностях жанра могут существовать автономно.

Общеязыковая компетенция, освоенная изученными БЯМ на уровне орфографической, грамматической и пунктуационной нормы, выступает необходимым, но не достаточным условием качественного академического текстопорождения. Дефицитарными оказываются параметры, требующие содержательной, а не формальной ответственности за высказывание (глубина раскрытия темы, доказательность, корректность научного цитирования, выраженность личного авторского вклада). Это та зона, где востребована эпистемическая позиция говорящего, по определению недоступная ИИ-системе.

Повышенная требовательность БЯМ к лексико-синтаксическому разнообразию при одновременной неспособности верифицировать фактическую достоверность и подлинность цитируемых источников указывает то, что БЯМ-оценщик распознает текст как соответствие усвоенному статистическому шаблону, тогда как эксперт-лингвист оценивает его как претендующее на истинность высказывание. Построенная типология БЯМ по соотношению общеязыковых и дискурсно-специфичных показателей демонстрирует, что данные измерения качества генерации относительно независимы, а степень их рассогласования может служить индикатором дискурсивной зрелости системы. Установлена уязвимость существующих методов оценки именно в отношении наиболее качественных генераций (чем ближе сгенерированный текст к жанровому эталону, тем хуже он поддается распознаванию). Проблема оценки ИСТ выходит за рамки лингвистической задачи и становится частью более широкого вопроса – обеспечения безопасности академической среды в условиях, когда ИСТ становится все более убедительным.

Совокупность полученных данных подтверждает продуктивность интегративной лингводискурсивной методики как инструмента, способного зафиксировать расхождение между ЕСТ и ИСТ, распределение зон уверенного освоения, функциональной адекватности и критического дефицита по конкретным лингвистическим параметрам. Дальнейшая перспектива иссле-

дования связана с расширением корпуса за счет иных жанров, что позволит установить, в какой мере выявленные закономерности носят универсальный характер или, напротив, обусловлены спецификой русскоязычного академического дискурса и привлеченных для анализа БЯМ.

ЛИТЕРАТУРА

1. *Liang, W.* Quantifying large language model usage in scientific papers / W. Liang, Y. Zhang, Z. Wu [et al.] // *Nature Human Behaviour*. – 2025. – Т. 9. – № 12. – P. 2599–2609.
2. *Schmidtova, P.* Automatic metrics in natural language generation: a survey of current evaluation practices / P. Schmidtova, S. Mahamood, S. Balloccu [et al.] // *Proceedings of the 17th International Natural Language Generation Conference*. – 2024. – P. 557–583.
3. *Van Miltenburg, E.* How reproducible are fluency ratings of generated text? / E. Van Miltenburg, A. Braggaar, N. Braun [et al.] // *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems*. – 2024. – P. 132–144.
4. *Fu, X.* How reliable is multilingual LLM-as-a-judge? / X. Fu, W. Liu // *Findings of the Association for Computational Linguistics: EMNLP 2025*. – 2025. – P. 11040–11053.
5. *Wei, H.* Systematic evaluation of LLM-as-a-judge in LLM alignment tasks: explainable metrics and diverse prompt templates / H. Wei, S. He, T. Xia [et al.] // *arXiv preprint arXiv:2408.13006*. – 2024.
6. *Гутовская, М. С.* Интегративная лингводискурсивная методика исследования продуктов искусственного текстопорождения / М. С. Гутовская, Д. В. Шайбакова // *Журн. Белорус. гос. ун-та. Филология*. – 2026. – № 2. – С. 32–40.
7. *Mueller, J.* The Authentic Assessment Toolbox: Enhancing Student Learning through Online Faculty Development / J. Mueller // *Journal of Online Learning and Teaching*. – 2005. – № 1. – URL: <http://jfmueLLer.faculty.noctrl.edu/toolbox/rubrics.htm> (date of access: 07.04.2026).

Поступила в редакцию 07.07.2026