

The Impact of Different Prompt Types on the Quality of Automatic Assessment of Student Answers by Artificial Intelligence Models

I. A. Meshcheryakov¹, N. S. Lagutina¹DOI: [10.18255/1818-1015-2025-4-396-416](https://doi.org/10.18255/1818-1015-2025-4-396-416)¹P.G. Demidov Yaroslavl State University, Yaroslavl, Russia

MSC2020: 68T50

Research article

Full text in Russian

Received September 30, 2025

Revised November 2, 2025

Accepted November 18, 2025

Artificial intelligence (AI) models can fully or partially automate the assessment of student assignments, making assessment methods more accurate and objective. The performance of such models depends not only on the underlying algorithms and training data but also on the effectiveness of the queries they formulate. The aim of the work is to investigate the possibility of using open artificial intelligence models to evaluate students' answers for compliance with the teacher's standard answer, as well as to increase the quality of problem solving using prompt engineering. The method for determining this quality was selected by statistical characteristics of the results of classifying answer texts into four categories: correct, partially correct, incorrect, inappropriate to the topic of the question, by GAI models using the following prompt options: simple prompt, role-playing prompt, "chain of thoughts" prompt, prompt generated by artificial intelligence. Models available for open use were selected for the study: ChatGPT o3-mini, DeepSeek V3, Mistral-Small-3.1-24B-Instruct-2503-IQ4.XS and Grok 3. Testing of the models was carried out on a corpus of student texts collected by teachers of Demidov Yaroslavl State University, from 507 answers to 8 questions. The best quality of answer assessment was shown by the ChatGPT o3-mini model. with the prompt it generated. The accuracy rate was 0.82, the mean square error (MSE) was 0.2, and the F-score reached 0.8, demonstrating the potential of GAI as not only an assessment tool but also a means of automatically generating instructions. The Fleiss coefficient was used to assess the consistency of the model's responses across 10 identical queries. For this model-prompt pair, it ranged from 0.48 for complex questions to 0.69 for simple questions.

Keywords: artificial intelligence; prompt engineering; automatic short answer grading; ChatGPT o-3 mini; DeepSeek V3; Mistral-Small-3.1-24B-Instruct-2503-IQ4.XS; Zero-Shot prompting; neural networks; Chain-of-Thought; Role prompting; NLP

INFORMATION ABOUT THE AUTHORS

Meshcheryakov, Ivan A. (corresponding author)	ORCID iD: 0009-0008-2758-206X . E-mail: meshcheryakov_it22@mail.ru Post-graduate Student
--	---

Lagutina, Nadezhda S.	ORCID iD: 0000-0002-6137-8643 . E-mail: lagutinans@rambler.ru PhD, Associate Professor
-----------------------	--

Funding: The study was supported by a grant from the Russian Science Foundation № 25-21-00196, <https://rscf.ru/project/25-21-00196/>.

For citation: I. A. Meshcheryakov and N. S. Lagutina, "The impact of different prompt types on the quality of automatic assessment of student answers by artificial intelligence models", *Modeling and Analysis of Information Systems*, vol. 32, no. 4, pp. 396–416, 2025. DOI: [10.18255/1818-1015-2025-4-396-416](https://doi.org/10.18255/1818-1015-2025-4-396-416).

Влияние различных типов промптов на качество автоматической оценки ответов учащихся моделями искусственного интеллекта

И. А. Мещеряков¹, Н. С. Лагутина¹

DOI: [10.18255/1818-1015-2025-4-396-416](https://doi.org/10.18255/1818-1015-2025-4-396-416)

¹Ярославский государственный университет им. П.Г. Демидова, Ярославль, Россия

УДК 004.891.3

Научная статья

Полный текст на русском языке

Получена 30 сентября 2025 г.

После доработки 2 ноября 2025 г.

Принята к публикации 18 ноября 2025 г.

Модели искусственного интеллекта (AI) могут полностью или частично автоматизировать проверку контрольных работ учащихся, делая методы экспертизы более точными и объективными. Качество работы таких моделей зависит не только от базовых алгоритмов и обучающих данных, но и от эффективности формулируемых запросов. Целью работы является исследование возможности применения открытых моделей искусственного интеллекта для оценивания ответов студентов на соответствие эталонному ответу преподавателя, а также увеличение качества решения задачи при помощи промпт-инжиниринга. Методом определения этого качества выбраны статистические характеристики результатов классификации текстов ответов на четыре категории: правильные, частично правильные, неверные, несоответствующие теме вопроса, моделями AI при использовании следующих вариантов промптов: простой промпт, ролевой промпт, промпт «цепочка мыслей», промпт, сгенерированный искусственным интеллектом. Для исследования были выбраны модели, доступные для открытого использования, ChatGPT o3-mini, DeepSeek V3, Mistral-Small-3.1-24B-Instruct-2503-IQ4_XS и Grok 3. Тестирование моделей проводилось на корпусе текстов студентов, собранном преподавателями ЯрГУ имени Демидова, из 507 ответов на 8 вопросов. Лучшее качество оценки ответов показала модель ChatGPT o3-mini со сгенерированным ей же промптом. Доля правильных ответов (ассурагу) составила 0,82, среднеквадратичная ошибка (MSE) — 0,2, а F-мера достигла 0,8, что показывает перспективность использования AI не только в качестве инструмента оценки, но и в качестве средства автоматической генерации инструкций. Для оценки согласованности ответов модели при 10 одинаковых запросах был использован коэффициент Флейсса. Для указанной пары модели и промпта он составил от 0,48 для сложных вопросов до 0,69 для простых вопросов.

Ключевые слова: искусственный интеллект; промпт-инжиниринг; автоматическая оценка ответов учащихся; ChatGPT o3-mini; DeepSeek V3; Mistral-Small-3.1-24B-Instruct-2503-IQ4_XS; Zero-Shot prompting; нейронные сети; Chain-of-Thought; Role prompting; NLP

ИНФОРМАЦИЯ ОБ АВТОРАХ

Мещеряков, Иван Андреевич
(автор для корреспонденции)

ORCID iD: [0009-0008-2758-206X](https://orcid.org/0009-0008-2758-206X). E-mail: meshcheryakov_it22@mail.ru
Аспирант

Лагутина, Надежда Станиславовна

ORCID iD: [0000-0002-6137-8643](https://orcid.org/0000-0002-6137-8643). E-mail: lagutinans@rambler.ru
Канд. физ.-мат. наук, доцент

Финансирование: Исследование выполнено за счет гранта Российского научного фонда № 25-21-00196, <https://rscf.ru/project/25-21-00196/>.

Для цитирования: I. A. Meshcheryakov and N. S. Lagutina, “The impact of different prompt types on the quality of automatic assessment of student answers by artificial intelligence models”, *Modeling and Analysis of Information Systems*, vol. 32, no. 4, pp. 396–416, 2025. DOI: [10.18255/1818-1015-2025-4-396-416](https://doi.org/10.18255/1818-1015-2025-4-396-416).

Введение

Одной из самых сложных задач в сфере образования является объективная оценка знаний учащихся. Такая оценка с одной стороны существенно повышает вовлечённость учеников в учебный процесс, их мотивацию, степень персонализации образовательной траектории, а с другой стороны, требует от преподавателя максимальной компетентности, методической грамотности, минимализации влияния субъективных факторов [1]. Для решения этой проблемы строятся автоматизированные системы на основе методов машинного и глубокого обучения, а в последние два года активно внедряется генеративный искусственный интеллект (Generative Artificial Intelligence, GAI) [2].

Генеративный искусственный интеллект обладает огромным потенциалом для выполнения конкретных задач и автоматизации различных процессов в образовании, таких как разработка образовательного контента, оценка знаний учащихся, организация обратной связи, прогнозирование успеваемости [3]. Технологии на основе искусственного интеллекта могут полностью или частично автоматизировать проверку контрольных работ учащихся, делая методы экспертизы более точными и объективными. Однако использование GAI для выставления оценки в процессе обучения требует обеспечения того, чтобы модели проверяли те же знания и навыки, которые оценивал бы преподаватель [4].

Основной подход к эффективному взаимодействию с GAI — создание и использование правильно сформулированных инструкций, или промпт-инжиниринг (prompt engineering). Качество ответов зависит не только от базовых алгоритмов и обучающих данных, но и от эффективности формулируемых запросов. Хорошо продуманная инструкция (промпт) может вызвать содержательный и информативный ответ, в то время как плохо продуманная может привести к нерелевантным или бессмысленным результатам [5]. Для преподавателей владение промпт-инжинирингом имеет решающее значение для эффективного взаимодействия с моделями GAI, поскольку является ключом к получению высококачественных результатов [6].

Авторы работы поставили цель проверить гипотезу о возможности применения открытых моделей искусственного интеллекта для оценивания ответов студентов на соответствие эталонному ответу преподавателя, а также о влиянии технологий промпт-инжиниринга на качество решения этой задачи. Методика исследования применения промптов находится в стадии формирования, поэтому основой для анализа были выбраны техники формулировки запросов к GAI, наиболее часто встречающиеся в работах последних лет. Для определения качества решения задачи применены статистические характеристики результатов классификации текстов ответов моделями GAI при использовании разных вариантов промптов.

1. Обзор связанных работ

В статье [7] рассмотрены различные технологии, связанные с промпт-инжинирингом, в сфере образования, такие как инструкции с закрытыми вопросами, открытые запросы, назначение конкретной роли искусственного интеллекта, четкое определение целей, обозначение ограничений, использование инструкций определённой структуры, уточнение ответов в ходе диалога. Авторы привели примеры конкретных промптов, однако не систематизировали эксперименты и не привели статистические оценки результатов. На основе опроса преподавателей был сделан вывод, что применение искусственного интеллекта улучшает процесс обучения за счёт адаптации запросов к индивидуальным потребностям пользователя, повышения вовлечённости учащихся в процесс получения знаний, развития навыков критического мышления и создания интерактивной среды обучения.

Авторы работы [8] представили инструмент CLEAR Framework для составления эффективных промптов для языковых моделей искусственного интеллекта, таких как ChatGPT. По замыслу исследователей во время семинаров по навыкам исследования или информационной грамотности

преподаватели могут использовать CLEAR Framework, чтобы продемонстрировать, как инструкции соответствующие пяти принципам: краткость, логичность, ясность, адаптивность и рефлексивность, повышают качество создаваемых GAI резюме, анализов и обзоров литературы. Авторы отмечают, что учащиеся также должны знать технические аспекты, которые могут влиять на генерируемый контент: минимальные единицы текста, которые обрабатывает модель (токены), случайность ответов модели (температура, более высокое значение ближе к 1 делает ответ более случайным, а более низкое значение ближе к 0 — более предсказуемым), вероятность выбора следующего токена ($\text{top-}p$, значение ближе к 1 означает, что модель будет учитывать все возможные токены, тогда как значение, близкое к 0, заставит модель выбирать только самые вероятные токены). Авторы утверждают, что внедрение принципов CLEAR в академическое образование способствует развитию информационной грамотности студентов, готовых ориентироваться в сложностях технологий, основанных на искусственном интеллекте. К сожалению, авторы не формализуют предложенные принципы и не предлагают способов объективной оценки достигнутых результатов.

Большая группа исследователей [9] показала, что качество ответов GAI в значительной степени зависит от формулировки запросов человеком и предложила методику разработки промптов. Авторы разработали алгоритм автоматического формирования промптов APEER для решения задачи ранжирования текстов по релевантности. Алгоритм итеративно генерирует уточнённые инструкции с помощью обратной связи и оптимизации предпочтений. Авторы сравнивают работу своего алгоритма с разными технологиями формулирования промптов: запрос от человека, генерация промежуточных шагов рассуждения (Chain-of-Thought Prompting, CoT), перефразирование человеческого запроса. Эксперименты с моделями GPT4, LLaMA3, Qwen2 и десятью наборами данных показала лучшее качество APEER по метрике nDCG сравнения прогнозируемого порядка элементов с идеальным, максимальное значение которой оказалось равно 0,86 для GPT4 (1 — идеальное ранжирование), хотя среднее между корпусами данных всего — 0,55. Разработанное решение продемонстрировало повышение производительности по сравнению с запросами от человека, кроме того, было обнаружено, что промпты, генерируемые APEER, лучше переносятся между различными моделями и предметными областями.

Интересные результаты получили исследователи технологий промпт-инжиниринга в медицинском образовании. Сразу же после появления модели GPT-3.5, предоставившей интерфейс онлайн-чата, была проанализирована эффективность ChatGPT на экзамене на получение медицинской лицензии США (USMLE) [10]. Эксперименты проводились на четырёх наборах вопросов и ответов, используемых для студентов-медиков. Авторы статьи обнаружили, что рассмотренная модель на наборе NBME-Free-Step1 показала долю правильных ответов (ассигасу), равную 0,64, что соответствует проходному баллу экзамена, однако для других наборов оказалась ниже (0,42–0,58). Важно, что модель использовалась без какой-либо специальной подготовки или дополнительного обучения. Morjaria и др. [11] использовали ChatGPT для оценки кратких ответов. Из программы подготовки к врачебной практике были случайным образом выбраны десять вопросов и для каждого из них собраны шесть ответов студентов, ранее оцененных преподавателем. Результаты экспериментов показали хороший коэффициент корреляции Спирмена оценок модели GAI и человека: 0,6–0,7. Вопрос согласованности оценок коротких вопросов не только между искусственным интеллектом и человеком, но и внутри самих моделей GAI, например, при повторении промпта для той же самой модели, поднимается в работе [12]. Авторы исследовали способность проверки ответов ChatGPT-4o, используя 557 ответов 215 студентов, зачисленных на онлайн-курс по искусственной вентиляции легких. Оказалось, что искусственный интеллект выставял более низкие оценки, чем оценщики-люди, со средней разницей равной $-1,34$ по 10-балльной шкале. Значения коэффициентов внутриклассовой корреляции ICC1, равные 0,086, и каппа Коэна, равная $-0,0786$, указали на низкую согласованность с экспертами. Дисперсия между пятью сессиями ChatGPT, равная 0,87, показала

хорошую внутреннюю согласованность. Различие результатов описанных исследований, скорее всего, является следствием различий предметных областей, к которым относятся вопросы из корпусов данных. В последнем случае это явно узкоспециализированная тематика, которая требует дообучения моделей.

Вопрос об оценке качества контента, генерируемого моделями искусственного интеллекта по запросу, также поднимается в работе [13]. В качестве количественных характеристик авторы предлагают использовать стандартные оценки качества задач автоматической обработки текстов, таких как классификация, определение сходства, аннотирование и т. д.

Gerd Kortemeyer рассмотрел оценку правильности ответов учащихся моделью GPT-4 [14]. Для экспериментов были взяты два набора данных: SciEntsBank и Beetle. Перед моделью была поставлена задача бинарной оценки на правильные и неправильные ответы, и трёхклассовой оценки, где была добавлена категория противоречивого ответа. Решение получалось с помощью ролевого промпта, который был отправлен через скрипт на языке Python для обработки GPT-4, ответ модель возвращала в виде csv-файла для дальнейшего анализа. Лучшие результаты были достигнуты для корпуса SciEntsBank, F-мера оказалась равна 0,74 для бинарной оценки и 0,73 для трёх категорий. Интересно, что предоставление эталонного ответа снизило качество решения задачи. Кроме того, в ответе оказались ошибки: 58 из 28 372 утверждений учащихся не получили оценок или были пропущены.

Автор исследования [15] сравнил оценку университетских экзаменов ChatGPT 3.5 и преподавателей. Корпус данных состоял из 463 ответов. По результатам эксперимента 70 % оценок ChatGPT находились в пределах 10 % от оценок учителей, а 31 % — в пределах 5 %. ChatGPT, как правило, давал немного более высокие баллы и лучше справлялся с более общими вопросами. В ходе опроса преподавателей и анализа результатов экспериментов сделан вывод, что GAI может выставять правдоподобные оценки, различия есть, но вполне вероятно, что проверка двумя разными экспертами может привести к аналогичным расхождениям. К сожалению, вопрос о роли собственно промпт-инжиниринга не поднимается.

Российские учёные также обращают внимание, что модели GAI открывают новые возможности для совершенствования учебного процесса и персонализации обучения [16]. А. А. Паскова [17] использовала ChatGPT 3.5 для составления тестовых заданий и предоставления обратной связи после их выполнения. С этой целью были построены специальные промпты, которые использовались как преподавателем, так и студентами. Группа учащихся, использовавшая GPT-3.5 для самоконтроля и получения индивидуальных рекомендаций по дальнейшему изучению курса, показала лучшие результаты итогового тестирования по сравнению с группой, изучавшей материалы классическими методами.

Анализ работ показывает, что в настоящий момент в полной мере не сформированы методы исследования результатов промпт-инжиниринга. Однако можно выделить определённые виды промптов, технология формирования которых поддаётся формализации и используется разными учёными.

В современном ландшафте исследований по управлению генеративными языковыми моделями (LLM) различают, по крайней мере, три перспективных подхода к формулированию промптов — ролевые промпты, цепочки мыслей (chain-of-thought, CoT) и самогенерируемые промпты, каждый из которых демонстрирует свою эффективность и ограничения в зависимости от задачи и контекста.

Ролевой промпт (role-play prompting) побуждает модель «принять роль», например, выступать в роли эксперта, судьи или преподавателя, что может активировать и структурировать её рассуждения вокруг заданной перспективы. В статье [18] авторы отмечают повышение качества решения задач на различных тестовых наборах без дообучения моделей.

Техника цепочек мыслей (Chain-of-Thought, CoT) стимулирует модель выдавать промежуточные шаги рассуждения перед финальным ответом, что часто улучшает производительность на задачах

с многошаговой логикой. Работа [19] демонстрирует, что включение таких рассуждений существенно повышает точность на задачах арифметики, здравого смысла и символических задачах. Более поздние исследования [20] анализируют, какие аспекты демонстраций рассуждений, например, релевантность, порядок шагов, действительно важны для эффективности CoT. Тем не менее, эта техника может проваливаться при решении субъективных задач [21], а также существует риск, что соответствующие рассуждения «скрывают» сигналы об ошибках или галлюцинациях [22], поэтому требует дополнительных исследований для определения границ применимости.

Подход с самогенерируемым промптом (self-prompting / derived prompt generation) подразумевает, что сама модель конструирует вспомогательные промпты или подзадачи, которые затем используются для усиления её собственного вывода. Примером может служить работа [23], где предложен механизм, при котором LLM генерирует «derived prompts» и использует их для построения более информативного контекстного окружения ответа. Этот метод открывает путь к более автономному и адаптивному поведению, но в то же время требует осторожности при контроле качества сгенерированных подсказок и исключении ошибок самоусиления.

Обзор работ даёт основание взять в основу метода исследования описанные технологии формирования промптов. Для оценки качества работы моделей GAI значительная часть учёных используют точность, полноту, F-меру, долю правильных ответов (ассурасу), как стандартные числовые метрики результатов классификации [24].

2. Метод оценки заданий с открытым ответом

Для исследования были выбраны четыре модели генеративного искусственного интеллекта разных разработчиков, доступные всем пользователям:

- ChatGPT o3-mini — модель от компании OpenAI, появившаяся в декабре 2024 года;
- DeepSeek V3 — китайская модель для анализа текста и генерации ответов, известная скоростью и точностью;
- Mistral-Small-3.1-24B-Instruct-2503-IQ4_XS — компактная, но высокопроизводительная модель от Mistral AI, оптимизированная для локального использования;
- Grok 3 — модель от компании Илона Маска xAI, ориентированная на диалоговые задачи и высокую степень контекстного понимания.

Тестирование моделей проводилось на корпусе ответов студентов, собранном преподавателями ЯрГУ имени П. Г. Демидова. Корпус состоял из 507 ответов студентов на 8 вопросов, перечисленных ниже.

1. Что такое модель в архитектуре модель — вид — контроллер?
2. Что такое архитектура программного приложения?
3. Что такое наследование в объектно-ориентированном программировании?
4. Дайте определение искусственной нейронной сети.
5. Что такое производная функции?
6. Какие алгоритмы сортировки массива с большим количеством чисел считаются самыми быстрыми и почему?
7. Что такое файл в области канцелярских принадлежностей?
8. Что такое файл в операционной системе компьютера?

Для достижения объективности и качества оценка работ проводилась группой из четырёх преподавателей. Итоговая оценка была согласована со всеми экспертами. Этот процесс оказался особенно трудоёмким. Для каждого вопроса был сформулирован эталонный ответ. Ответы учащихся получены в условиях реальной учебной среды. Аналогичных открытых корпусов не было обнаружено, что во многом объясняется большими временными затратами на их создание, поэтому в рамках работы используется один набор данных.

Оценивание ответов студентов осуществлялось по следующей системе:

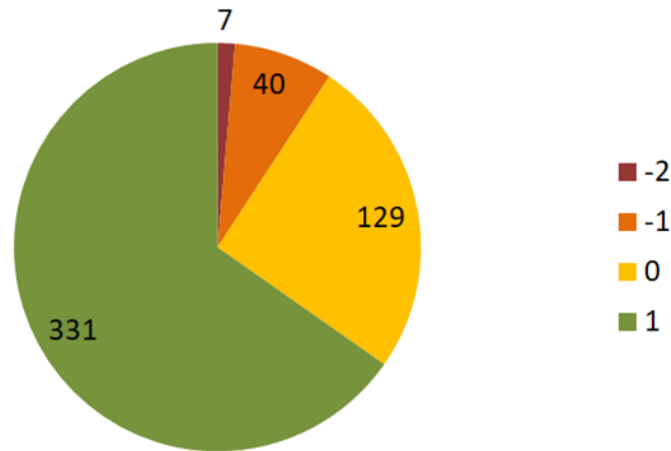


Fig. 1. Distribution of data according to expert estimates

Рис. 1. Распределение данных по оценкам экспертов

- -2 — ответ не соответствует теме вопроса;
- -1 — ответ неправильный (соответствие эталонному ответу менее 50 %);
- 0 — ответ частично правильный (соответствие эталонному ответу от 50 % до 90%);
- 1 — ответ правильный (соответствие эталонному ответу более 90 %).

При первоначальной экспертной оценке использовалось исключительно текстовое описание оценок. Указанные проценты соответствия эталонному ответу были добавлены позже, в процессе формирования промптов. Конкретные значения были подобраны в ходе экспериментов. Они не являются строгой метрикой. Данные значения используются, чтобы задать желаемый уровень строгости модели. Приведя данный уровень строгости в соответствии с уровнем строгости реальных экспертов, можно добиться более релевантных оценок. Без такого уточнения модель интерпретирует любую неточность в ответе как ошибку, в то время как реальные эксперты склонны прощать небольшое отклонение от эталона.

Распределение данных по классам в соответствии с оценками преподавателей представлено на рисунке 1.

Перед моделью стояла задача классифицировать ответы студентов на четыре категории оценок, перечисленных выше. Каждая модель в данном исследовании была протестирована с различными вариантами промптов. Во всех экспериментах использовался подход «Zero-Shot Prompting», при котором модель не обучалась для данной задачи, а сразу получала на вход промпт со словесным описанием и затем ответы для оценивания.

Первоначальный промпт объяснял модели, что нужно сделать, включал в себя список вопросов, эталонных ответов и систему оценивания:

Первоначальный промпт. «Необходимо оценить соответствие ответов студентов эталонным ответам преподавателя. Далее приводится список вопросов по номерам и эталонные ответы к ним, а также список возможных оценок и их описание. Запомни их, после чего мы начнём работу.» (Далее в промпте шёл пронумерованный список вопросов с эталонными ответами, а в самом конце — список возможных оценок.)

На следующем этапе модель получала однотипные промпты для каждого вопроса:

Простой промпт. «Привожу список ответов студентов на вопрос 1. Оценки предоставь мне в виде колонки для копирования в Excel.» (Далее шёл пронумерованный список ответов на вопрос.)

Из-за ограничения на количество токенов в промпте список ответов приходилось разбивать на части (батчи) по 50 ответов. На основе экспериментов с ChatGPT o3-mini данный промпт был признан неэффективным, так как в результате модель выдавала колонку с оценками, но количество оценок в колонке не соответствовало ожидаемому, а сами оценки больше напоминали случайные числа. Для преодоления проблемы был использован уточнённый простой промпт, который в дальнейшем применялся и при тестировании других моделей.

Уточнённый простой промпт. *«Привожу список ответов студентов на вопрос 1. Сначала вспомни сам вопрос, который соответствует указанному номеру. Затем тебе необходимо оценить КАЖДЫЙ из ответов. Не пропускай ответы и не заканчивай раньше времени — мне нужны оценки для КАЖДОГО ответа. Оценки предоставь в виде колонки, которую я скопирую в Excel: ты должен взять каждую оценку из списка и составить соответствующую им колонку. В этой колонке должно быть ровно столько же оценок, сколько и в изначальном списке. Оценки должны быть те же самые, в том же порядке.»*

Для исследования возможностей повышения качества оценивания был использован ролевой промпт:

Ролевой промпт. *«Вы — преподаватель университета, специализирующийся на дисциплинах в области информационных технологий. Ваша цель — профессионально и беспристрастно оценить по шкале от –2 до 1 ответы студентов на вопросы из приведённого ниже списка.»* (Далее был приведён список вопросов, эталонных ответов и описание системы оценивания.)

Ролевой промпт (Role prompting) — это инструкция, в которой модели задаётся определённая роль, стиль поведения и контекст для выполнения задания [18]. В отличие от простого промпта, ролевой подход помогает модели действовать более целенаправленно, что повышает качество и согласованность ответов. В данном случае модель принимает роль преподавателя, чтобы оценка студенческих ответов была максимально близка к экспертной и соответствовала профессиональной этике.

Далее была применена техника «цепочка мыслей» (Chain-of-Thought). В промпт было добавлено требование к модели обосновывать каждый свой ответ:

Промпт «Цепочка мыслей». *«Привожу список ответов студентов на вопрос 1. Сначала вспомни сам вопрос, который соответствует указанному номеру. Затем тебе необходимо оценить КАЖДЫЙ из ответов и обосновать своё решение о выставяемой оценке. Не пропускай ответы и не заканчивай раньше времени — мне нужны оценки для КАЖДОГО ответа. Для каждого ответа сначала повтори сам ответ, проанализируй его и укажи оценку с обоснованием, почему оценка именно такая. Наконец, предоставь мне получившиеся оценки в виде колонки, которую я скопирую в Excel: ты должен взять каждую оценку из списка и составить соответствующую им колонку. В этой колонке должно быть ровно столько же оценок, сколько и в изначальном списке. Оценки должны быть те же самые, в том же порядке.»*

Промпт «цепочка мыслей» — это техника, при которой модель явно побуждают рассуждать пошагово, объяснять ход своих мыслей и показывать логику, ведущую к ответу [25]. Такой подход помогает повысить точность при сложных или неоднозначных заданиях, так как модель не просто выдает результат, а последовательно анализирует ввод и проверяет свои рассуждения. В данном случае требование обосновывать каждую выставленную оценку стимулирует более вдумчивую проверку ответов студентов и снижает риск случайных или поверхностных решений.

После тестирования различных вариантов промпта, созданных вручную, был применен альтернативный подход — создание промпта при помощи модели искусственного интеллекта. Для этого был использован следующий промпт:

«Привет. Я хочу использовать ИИ для оценки ответов студентов на разные вопросы. Требуется использовать следующую систему оценивания:» (далее была описана система оценивания). «Оценки я бы хотел получать в виде колонки с цифрами, которую легко можно было бы скопировать и вставить в Excel таблицу. Напиши эффективный промпт, который я мог бы использовать для решения данной задачи. Используй комбинацию различных методов промпт-инжиниринга, таких как цепочка мыслей, ролевой промпт и другие.»

Модель ChatGPT o3-mini сгенерировала следующий вариант промпта:

Промпт ИИ. «Ты — строгий, но справедливый преподаватель-эксперт с 20-летним стажем. Твоя задача — оценить ответы студентов по 4-балльной шкале (–2 до 1) максимально объективно.

****Алгоритм оценки:****

1. Определи, соответствует ли ответ теме вопроса. Если нет → –2.
2. Если соответствует, оцени полноту и точность:
 - ≥ 91 % правильной информации → 1.
 - 50–90 % правильной информации → 0.
 - <50 % правильной информации → –1.

****Требования к выводу:****

- Выводи ТОЛЬКО цифры (по одной на строку), без пояснений.
- Формат: одна оценка на строку, без точек, скобок и других символов.
- Сохраняй порядок исходных ответов.

****Контекст:****

Вопрос: «{вставить_вопрос}».

Эталонный ответ: «{вставить_эталонный_ответ}».

****Ответы студентов для оценки (начинай обработку после двоеточия):****»

Этот промпт ИИ был применён ко всем моделям данного исследования.

Для оценки соответствия ответов студентов эталонным ответам преподавателя применяются как классификационные, так и числовые метрики качества. Accurasy позволяет определить долю ответов, оценённых верно, то есть совпадающих с эталоном. Precision и Recall дают возможность глубже проанализировать качество оценивания для отдельных категорий ответов: первая показывает точность выставления конкретной оценки (сколько из ответов с данной оценкой действительно соответствовали эталонной оценке), а вторая — полноту (сколько из действительно соответствующих эталону оценок были правильно распознаны). F-мера позволяет сбалансировать эти два показателя. Использование MSE (Mean Squared Error) в данном контексте может показаться неочевидным, однако оно оправдано: оценки студентов формируют линейную шкалу (от –2 до 1 баллов с порядковой интерпретацией шагов). В этом случае разность между поставленной оценкой и эталонной отражает степень отклонения, а возведение в квадрат подчёркивает более серьёзные ошибки. Таким образом, MSE позволяет количественно зафиксировать среднюю величину отклонения оценок студента от эталонных, что дополняет классификационные метрики, фиксирующие лишь сам факт совпадения или несовпадения [15].

3. Модель ChatGPT o3-mini

Первой моделью, с которой проводились эксперименты, стал ChatGPT o3-mini. Модель OpenAI o3-mini представляет собой компактную версию серии моделей с улучшенными способностями к логико-математическим рассуждениям (reasoning). Она ориентирована на решение задач в области точных наук, программирования и анализа данных при сохранении высокой вычислительной эффективности и низкой задержки отклика. Архитектурные и обучающие параметры модели не раскрываются в полном объёме, однако известно, что в процессе обучения применялись масштабные датасеты, включающие разнообразные текстовые источники, а также процедуры оптимизации

Table 1. ChatGPT o3-mini testing results with simple prompt

Номер вопроса	Кол-во данных	Accuracy	MSE	Precision	Recall	F-мера
Вопрос 1	137	0,6423	0,7591	0,4512	0,5234	0,3927
Вопрос 2	49	0,5918	0,4082	0,5076	0,5983	0,5421
Вопрос 3	59	0,7119	0,3390	0,4875	0,6152	0,4286
Вопрос 4	58	0,6379	0,4138	0,6087	0,6891	0,6424
Вопрос 5	59	0,5763	0,6780	0,4219	0,5042	0,3548
Вопрос 6	58	0,5000	0,6552	0,3728	0,4567	0,3968
Вопрос 7	44	0,7500	0,2500	0,5684	0,6229	0,4671
Вопрос 8	43	0,5814	0,4884	0,4021	0,4872	0,3710
Все вопросы	507	0,6252	0,5483	0,4725	0,5560	0,4406

Таблица 1. Результаты тестирования ChatGPT o3-mini с простым промптом**Table 2.** ChatGPT o3-mini testing results with role prompt

Номер вопроса	Кол-во данных	Accuracy	MSE	Precision	Recall	F-мера
Вопрос 1	137	0,7956	0,2482	0,6317	0,7034	0,6625
Вопрос 2	49	0,7347	0,2653	0,5412	0,6647	0,5969
Вопрос 3	59	0,6949	0,3051	0,7195	0,8431	0,7783
Вопрос 4	58	0,4655	0,5345	0,3526	0,4821	0,4088
Вопрос 5	59	0,2881	1,3220	0,1524	0,2497	0,1897
Вопрос 6	58	0,4483	0,6034	0,2789	0,3995	0,3333
Вопрос 7	44	0,5909	0,4773	0,6031	0,7896	0,6898
Вопрос 8	43	0,6977	0,3023	0,7012	0,8019	0,7463
Все вопросы	507	0,6154	0,4793	0,5085	0,6189	0,5574

Таблица 2. Результаты тестирования ChatGPT o3-mini с ролевым промптом

под задачи многошаговых рассуждений. По сравнению с более крупными аналогами, o3-mini характеризуется сниженной вычислительной стоимостью и повышенной скоростью работы, что делает её пригодной для массового применения в интерактивных системах и встраиваемых аналитических решениях.

После получения оценок для всех ответов, были рассчитаны метрики качества. Доля правильных ответов (Ассигасу) в среднем для всех вопросов составила 0,62, средняя квадратичная ошибка (MSE) оказалась равна 0,55, а F-мера имела значение 0,44. Полные результаты тестирования представлены в таблице 1.

Для повышения качества оценивания был использован ролевой промпт. В результате, Ассигасу в среднем для всех вопросов осталась на уровне 0,62, средняя квадратичная ошибка снизилась до 0,48, а F-мера увеличилась до 0,58. Полные результаты тестирования представлены в таблице 2.

Далее была применена техника «цепочка мыслей». Модель выполнила поставленную задачу. Ассигасу достигла 0,79, средняя квадратичная ошибка снизилась до 0,22, а F-мера приняла значение 0,77. Оценки модели крайне редко отклонялись от оценок преподавателя более чем на 1 балл, а в спорных случаях предоставлялись аргументированные объяснения. Наибольшее расхождение наблюдалось при оценке частично правильных ответов. Полные результаты тестирования представлены в таблице 3.

Несмотря на то, что первоначальные оценки модели были не верными даже по форме, уточнив промпт, авторы получили корректный список с оценками, при этом ассигасу составила 0,62, а средняя квадратичная ошибка (MSE) оказалась равна 0,55. Применив технику «Chain-of-Thought», ассигасу удалось увеличить до 0,79, а среднюю квадратичную ошибку уменьшить до 0,22. Таким

Table 3. ChatGPT o3-mini testing results with the "chain of thoughts" prompt

Номер вопроса	Кол-во данных	Accuracy	MSE	Precision	Recall	F-мера
Вопрос 1	137	0,7445	0,2993	0,5723	0,6187	0,5897
Вопрос 2	49	0,8367	0,1633	0,8912	0,9407	0,9140
Вопрос 3	59	0,8136	0,1864	0,7421	0,8196	0,7783
Вопрос 4	58	0,7241	0,2759	0,7124	0,8043	0,7564
Вопрос 5	59	0,9831	0,0169	0,9652	0,9810	0,9725
Вопрос 6	58	0,6724	0,3276	0,6021	0,7482	0,6747
Вопрос 7	44	0,9773	0,0227	0,9657	0,9888	0,9773
Вопрос 8	43	0,6977	0,3721	0,7421	0,8907	0,8094
Все вопросы	507	0,7949	0,2229	0,7366	0,8066	0,7686

Таблица 3. Результаты тестирования ChatGPT o3-mini с промптом «цепочка мыслей»**Table 4.** ChatGPT o3-mini testing results with AI prompt

Номер вопроса	Кол-во данных	Accuracy	MSE	Precision	Recall	F-мера
Вопрос 1	137	0,7664	0,2774	0,6124	0,6805	0,6442
Вопрос 2	49	0,8367	0,1633	0,8803	0,9415	0,9099
Вопрос 3	59	0,8305	0,1695	0,7402	0,8553	0,7947
Вопрос 4	58	0,7586	0,2414	0,7251	0,8296	0,7740
Вопрос 5	59	0,9831	0,0169	0,9812	0,9909	0,9860
Вопрос 6	58	0,7586	0,2414	0,7107	0,7634	0,7333
Вопрос 7	44	0,9773	0,0227	0,9684	0,9914	0,9798
Вопрос 8	43	0,7442	0,3256	0,7421	0,9053	0,8184
Все вопросы	507	0,8205	0,1972	0,7621	0,8348	0,7961

Таблица 4. Результаты тестирования ChatGPT o3-mini с ИИ промптом**Table 5.** General results of testing ChatGPT o3-mini

Номер вопроса	Accuracy	MSE	Precision	Recall	F-мера
Простой промпт	0,5858	0,5680	0,4725	0,5560	0,5408
Ролевой промпт	0,6489	0,4379	0,5085	0,6189	0,5660
Цепочка мыслей	0,7101	0,2209	0,7366	0,8066	0,6450
ИИ промпт	0,7160	0,1972	0,7621	0,8348	0,6548

Таблица 5. Общие результаты тестирования ChatGPT o3-mini

образом, качество промпта сильно влияет на качество ответов модели. Метод «цепочка мыслей» показал наилучшие результаты, значительно обогнав как простой, так и ролевой промпт.

Далее модель была протестирована с промптом ИИ. Accuracy достигла 0,82, средняя квадратичная ошибка снизилась до 0,2, а F-мера приняла значение 0,8. Полные результаты тестирования представлены в таблице 4. Создание промпта при помощи самой модели оказалось самым быстрым и удобным способом формулирования запроса и самым эффективным с точки зрения итоговых метрик качества модели. Общие результаты экспериментов с моделью ChatGPT o3-mini для каждого варианта промпта представлены в таблице 5.

На рисунке 2 представлено количественное сравнение оценок экспертов и оценок модели ChatGPT o3-mini с ИИ промптом. Наблюдается схожее распределение оценок между экспертами и моделью, однако с рядом отклонений. Модель в большей степени склонна давать нейтральные оценки: при оценке «0» и «1» она показывает соответственно 155 и 301 случаев, тогда как эксперты — 129 и 331. В отрицательной зоне различия минимальны: для оценки «-1» показатели почти совпадают (40 у экспертов против 41 у модели), тогда как редкая оценка «-2» моделью используется чаще (10 против 7). В целом модель демонстрирует близкую к экспертной структуру распределе-

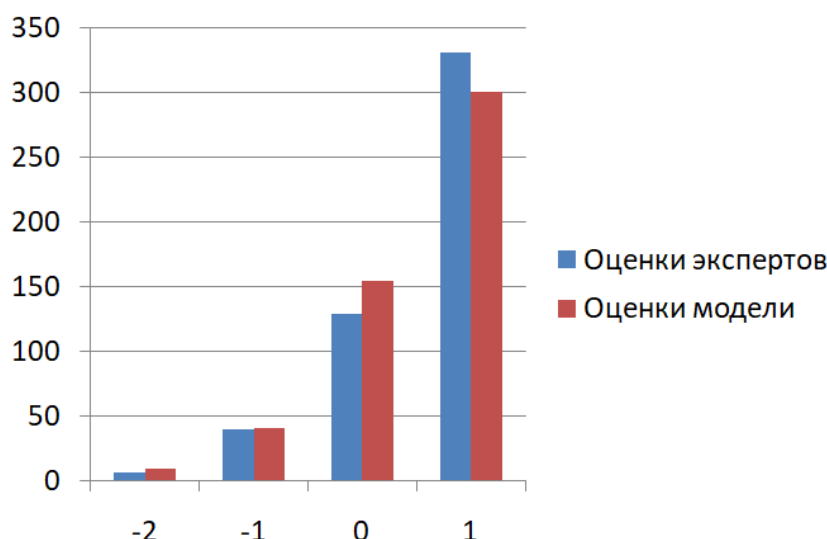


Fig. 2. Error Analysis of the ChatGPT o3-mini Model with AI Prompt

Рис. 2. Анализ ошибок модели ChatGPT o3-mini с ИИ промптом

Table 6. DeepSeek-V3 testing results with simple prompt

Номер вопроса	Кол-во данных	Accuracy	MSE	Precision	Recall	F-мера
Вопрос 1	137	0,5474	0,5620	0,5204	0,5837	0,5507
Вопрос 2	49	0,5918	0,6531	0,4981	0,6225	0,5556
Вопрос 3	59	0,6102	0,5763	0,5452	0,5653	0,5538
Вопрос 4	58	0,5517	0,6034	0,4917	0,5614	0,5246
Вопрос 5	59	0,6949	0,5085	0,5732	0,6115	0,5915
Вопрос 6	58	0,5000	0,5000	0,4682	0,5595	0,5172
Вопрос 7	44	0,7273	0,3864	0,5712	0,6173	0,5926
Вопрос 8	43	0,5349	0,7907	0,5027	0,5675	0,5333
Все вопросы	507	0,5858	0,5680	0,5209	0,5848	0,5408

Таблица 6. Результаты тестирования DeepSeek-V3 с простым промптом

ния, но с тенденцией к сглаживанию крайних значений и смещением в сторону более нейтральных оценок.

Следует отметить, что у других моделей были обнаружены аналогичные тенденции, поэтому подробное сравнение результатов с экспертными приведено только для ChatGPT o3-mini, оказавшейся в итоге лучшей по качеству.

4. Модель DeepSeek V3

Модель DeepSeek-V3, разработанная исследовательской группой DeepSeek, реализует архитектуру типа Mixture-of-Experts (MoE), сочетающую масштабируемость и эффективность за счёт активации лишь части параметров при обработке каждого токена. Согласно технической документации, модель обладает суммарным параметрическим объёмом порядка сотен миллиардов (порядка 671 млрд), при этом активно используются лишь около 37 млрд параметров в каждый момент времени. В архитектуре применены методы Multi-head Latent Attention, а также усовершенствованные механизмы балансировки нагрузки между экспертами. Обучение осуществлялось на крупных корпусах текстов высокого качества, что позволило достичь значительных показателей производительности в сравнении с предыдущей версией (DeepSeek-V2). Модель предназначена для выполнения

Table 7. DeepSeek-V3 testing results with role prompt**Таблица 7.** Результаты тестирования DeepSeek-V3 с ролевым промптом

Номер вопроса	Кол-во данных	Accuracy	MSE	Precision	Recall	F-мера
Вопрос 1	137	0,6277	0,4161	0,5904	0,6757	0,6324
Вопрос 2	49	0,6735	0,3673	0,5402	0,6351	0,5862
Вопрос 3	59	0,6441	0,3898	0,5083	0,6359	0,5672
Вопрос 4	58	0,6034	0,5517	0,4898	0,6463	0,5625
Вопрос 5	59	0,7966	0,4068	0,5731	0,6817	0,6234
Вопрос 6	58	0,5345	0,4483	0,4982	0,5777	0,5333
Вопрос 7	44	0,7727	0,4318	0,5472	0,6735	0,6071
Вопрос 8	43	0,5814	0,5349	0,5109	0,6022	0,5532
Все вопросы	507	0,6489	0,4379	0,5414	0,6468	0,5660

Table 8. DeepSeek-V3 testing results with the "chain of thoughts" prompt**Таблица 8.** Результаты тестирования DeepSeek-V3 с промптом «цепочка мыслей»

Номер вопроса	Кол-во данных	Accuracy	MSE	Precision	Recall	F-мера
Вопрос 1	137	0,6861	0,2920	0,5782	0,6721	0,6204
Вопрос 2	49	0,7347	0,1633	0,6034	0,7082	0,6531
Вопрос 3	59	0,7288	0,1864	0,5912	0,7043	0,6441
Вопрос 4	58	0,7241	0,2759	0,5641	0,6879	0,6207
Вопрос 5	59	0,8475	0,0169	0,7234	0,8418	0,7797
Вопрос 6	58	0,5517	0,3276	0,4521	0,5567	0,5000
Вопрос 7	44	0,9091	0,0227	0,8412	0,9367	0,8864
Вопрос 8	43	0,5349	0,3721	0,4623	0,5658	0,5116
Все вопросы	507	0,7101	0,2209	0,5960	0,7016	0,6450

Table 9. DeepSeek-V3 testing results with AI prompt**Таблица 9.** Результаты тестирования DeepSeek-V3 с ИИ промптом

Номер вопроса	Кол-во данных	Accuracy	MSE	Precision	Recall	F-мера
Вопрос 1	137	0,6715	0,2774	0,5721	0,6592	0,6131
Вопрос 2	49	0,7143	0,1633	0,5924	0,6761	0,6327
Вопрос 3	59	0,7119	0,1695	0,5837	0,6746	0,6271
Вопрос 4	58	0,6724	0,2414	0,5408	0,6371	0,5862
Вопрос 5	59	0,8814	0,0169	0,7632	0,8673	0,8136
Вопрос 6	58	0,6034	0,2414	0,4871	0,6243	0,5517
Вопрос 7	44	0,9545	0,0227	0,9082	0,9581	0,9318
Вопрос 8	43	0,6047	0,3256	0,5294	0,6403	0,5814
Все вопросы	507	0,7160	0,1972	0,6099	0,7047	0,6548

широкого спектра задач, включая диалоговую генерацию, анализ данных и решение прикладных инженерных проблем.

Модель DeepSeek-V3 была протестирована со всеми промптами, описанными в разделе 2. Результаты тестирования представлены в таблицах 6, 7, 8 и 9. Для модели DeepSeek-V3 наблюдается устойчивое улучшение качества оценивания по мере усложнения промпта. При использовании простого промпта Аккуратность составила 0,59, а MSE была относительно высокой (0,57), что говорит о частых отклонениях от эталонных оценок. Переход к ролевому промпту позволил снизить среднюю квадратичную ошибку до 0,44 и немного повысить точность и F-меру, что указывает на более осмысленное распределение оценок. Применение техники «цепочка мыслей» значительно улуч-

Table 10. General results of testing DeepSeek-V3**Таблица 10.** Общие результаты тестирования DeepSeek-V3

Номер вопроса	Accuracy	MSE	Precision	Recall	F-мера
Простой промпт	0,5858	0,5680	0,5209	0,5848	0,5408
Ролевой промпт	0,6489	0,4379	0,5414	0,6468	0,5660
Цепочка мыслей	0,7101	0,2209	0,5960	0,7016	0,6450
ИИ промпт	0,7160	0,1972	0,6099	0,7047	0,6548

Table 11.Mistral-Small-3.1-24B-Instruct-2503-IQ4-XS testing
results with simple prompt**Таблица 11.** Результаты тестирования
Mistral-Small-3.1-24B-Instruct-2503-IQ4-XS
с простым промптом

Номер вопроса	Кол-во данных	Accuracy	MSE	Precision	Recall	F-мера
Вопрос 1	137	0,5474	0,4745	0,3128	0,3924	0,3838
Вопрос 2	49	0,6531	0,4490	0,2423	0,3987	0,3077
Вопрос 3	59	0,6610	0,3390	0,2412	0,4110	0,3125
Вопрос 4	58	0,5690	0,5690	0,2205	0,4302	0,3000
Вопрос 5	59	0,7797	0,3559	0,2984	0,4359	0,3582
Вопрос 6	58	0,5000	0,6552	0,1951	0,3988	0,2759
Вопрос 7	44	0,7500	0,3182	0,2980	0,4436	0,3600
Вопрос 8	43	0,5581	0,4419	0,1980	0,4210	0,2727
Все вопросы	507	0,6134	0,4576	0,2609	0,4122	0,2910

Table 12.Mistral-Small-3.1-24B-Instruct-2503-IQ4-XS testing
results with role prompt**Таблица 12.** Результаты тестирования
Mistral-Small-3.1-24B-Instruct-2503-IQ4-XS
с ролевым промптом

Номер вопроса	Кол-во данных	Accuracy	MSE	Precision	Recall	F-мера
Вопрос 1	137	0,6204	0,4453	0,3541	0,6025	0,4583
Вопрос 2	49	0,7143	0,3878	0,2102	0,5798	0,3333
Вопрос 3	59	0,6441	0,4915	0,2001	0,5623	0,3175
Вопрос 4	58	0,6207	0,4138	0,1857	0,4998	0,2951
Вопрос 5	59	0,8136	0,3220	0,2635	0,5197	0,3529
Вопрос 6	58	0,5172	0,4828	0,1650	0,4521	0,2712
Вопрос 7	44	0,7727	0,4318	0,2773	0,4949	0,3600
Вопрос 8	43	0,5349	0,6512	0,1523	0,5038	0,2727
Все вопросы	507	0,6489	0,4477	0,2471	0,5393	0,3046

шило все показатели: Ассурасу выросла до 0,71, F-мера — до 0,65, а MSE сократилась более чем в два раза (до 0,22), что означает сокращение величины ошибок в оценивании. Наилучший результат показал промпт, сгенерированный самой моделью (ИИ-промпт): Ассурасу достигла 0,72, F-мера — 0,65, а MSE снизилась до 0,20. Таким образом, и в случае DeepSeek-V3 метод пошагового рассуждения и использование автоматически сформулированного промпта оказываются наиболее эффективными для повышения качества оценивания. В таблице 10 приведены общие результаты экспериментов с данной моделью. DeepSeek-V3 справилась с данной задачей хуже, чем ChatGPT o3-mini. Любопытным фактом является то, что в процессе проведения тестирования, когда модели был задан вопрос «какую модель я в данный момент использую», она ответила, что является ChatGPT-4, что может указывать на то, что DeepSeek-V3 обучалась на диалогах пользователей с моделью ChatGPT-4.

Table 13.
Mistral-Small-3.1-24B-Instruct-2503-IQ4-XS testing
results with the “chain of thoughts” prompt

Номер вопроса	Кол-во данных	Accuracy	MSE	Precision	Recall	F-мера
Вопрос 1	137	0,6204	0,2920	0,2752	0,5497	0,3796
Вопрос 2	49	0,6531	0,1633	0,1425	0,5174	0,2653
Вопрос 3	59	0,6441	0,1864	0,1550	0,4921	0,2712
Вопрос 4	58	0,6207	0,2759	0,1591	0,5027	0,2759
Вопрос 5	59	0,7797	0,0169	0,0854	0,4025	0,1525
Вопрос 6	58	0,5000	0,3276	0,2920	0,6254	0,4483
Вопрос 7	44	0,8864	0,0227	0,0521	0,1633	0,0909
Вопрос 8	43	0,5116	0,3721	0,2234	0,7252	0,4651
Все вопросы	507	0,6450	0,2209	0,1912	0,5074	0,3077

Таблица 13. Результаты тестирования
Mistral-Small-3.1-24B-Instruct-2503-IQ4-XS
с промптом «цепочка мыслей»

Table 14.
Mistral-Small-3.1-24B-Instruct-2503-IQ4-XS results
with AI prompt

Номер вопроса	Кол-во данных	Accuracy	MSE	Precision	Recall	F-мера
Вопрос 1	137	0,6131	0,2774	0,2421	0,6712	0,3942
Вопрос 2	49	0,6327	0,1633	0,1357	0,6214	0,2857
Вопрос 3	59	0,6271	0,1695	0,1462	0,5847	0,2881
Вопрос 4	58	0,5862	0,2414	0,1923	0,5789	0,3276
Вопрос 5	59	0,8136	0,0169	0,0667	0,2431	0,1186
Вопрос 6	58	0,5517	0,2414	0,1982	0,8765	0,3966
Вопрос 7	44	0,9318	0,0227	0,0241	0,1035	0,0455
Вопрос 8	43	0,5814	0,3256	0,2003	0,8912	0,3953
Все вопросы	507	0,6548	0,1972	0,1671	0,5888	0,3018

Таблица 14. Результаты тестирования
Mistral-Small-3.1-24B-Instruct-2503-IQ4-XS с ИИ
промптом

5. Модель Mistral-Small-3.1-24B-Instruct-2503-IQ4_XS

Модель Mistral-Small-3.1-24B-Instruct-2503, разработанная компанией Mistral AI, представляет собой инструкционно-ориентированную языковую модель среднего размера (около 24 млрд параметров), оптимизированную для снижения задержки при генерации текста и работы в условиях ограниченных вычислительных ресурсов. В процессе разработки модель подвергалась инструкционному дообучению (instruction tuning) для улучшения соответствия ответов пользовательским запросам, включая задачи программирования, работы с длинными документами и обработки многошаговых рассуждений. Одной из особенностей данной версии является возможность эффективного квантования (включая конфигурацию IQ4_XS), что обеспечивает развёртывание на графических процессорах потребительского уровня или серверах с ограниченным объёмом оперативной памяти. По своей функциональности модель ориентирована на прикладное использование в локальных системах, аналитических приложениях и специализированных чат-ботах.

Особенностью данной модели является ее относительная компактность — в отличие от двух предыдущих моделей, тестирование которых осуществлялось через их публичные веб-интерфейсы, модель Mistral удалось запустить на персональном компьютере с 12 ГБ видеопамати и 32 ГБ оперативной памяти. Результаты тестирования с различными промптами представлены в таблицах 11, 12, 13 и 14. В таблице 15 приведены общие результаты экспериментов с данной моделью. Для модели Mistral-Small-3.1-24B-Instruct-2503-IQ4-XS прирост качества при изменении стратегий промптирования выражен менее отчётливо. Ассурасу повышается умеренно: с 0,61 при простом промпте до 0,65 при использовании ИИ-промпта. Однако средняя квадратичная ошибка замет-

Table 15. General results of testing
Mistral-Small-3.1-24B-Instruct-2503-IQ4-XS

Номер вопроса	Accuracy	MSE	Precision	Recall	F-мера
Простой промпт	0,6134	0,4576	0,2609	0,4122	0,2910
Ролевой промпт	0,6489	0,4477	0,2471	0,5393	0,3046
Цепочка мыслей	0,6450	0,2209	0,1912	0,5074	0,3077
ИИ промпт	0,6548	0,1972	0,1671	0,5888	0,3018

Таблица 15. Общие результаты тестирования
Mistral-Small-3.1-24B-Instruct-2503-IQ4-XS**Table 16.** Grok 3 testing results with simple prompt

Номер вопроса	Кол-во данных	Accuracy	MSE	Precision	Recall	F-мера
Вопрос 1	137	0,4088	0,5839	0,2743	0,6718	0,4118
Вопрос 2	49	0,4694	0,6327	0,3121	0,8245	0,5000
Вопрос 3	59	0,4237	0,7288	0,3257	0,7345	0,4727
Вопрос 4	58	0,3966	0,7414	0,2876	0,7821	0,4615
Вопрос 5	59	0,5085	0,6271	0,4123	0,6627	0,5085
Вопрос 6	58	0,3621	0,7759	0,2412	0,6673	0,4400
Вопрос 7	44	0,5455	0,5682	0,4012	0,7412	0,5217
Вопрос 8	43	0,3721	0,7442	0,2387	0,8421	0,4324
Все вопросы	507	0,4300	0,6627	0,3057	0,7254	0,4628

Таблица 16. Результаты тестирования Grok 3
с простым промптом

но снижается — с 0,46 до 0,20 — что говорит о сокращении величины расхождений между оценками модели и преподавателя. При этом показатели Precision остаются низкими во всех сценариях, что означает склонность модели к избыточно «широкому» распределению оценок. Recall, напротив, растёт при усложнении промпта, что свидетельствует о повышении способности модели находить корректные ответы среди всех возможных. F-мера улучшается незначительно, оставаясь около 0,30. Таким образом, данная модель становится более стабильной в плане величины ошибок при использовании продвинутых стратегий промптирования, однако её способность точно различать категории оценок остаётся ограниченной. Несмотря на то, что модель справляется с задачей хуже, чем ChatGPT o3-mini, тем не менее она обеспечивает приемлемые результаты. При этом ее преимуществом является доступность. Локальная версия может работать без подключения к интернету.

6. Модель Grok 3

Grok 3 является продуктом компании xAI, основанной Илоном Маском, и представляет собой третье поколение линейки Grok, ориентированное на улучшенные способности к логическим рассуждениям и самокоррекции. Архитектурные детали и точные параметры модели в открытых источниках не раскрываются, однако заявляется о применении методов обучения с подкреплением (reinforcement learning) для формирования устойчивых цепочек рассуждений (chain-of-thought), обратного анализа шагов (backtracking) и минимизации ошибок фактического характера. Модель интегрируется с системами получения актуальной информации из внешних источников, что повышает её релевантность при ответах на вопросы, требующие знания текущих событий. Модель Grok 3 позиционируется как инструмент для генерации ответов с повышенным уровнем достоверности, предназначенный как для исследовательских задач, так и для массового пользовательского применения.

Результаты тестирования со всеми промптами представлены в таблицах 16, 17, 18 и 19. В таблице 20 приведены общие результаты экспериментов с данной моделью. Для модели Grok 3 наблюдается последовательное улучшение качества оценивания по мере усложнения промпта. При использовании простого промпта Accuracy составила лишь 0,43, а MSE была высокой — 0,66, что указывает

Table 17. Grok 3 testing results with role prompt

Номер вопроса	Кол-во данных	Accuracy	MSE	Precision	Recall	F-мера
Вопрос 1	137	0,4672	0,5474	0,3124	0,6821	0,4706
Вопрос 2	49	0,5306	0,5306	0,3782	0,7171	0,5200
Вопрос 3	59	0,5254	0,4746	0,3825	0,7301	0,5246
Вопрос 4	58	0,4310	0,7241	0,2910	0,8234	0,4815
Вопрос 5	59	0,6102	0,4237	0,4592	0,6714	0,5538
Вопрос 6	58	0,4138	0,6724	0,2857	0,7021	0,4528
Вопрос 7	44	0,5682	0,6136	0,4318	0,7541	0,5532
Вопрос 8	43	0,4419	0,6047	0,3125	0,7561	0,4878
Все вопросы	507	0,4931	0,5680	0,3489	0,7208	0,4970

Таблица 17. Результаты тестирования Grok 3 с ролевым промптом**Table 18.** Grok 3 testing results with the "chain of thoughts" prompt

Номер вопроса	Кол-во данных	Accuracy	MSE	Precision	Recall	F-мера
Вопрос 1	137	0,5401	0,5255	0,4215	0,7283	0,5441
Вопрос 2	49	0,6327	0,4082	0,4921	0,6785	0,5714
Вопрос 3	59	0,6102	0,3729	0,4605	0,6823	0,5538
Вопрос 4	58	0,5172	0,6379	0,3712	0,7205	0,5085
Вопрос 5	59	0,6780	0,3051	0,5128	0,6593	0,5797
Вопрос 6	58	0,5172	0,6379	0,3927	0,6802	0,5085
Вопрос 7	44	0,7273	0,4545	0,5386	0,6651	0,5926
Вопрос 8	43	0,5349	0,5116	0,4105	0,6872	0,5333
Все вопросы	507	0,5838	0,4892	0,4437	0,6947	0,5392

Таблица 18. Результаты тестирования Grok 3 с промптом «цепочка мыслей»**Table 19.** Grok 3 testing results with AI prompt

Номер вопроса	Кол-во данных	Accuracy	MSE	Precision	Recall	F-мера
Вопрос 1	137	0,5839	0,4380	0,5204	0,6653	0,5882
Вопрос 2	49	0,6735	0,4898	0,5125	0,6827	0,5862
Вопрос 3	59	0,6102	0,4237	0,4903	0,6305	0,5538
Вопрос 4	58	0,5517	0,5000	0,4602	0,6115	0,5246
Вопрос 5	59	0,7627	0,3390	0,5408	0,7000	0,6133
Вопрос 6	58	0,5000	0,5000	0,4705	0,5703	0,5172
Вопрос 7	44	0,7727	0,3636	0,5750	0,6500	0,6071
Вопрос 8	43	0,5349	0,4419	0,4820	0,6155	0,5333
Все вопросы	507	0,6154	0,4379	0,5074	0,6444	0,5522

Таблица 19. Результаты тестирования Grok 3 с ИИ промптом**Table 20.** General results of testing Grok 3

Номер вопроса	Accuracy	MSE	Precision	Recall	F-мера
Простой пром프트	0,4300	0,6627	0,3057	0,7254	0,4628
Ролевой пром프트	0,4931	0,5680	0,3489	0,7208	0,4970
Цепочка мыслей	0,5838	0,4892	0,4437	0,6947	0,5392
ИИ пром프트	0,6154	0,4379	0,5074	0,6444	0,5522

Таблица 20. Общие результаты тестирования Grok 3

на существенные расхождения между ответами модели и эталоном. Переход к ролевому пром프트 обеспечил небольшое повышение точности и снижение ошибки, при этом F-мера также улучши-

Table 21. Consistency of answers for simple question (Question 7: What is a file in the office supplies area?)

Модель	Простой промпт	Ролевой промпт	Промпт «цепочка мыслей»	ИИ промпт
ChatGPT o-3	0,6687	0,6917	0,7211	0,7666
DeepSeek-V3	0,6603	0,7045	0,7226	0,7412
Mistral-Small-3	0,6359	0,6779	0,6786	0,6958
Grok 3	0,6769	0,6692	0,6697	0,6953

Таблица 21. Устойчивость ответов для простого вопроса (Вопрос 7: Что такое файл в области канцелярских принадлежностей?)**Table 22.** Consistency of answers for a complex question (Question 6: Which algorithms for sorting an array with a large number of numbers are considered the fastest and why?)

Модель	Простой промпт	Ролевой промпт	Промпт «цепочка мыслей»	ИИ промпт
ChatGPT o-3	0,4883	0,477	0,4948	0,5021
DeepSeek-V3	0,4768	0,4833	0,4857	0,4914
Mistral-Small-3	0,4676	0,4726	0,4842	0,4989
Grok 3	0,4635	0,4751	0,4775	0,4833

Таблица 22. Устойчивость ответов для сложного вопроса (Вопрос 6: Какие алгоритмы сортировки массива с большим количеством чисел считаются самыми быстрыми и почему?)

лась, отражая более уверенное распознавание корректных ответов. Применение техники «цепочка мыслей» значительно улучшило результаты: Ассурасу выросла до 0,58, а F-мера до 0,54, при заметном сокращении MSE. Наилучшие показатели достигнуты при использовании ИИ-промпта — Ассурасу достигла 0,62, MSE снизилась до 0,44, а F-мера выросла до 0,55. Таким образом, Grok 3 особенно выигрывает от более детального и структурированного формулирования запроса, а автоматическое построение промпта моделью приводит к наилучшему балансу точности и устойчивости оценивания.

7. Устойчивость ответов моделей

Для оценки устойчивости ответов моделей использовался коэффициент Флейсса (Fleiss Kappa). Он представляет собой статистическую меру согласованности, применяемую для оценки степени согласия между несколькими независимыми наблюдателями (или источниками оценок) при классификации объектов по дискретным категориям. В контексте языковых моделей данный коэффициент может использоваться для измерения устойчивости ответов путём анализа согласованности между результатами нескольких запусков модели на идентичных запросах. Расчет данного коэффициента на основе десяти экспериментов для каждой комбинации модели и промпта представлен в таблицах 21 и 22. В таблице 21 представлены результаты экспериментов с наиболее простым вопросом (Вопрос 7: Что такое файл в области канцелярских принадлежностей?), для которого модели продемонстрировали наивысшие метрики качества, в то время как в таблице 22 представлены результаты экспериментов с самым сложным вопросом (Вопрос 6: Какие алгоритмы сортировки массива с большим количеством чисел считаются самыми быстрыми и почему?), с которым модели справлялись хуже всего. Для простого вопроса все модели продемонстрировали существенный уровень согласованности ответов, в то время как для сложного вопроса уровень согласованности оказался умеренным.

Заключение

По результатам экспериментов по оценке открытых ответов учащихся с помощью моделей ChatGPT o3-mini, DeepSeek V3, Mistral-Small-3.1-24B-Instruct-2503-IQ4_XS и Grok 3 с различными вариантами промптов лучшее качество оценки ответов показала модель ChatGPT o3-mini со сгенерированным ей же проммптом. Доля правильных ответов (ассурасу) составила 0,82, среднеквадра-

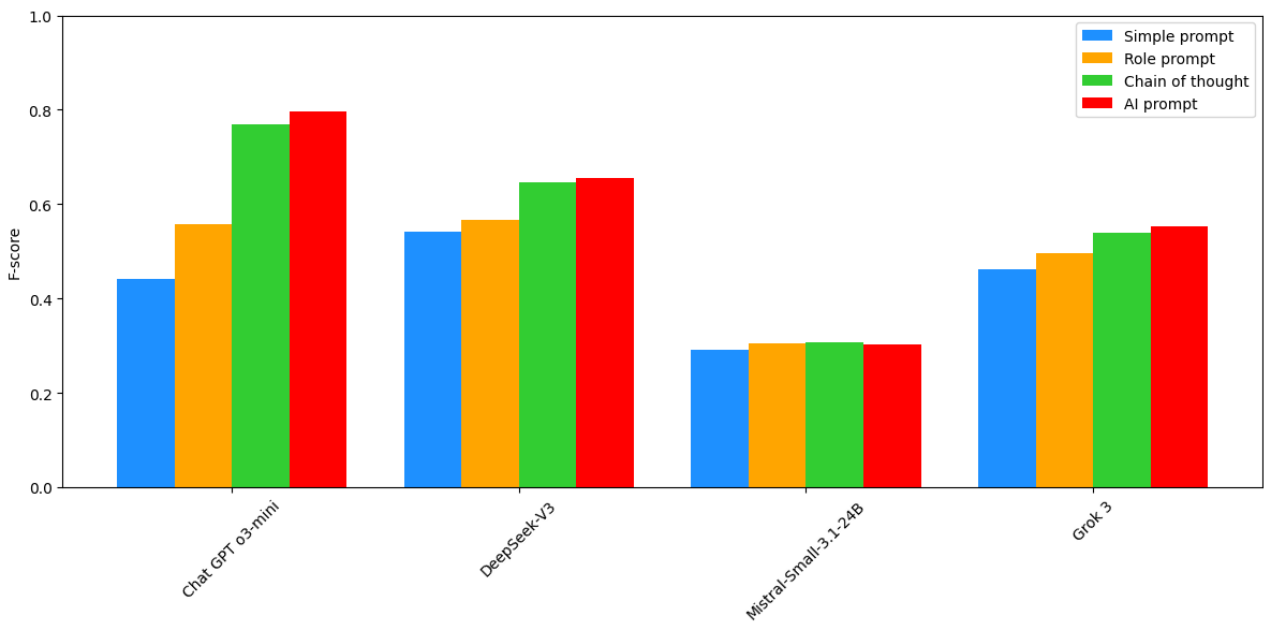


Fig. 3. General results of experiments with different models and prompts.

Рис. 3. Общие результаты экспериментов с разными моделями и промптами

тичная ошибка (MSE) оказалась равна 0,2, а F-мера достигла 0,8, что подтверждает перспективность использования ИИ не только в качестве инструмента оценки, но и в качестве средства автоматической генерации инструкций. Коэффициент Флейсса для данной пары модели и промпта составляет от 0,48 для сложных вопросов до 0,69 для простых вопросов, что в среднем указывает на умеренно-существенный уровень согласованности оценок модели. Общие результаты экспериментов представлены на рисунке 3.

Целью дальнейших исследований является поиск новых подходов к созданию промптов, которые бы могли улучшить качество оценок моделей, а также систематизация и формализация таких подходов.

References

- [1] S. Khan, L. Blessing, and Y. Ndiaye, "Artificial intelligence for competency assessment in design education: A review of literature", in *International Conference on Research into Design*, Springer, 2023, pp. 1047–1058. DOI: [10.1007/978-981-99-0428-0_85](https://doi.org/10.1007/978-981-99-0428-0_85).
- [2] J. Lu, B. K. Balasubramanian, M. Joy, and Q. Xu, "Survey and analysis for the challenges in computer science to the automation of grading systems", *ACM Computing Surveys*, p. 3 748 521, 2025. DOI: [10.1145/3748521](https://doi.org/10.1145/3748521).
- [3] L. Yan *et al.*, "Practical and ethical challenges of large language models in education: A systematic scoping review", *British Journal of Educational Technology*, vol. 55, no. 1, pp. 90–112, 2024. DOI: doi.org/10.1111/bjet.13370.
- [4] L. Kaldaras, H. O. Akazeze, and M. D. Reckase, "Developing valid assessments in the era of generative artificial intelligence", in *Frontiers in Education*, vol. 9, 2024, p. 1 399 377. DOI: [10.3389/feduc.2024.1399377](https://doi.org/10.3389/feduc.2024.1399377).
- [5] L. S. Lo, "The art and science of prompt engineering: A new literacy in the information age", *Internet Reference Services Quarterly*, vol. 27, no. 4, pp. 203–210, 2023. DOI: [10.1080/10875301.2023.2227621](https://doi.org/10.1080/10875301.2023.2227621).

- [6] J. Park and S. Choo, “Generative AI prompt engineering for educators: Practical strategies”, *Journal of Special Education Technology*, p. 01 626 434 241 298 954, 2024. DOI: [10.1177/01626434241298954](https://doi.org/10.1177/01626434241298954).
- [7] C. H. Leung, “Promoting optimal learning with ChatGPT: A comprehensive exploration of prompt engineering in education”, *Asian Journal of Contemporary Education*, vol. 8, no. 2, pp. 104–114, 2024. DOI: [10.55493/5052.v8i2.5101](https://doi.org/10.55493/5052.v8i2.5101).
- [8] L. S. Lo, “The CLEAR path: A framework for enhancing information literacy through prompt engineering”, *The Journal of Academic Librarianship*, vol. 49, no. 4, p. 102 720, 2023. DOI: [10.1016/j.acalib.2023.102720](https://doi.org/10.1016/j.acalib.2023.102720).
- [9] C. Jin *et al.*, “Apeer: Automatic prompt engineering enhances large language model reranking”, in *Companion Proceedings of the ACM on Web Conference 2025*, 2025, pp. 2494–2502. DOI: [10.1145/3701716.3717574](https://doi.org/10.1145/3701716.3717574).
- [10] A. Gilson *et al.*, “How does ChatGPT perform on the United States medical licensing examination (USMLE)? the implications of large language models for medical education and knowledge assessment”, *JMIR Medical Education*, vol. 9, no. 1, e45312, 2023. DOI: [10.2196/45312](https://doi.org/10.2196/45312).
- [11] L. Morjaria *et al.*, “Examining the efficacy of ChatGPT in marking short-answer assessments in an undergraduate medical program”, *International Medical Education*, vol. 3, no. 1, pp. 32–43, 2024. DOI: [10.3390/ime3010004](https://doi.org/10.3390/ime3010004).
- [12] T. Jade and A. Yartsev, *ChatGPT for automated grading of short answer questions in mechanical ventilation*, 2025. DOI: [10.48550/arXiv.2505.04645](https://doi.org/10.48550/arXiv.2505.04645). arXiv: [2505.04645 \[cs.CL\]](https://arxiv.org/abs/2505.04645).
- [13] L. Henrickson and A. Meroño-Peñuela, “Prompting meaning: A hermeneutic approach to optimising prompt engineering with ChatGPT”, *AI & SOCIETY*, vol. 40, no. 2, pp. 903–918, 2025. DOI: [10.1007/s00146-023-01752-8](https://doi.org/10.1007/s00146-023-01752-8).
- [14] G. Kortemeyer, “Performance of the pre-trained large language model GPT-4 on automated short answer grading”, *Discover Artificial Intelligence*, vol. 4, no. 1, p. 47, 2024. DOI: [10.1007/s44163-024-00147-y](https://doi.org/10.1007/s44163-024-00147-y).
- [15] J. Flodén, “Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT”, *British Educational Research Journal*, vol. 51, no. 1, pp. 201–224, 2025. DOI: [10.1002/berj.4069](https://doi.org/10.1002/berj.4069).
- [16] A. V. Rezaev and N. D. Tregubova, “ChatGPT and AI in the universities: An introduction to the near future”, *Higher Education in Russia*, vol. 32, no. 6, pp. 19–37, 2023, in Russian. DOI: [10.31992/0869-3617-2023-32-6-19-37](https://doi.org/10.31992/0869-3617-2023-32-6-19-37).
- [17] P. A. A., “Potentials of integrating generative artificial intelligence technologies into formative assessment processes in higher education”, *Vestnik Majkopskogo Gosudarstvennogo Tehnologiceskogo Universiteta*, vol. 16, no. 2, pp. 98–109, 2024, in Russian. DOI: [10.47370/2078-1024-2024-16-2-98-109](https://doi.org/10.47370/2078-1024-2024-16-2-98-109).
- [18] A. Kong *et al.*, “Better zero-shot reasoning with role-play prompting”, in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2024, pp. 4099–4113. DOI: [10.18653/v1/2024.naacl-long.228](https://doi.org/10.18653/v1/2024.naacl-long.228).
- [19] J. Wei *et al.*, “Chain-of-thought prompting elicits reasoning in large language models”, *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.
- [20] B. Wang *et al.*, “Towards understanding chain-of-thought prompting: An empirical study of what matters”, in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 2717–2739.

- [21] G. Chochlakis, N. M. Pandiyan, K. Lerman, and S. Narayanan, “Larger language models don’t care how you think: Why chain-of-thought prompting fails in subjective tasks”, in *Proceedings of the 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2025, pp. 1–5.
- [22] J. Cheng *et al.*, *Chain-of-thought prompting obscures hallucination cues in large language models: An empirical evaluation*, 2025. arXiv: [2506.17088](#) [[cs.CL](#)].
- [23] Z. Li, Y. Du, J. Hu, X. Wan, and A. Gao, *Self-instructed derived prompt generation meets in-context learning: Unlocking new potential of black-box LLMs*, 2024. arXiv: [2409.01552](#) [[cs.CL](#)].
- [24] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha, *A systematic survey of prompt engineering in large language models: Techniques and applications*, 2024. arXiv: [2402.07927](#) [[cs.AI](#)].
- [25] X. Wang and D. Zhou, “Chain-of-thought reasoning without prompting”, *Advances in Neural Information Processing Systems*, vol. 37, pp. 66 383–66 409, 2024.