Научная статья
УДК 004.65
https://doi.org/10.17586/2713-1874-2025-4-52-67

ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА ПОДДЕРЖКИ КАДРОВЫХ РЕШЕНИЙ: АДАПТАЦИЯ МОДЕЛЕЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ РЕКРУТИНГА В СФЕРЕ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

Софья Александровна Александрова[1], Диана Андреевна Бармина[2]✉, Кирилл Георгиевич Глушков[3]

[1,2]Университет ИТМО, Санкт-Петербург, Россия
[3]ООО «ТМХ ТМ», Санкт-Петербург, Россия
[1]alexandrova_sophie@mail.ru
[2]diana.barmina.2003@mail.ru✉
[3]kirill6b@gmail.com
Язык статьи – русский

Аннотация: В статье предложен подход к решению одной из ключевых проблем цифровой экономики: эффективного подбора высококвалифицированных ИТ-специалистов. Для повышения качества и скорости оценки кандидатов в ИТ-рекрутинге предлагается к рассмотрению интеллектуальная система поддержки кадровых решений, адаптированная с применением методов LoRA и GRPO. Цель исследования – разработка и валидация концепции интерпретируемого ИИ-ассистента, способного автоматизировано оценивать технические компетенции соискателей с высокой точностью анализа кода и логической согласованностью рассуждений при минимальных вычислительных требованиях. Предложенный подход снижает когнитивную нагрузку на HR-специалистов, минимизирует субъективность отбора и позволяет инженерным командам сосредоточиться на экспертной оценке кандидатов, предварительно отобранных системой. В качестве базовой модели выбрана Qwen2.5-Coder-7B-Instruct, оптимизированная для анализа кода и технических вопросов. Применены методы LoRA для эффективной тонкой настройки и GRPO для улучшения логической согласованности ответов. Проведено сравнительное тестирование дообученной модели с базовой версией и более крупными аналогами (Qwen2.5-Coder-14B, QWQ-32b). Результаты показали улучшение технической корректности (+6,84%) и глубины анализа (+5,52%), а также снижение дисперсии оценок. Разработанный подход позволяет внедрять систему даже на маломощных GPU (T4), что делает его доступным для стартапов и небольших компаний.

Ключевые слова: ИТ-рекрутинг, оценка кандидатов, технические собеседования, тонкая настройка, языковые модели, GRPO, LoRA, Qwen2.5

Ссылка для цитирования: Александрова С. А., Бармина Д. А., Глушков К. Г. Интеллектуальная система поддержки кадровых решений: адаптация моделей искусственного интеллекта для рекрутинга в сфере информационных технологий // Экономика. Право. Инновации. – 2025. – Т. 13. – № 4. – С. 52–67. – https://doi.org/10.17586/2713-1874-2025-4-52-67

INTELLIGENT HR DECISION SUPPORT SYSTEM: ADAPTATION OF ARTIFICIAL INTELLIGENCE MODELS FOR INFORMATION TECHNOLOGY RECRUITING

Sofia A. Aleksandrova[1], Diana A. Barmina[2]✉, Kirill G. Glushkov[3]

[1,2]ITMO University, Saint Petersburg, Russia
[3]TMH TM LLC, Saint Petersburg, Russia
[1]alexandrova_sophie@mail.ru
[2]diana.barmina.2003@mail.ru✉
[3]kirill6b@gmail.com
Article in Russian

Abstract: This paper presents a methodological framework aimed at tackling a critical challenge in the digital economy: the efficient recruitment of highly qualified IT professionals. To improve the quality and speed of candidate evaluation in IT recruiting, we propose to consider an intelligent system for supporting personnel decisions, adapted using the LoRA and GRPO methods. The aim of this research is to develop and validate the concept of an interpretable AI assistant capable of automatically evaluating candidates' technical competencies with high code-analysis accuracy and logically

coherent reasoning, while imposing minimal computational requirements. The proposed solution reduces cognitive load on HR specialists, minimizes subjectivity in the selection process, and enables engineering teams to focus their expertise on in-depth evaluation of candidates pre-screened by the system. The Qwen2.5-Coder-7B-Instruct model was selected as the base architecture due to its code analysis capabilities. LoRA was applied for parameter-efficient fine-tuning, while GRPO enhanced logical reasoning in model responses. Comparative testing was conducted between the fine-tuned model, its base version, and larger counterparts (Qwen2.5-Coder-14B, QWQ-32b). Results demonstrated improvements in technical correctness (+6.84%) and analysis depth (+5.52%), along with reduced score variance. The proposed solution is deployable on low-resource GPUs (T4), making it accessible for startups and small businesses.

Keywords: IT recruitment, candidates' evaluation, technical interviews, fine-tuning, language models, GRPO, LoRA, Qwen2.5

For citation: Aleksandrova S. A., Barmina D. A., Glushkov K. G. Intelligent HR Decision Support System: Adaptation of Artificial Intelligence Models for Information Technology Recruiting. *Ekonomika. Pravo. Innovacii.* 2025. Vol. 13. No. 4. pp. 52–67. (In Russ.). https://doi.org/10.17586/2713-1874-2025-4-52-67

Введение. Сфера информационных технологий (ИТ) демонстрирует устойчивый тренд развития в условиях цифровой экономики, что подтверждается количественными и качественными показателями роста отрасли. В рамках государственной поддержки стартапов и внедрения инновационных решений, ИТ-сфера становится ключевым драйвером цифровизации бизнес-процессов и экономической активности в целом. Однако несмотря на значительные темпы развития, одной из наиболее острых проблем, ограничивающих потенциал отрасли, является дефицит высококвалифицированных специалистов. Нехватка кадров негативно влияет на масштабирование проектов, внедрение новых технологических решений и конкурентоспособность отечественных компаний. Согласно исследованиям [1], проведенным в рамках анализа состояния рынка труда и подготовки ИТ-кадров, спрос на специалистов в области ИТ растет год за годом. Так, на начало 2023 года дефицит ИТ-кадров в России составлял около 1 млн человек, а к 2027 году эта цифра может достигнуть 2 млн. Это связано с повсеместным внедрением информационно-коммуникационных технологий (ИКТ), оцифровкой бизнес-процессов и цифровыми трансформационными изменениями в экономической деятельности российских предприятий. Работодатели все чаще нуждаются в сотрудниках, обладающих высоким уровнем цифровых компетенций, поскольку интеллектуализация труда выходит на первый план, постепенно вытесняя физический труд. Таким образом, в цифровой экономике ИКТ станет неотъемлемой частью практически всех существующих рабочих мест, следовательно, цифровые навыки в ближайшее время потребуются для каждого вида работ и доля таких рабочих мест приблизится к 90% [2].

В эпоху искусственного интеллекта (ИИ) требования к компетенциям будут смещаться от «твёрдых» технических навыков к «мягким» навыкам, таким как адаптивность, способность к обучению, коммуникация и лидерство, поскольку ИИ берет на себя рутинные и вычислительные задачи, а ценность человеческого капитала все больше определяется способностью взаимодействовать с ИИ и решать нестандартные, творческие и социально-ориентированные задачи [3].

Так за 2023–2025 годы компетентностная модель современного ИТ-специалиста изменилась: к классическим языкам программирования, алгоритмам и знанию баз данных добавились ИИ-грамотность, cloud-native/DevOps-практики и компетенции безопасной разработки. Эффект массового внедрения ИИ оценило ежегодное исследование Stack Overflow 2025 [4]. Опрос показал, что 84% разработчиков уже используют или планируют использовать ИИ-инструменты (у профессиональных разработчиков 51% – ежедневно), причем сферы применения выходят за рамки автодополнения кода. По данным JetBrains [5], регулярное использование ИИ стало нормой, однако доверие к точности ИИ-инструментов снижается. В опросе Stack Overflow 2025 [6] лишь 29% респондентов заявили, что доверяют точности ИИ-выходов, тогда как 46% активно не доверяют, что является прямым аргументом в пользу проверки умений чтения, отладки и тестирования кода у кандидатов. Исследования показывают, что разработчики, использующие ИИ-инструменты, нередко пишут менее безопасный код и переоценивают его безопасность, а значит,

остаются критичны компетенции безопасной разработки и объективной валидации решений [7]. Поэтому, все еще является актуальным тестирование ключевых технических компетенций, таких как знание современных языков программирования, понимание алгоритмов и структур данных, навыки работы с базами данных, а также владение методологиями разработки ПО, так как они являются основой профессиональной подготовки, обеспечивающей соответствие специалиста требованиям рынка труда и профессиональное взаимодействие с ИИ-инструментами.

На практике наблюдается несоответствие между требованиями работодателей и фактическим уровнем квалификации как выпускников, так и действующих специалистов, что усугубляется значительным оттоком ИТ-кадров и высокой текучестью персонала [8]. Это обуславливает необходимость внедрения объективных механизмов оценки квалификации при найме ИТ-кадров на базе ИИ, так как в долгосрочной перспективе внедрение ИИ в управлении персоналом перестанет быть опциональным трендом и станет необходимым условием устойчивого развития и конкурентоспособности любой организации [9].

По данным исследования [1], с сентября 2021 года по май 2023 года количество проектов по разработке и внедрению решений в сфере ИТ, реализуемых стартапами, получившими государственную поддержку, увеличилось со 103 до 706 единиц, то есть примерно в 7 раз. Рост числа стартапов свидетельствует о повышении интереса к инновационным направлениям и развитию собственных технологических решений внутри страны. Одновременно с этим стартапы сталкиваются с рядом ограничений, особенно на ранних этапах своего развития. Одним из ключевых факторов, сдерживающих эффективное развитие таких компаний, является ограниченность ресурсов в сфере управления персоналом (HR-ресурсов). Стартапы зачастую не имеют возможности выделить достаточное количество времени и финансов на тщательный отбор кандидатов, в том числе проведение многоэтапного интервью и профессионального тестирования, а привлечение внешнего рекрутмента сопряжено с большими издержками и рисками субъективной оценки кандидатов, недостаточной экспертизы в узкопрофильных технических областях и, как следствие, подбором специалистов, не соответствующих реальным требованиям вакансии [10]. Это делает особенно актуальным автоматизацию процессов тестирования кандидатов по ключевым техническим знаниями навыкам.

Современные методы поиска и привлечения ИТ-кадров должны включать:

– автоматизированные системы скрининга резюме;

– профессиональное тестирование технических компетенций;

– оценку soft skills и способности к обучению;

– анализ портфолио реализованных проектов.

Также автоматизация позволяет значительно сократить время на подбор персонала, повысить точность оценки квалификации кандидатов и минимизировать влияние человеческого фактора в процессе найма. Для стартапов это особенно важно, поскольку они нуждаются в оперативном привлечении квалифицированных специалистов, способных быстро включиться в проектную работу и внести вклад в развитие продукта.

Таким образом, решение проблемы кадрового дефицита и обеспечение оперативного поиска высококвалифицированных специалистов с релевантным опытом становится критически важным условием дальнейшего развития ИТ-сферы и цифровой экономики в целом.

На практике объективная оценка процессов в сфере ИТ-рекрутинга осложняется ростом сложности кадровых процессов. Это связано с увеличением числа технических вакансий, предъявляющих высокие и специализированные требования, а также значительным числом кандидатов, не обладающих достаточной квалификацией и необходимым уровнем знаний.

Современные HR-процессы сталкиваются с рядом существенных проблем.

1) Низкое качество предварительного отбора кандидатов, связанное с ограниченными ресурсами HR-отделов, которые вынуждены оценивать большое количество резюме в условиях временного давления для достижения KPI компании, что приводит к снижению точности оценки компетенций соискателей или целенаправленному росту собеседований заведомо неподходящих кандидатур.

2) Высокая степень субъективности при интерпретации результатов тестирования и собеседований, что негативно влияет на объективность отбора и может приводить к дискриминации по несущественным признакам.

3) Значительные задержки в передаче информации между HR-менеджерами и техническими руководителями, вызванные медленными или недостаточно автоматизированными коммуникационными цепочками и высокой загруженностью последних, что часто приводит к утрате потенциально подходящих кандидатов, особенно в условиях высокой конкуренции за квалифицированные кадры.

4) Особые сложности для малых компаний и стартапов, сталкивающихся с высокой потребностью в квалифицированных специалистах, но имеющих ограниченные ресурсы для построения полноценной системы подбора персонала.

В этих условиях возникает необходимость внедрения интеллектуальных решений, способных обеспечить оперативную и объективную оценку кандидатов, снизить нагрузку на HR-специалистов и ускорить процесс принятия решений по кандидатам.

При этом важно учитывать, что вопросы справедливости и законности применения искусственного интеллекта в рекрутинге остаются предметом активных дискуссий. Как отмечается в статье [11], ключевые правовые риски связаны с алгоритмической дискриминацией, нарушением конфиденциальности персональных данных и отсутствием прозрачности в принятии решений. В России, несмотря на отсутствие специализированного закона, регулирующего ИИ в сфере рекрутинга, работодатели и разработчики автоматизированных систем оценки кандидатов обязаны соблюдать:

1) Закон № 152-ФЗ «О персональных данных», требующий согласия кандидатов на обработку их информации и минимизации её объема.

2) Трудовой кодекс РФ, запрещающий дискриминацию по полу, возрасту и другим нерелевантным критериям (ст. 3 ТК РФ), что особенно актуально при использовании ИИ, решения которого сложно объяснить.

3) Кодекс этики в сфере ИИ, разработанный Минэкономразвития, который рекомендует проверку алгоритмов на предвзятость и сохранение человеческого контроля над итоговыми решениями.

Литературный обзор. Существующие решения на рынке подбора персонала включают в себя ассистентов на основе ИИ, которые могут помочь отобрать кандидатов с помощью анализа резюме и создать опросник, включающий вопросы о коммуникативных навыках и опыте работы человека и позволяющий отсеять незаинтересованных или конфликтных соискателей.

В статье [12] исследуется трансформация рекрутинга под воздействием ИИ-технологий. Авторы, опираясь на теорию ресурсно-ориентированного подхода, доказывают, что ИИ по отчетам Unilever повышает эффективность найма и сокращает время на 50%, снижает затраты до 30% и способствует разнообразию за счет «слепого» отбора, однако отмечают риски алгоритмической предвзятости. Исследование подтверждает необходимость баланса между автоматизацией и человеческим фактором.

В исследовании [13] анализируется эффективность ИИ-инструментов для подбора персонала по анализу резюме в сравнении с традиционными методами. Результаты показали, что ИИ-инструменты в среднем на 20–30% точнее традиционных методов в подборе релевантных профилей, и выявили согласованность мнений экспертов-людей и ИИ-оценки в выборе кандидатов.

Большие языковые модели (БЯМ) используются для генерации разнообразных тестов для оценивания знаний в ИТ-сфере и для дачи обратной связи или пояснений о правильном варианте ответа, что может существенно облегчить работу преподавателей и рекрутеров [14]. Быстрая и простая генерация большого количества вопросов позволяет поддерживать актуальность тестов и избегать повторяющихся вопросов.

Постановка задачи исследования. Традиционные методы, такие как резюме и неструктурированные интервью не позволяют объективно оценить глубину знаний для кадрового специалиста, не обладающего достаточным уровнем экспертизы в узкоспециализированной предметной области. Важно учитывать, что сфера ИТ-рекрутинга обладает следующими особенностями:

– специалистам по подбору персонала

желательно разбираться в технологиях, языках программирования, фреймворках и т.д., в ином случае компетенции по найму передаются уже работающим в компании командам по разработке;

– активно используются GitHub, Stack Overflow, данные хакатонов и ИТ-сообществ;

– используются технические собеседования, тестовые задания, live-кодинг.

Система должна учитывать особенности узкоспециализированных областей, где даже опытные рекрутеры сталкиваются с трудностями в интерпретации ответов кандидатов на типичные вопросы, например, о работе с микроконтроллерами, об алгоритмах машинного обучения или о проектировании распределенных архитектур. Ключевым требованием выступает способность ассистента моделировать логику анализа решений, характерную для Senior-разработчиков. Это предполагает комплексную оценку не только корректности кода, но и таких аспектов, как качество архитектурных решений, эффективность с точки зрения производительности и соответствие лучшим общепринятым практикам в разработке программного обеспечения.

Цель исследования – разработка и валидация концепции интерпретируемого ИИ-ассистента, способного автоматизировано оценивать технические компетенции соискателей с высокой точностью анализа кода и логической согласованностью рассуждений при минимальных вычислительных требованиях.

Для достижения поставленной цели требуется реализация двух взаимодополняющих подходов:

1) обучение модели на репрезентативных эмпирических данных, обеспечивающее адекватность предсказаний;

2) обеспечение интерпретируемости принимаемых решений: система должна предоставлять не только итоговые оценки, но и развернутые пояснения по каждому критерию, что позволит снизить когнитивную нагрузку на рекрутеров, повысить прозрачность автоматизированного скрининга и усилить доверие HR-специалистов к результатам анализа.

Параллельно требуется решить проблему масштабируемости: традиционная передача найма командам разработки заменяется на инструмент, который позволяет инженерам фокусироваться на экспертной оценке лишь лучших кандидатов, отобранных ИИ-ассистентом на ранних этапах. Это достигается через модульную архитектуру системы, где компоненты анализа кода, обработки естественного языка и генерации отчетов взаимодействуют автономно, но интегрируются в единую платформу.

Методы и материалы исследования. В рамках данного исследования был реализован многоэтапный подход к разработке и оценке системы автоматизированной оценки технических компетенций ИТ-кандидатов на основе малых языковых моделей. Методология включает в себя:

1) Обоснование выбора базовой модели. Для решения задачи анализа технических ответов была выбрана модель Qwen2.5-Coder-7B-Instruct, 7-миллиардная версия из семейства код-ориентированных языковых моделей Qwen2.5, разработанная Alibaba Cloud.

2) Формирование обучающего датасета. Был создан специализированный датасет в формате «вопрос–ответ–оценка», ориентированный на вакансию Python-разработчика. Python демонстрирует устойчивое ускорение распространения и выступает ключевым языком для задач искусственного интеллекта, анализа данных и серверной разработки, что обеспечивает массовость пула кандидатов и подтверждает актуальность данной роли для верификации предлагаемого подхода.

Сбор данных включал:

– синтетическую генерацию с использованием моделей Qwen2.5-Max и DeepSeek-R1 для создания пар «вопрос–эталонный ответ»;

– аннотацию ответов по трёхбалльной шкале (корректность, формат, полнота рассуждения);

– инжекцию негативных примеров: в датасет включены заведомо некорректные ответы, а также фрагменты нерелевантного текста для повышения устойчивости модели к шуму.

Итоговый датасет содержит 2789 записей. Средняя длина вопроса составила 16.1 слова, ответа – 24.0 слова. Все данные представлены в формате ChatML, совместимом с инструкт-файнтюнингом модели Qwen.

1) Для адаптации выбранной модели к задаче оценки технических ответов применены два комплементарных подхода:

– LoRA – параметр-эффективный метод тонкой настройки, при котором обучению подвергаются только низкоранговые матрицы, встроенные в ключевые проекционные слои трансформера;

– для улучшения логической последовательности ответов применён алгоритм GRPO (Group Relative Policy Optimization) – метод обучения с подкреплением, оптимизирующий не абсолютную награду, а относительный порядок между несколькими вариантами ответа. Награда рассчитывалась как взвешенная сумма компонентов: correctness (соответствие эталону), format (соблюдение структуры).

В качестве оценщика (rewards model) использовалась модель Qwen2.5-Coder-14B, что обеспечило высокую корреляцию с экспертным суждением.

Оценка эффективности. Оценка проводилась в два этапа:

– сравнительное тестирование обученной модели (Qwen2.5-Coder-7B + LoRA + GRPO) против базовой (Qwen2.5-Coder-7B) и более крупных аналогов (Qwen2.5-Coder-14B, QWQ-32B) по метрикам: Technical Correctness, Depth of Analysis, Accuracy of Analysis, Overall Score;

– реализация ИИ-ассистента рекрутера посредством Telegram-бота: был запущен тестовый отбор по вакансии Python-разработчика на hh.ru. За 72 часа система обработала 50 кандидатов, автоматически собрав и проанализировав 1527 ответов. Результаты верифицировались тремя независимыми моделями-экспертами.

Результаты исследования. При выборе языковой модели для решения задачи верификации ответов на технические вопросы главным критерием является специализация модели на кодогенерации и технических доменах.

Модели семейства Qwen2.5 демонстрируют значительное превосходство в обработке технических запросов, включая анализ кода и синтез решений, что подтверждается тестами MMLU (Massive Multitask Language Understanding) и HumanEval (HE). Согласно исследованию [15], Qwen2.5-Coder показывает 88.4% на HumanEval и 84.1% на HumanEval+ (HE+) на задачах, связанных с программированием, что на 12% выше, чем у аналогичных моделей с открытым исходным кодом, сравнение характеристик моделей представлено в таблице 1. Важным критерием выбора являлись доступность и лицензирование моделей: Qwen2.5 распространяется под открытой лицензией Apache 2.0, что позволяет свободно использовать и модифицировать модель для исследовательских и коммерческих целей.

Таблица 1

Сравнение моделей для задач кодогенерации

Источник:[15]

Модель	Лицензия	HE	HE+	MBPP	MBPP+	BCB (full/hard)	LCB (pass@1)
DeepSeek-Coder-6.7B-Instruct	пермиссивная (коммерч. доп.)	78.6	70.7	75.1	66.1	35.5/10.1	20.5
Qwen2.5-Coder-7B-Instruct	Apache 2.0	88.4	84.1	83.5	71.7	41.0/18.2	37.6
CodeQwen1.5-7B-Chat	Apache 2.0	86.0	79.3	83.3	71.4	39.6/17.2	20.1
CodeStral-22B	(см. Mistral)	78.1	74.4	73.3	68.2	47.1/20.6	32.9
DeepSeek-Coder-33B-Instruct	пермиссивная (коммерч. доп.)	79.3	68.9	81.2	70.1	46.5/17.6	27.7

В условиях ограниченных вычислительных ресурсов важно выбрать такую модель, чтобы соблюдался баланс между производительностью и вычислительными затратами. После анализа доступных размеров был сделан выбор: модель с 7 миллиардами параметров. 7B-класс обеспечивает сопоставимые с 20-33B открытыми моделями результаты на ряде бенчмарков при существенно меньших требованиях к VRAM/инференсу и лучшей гибкости дообучения, обеспечивая этим возможность адаптации под специфические задачи.

Итоговым выбором стала Qwen/Qwen2.5-Coder-7B-Instruct, улучшенная в разработке кода и синтезе данных, что делает ее современной БЯМ с открытым исходным кодом и возможностями кодирования, сравнимыми с GPT-4.

Qwen2.5-Coder-7B-Instruct также обладает такими преимуществами как:

1) поддержка контекста до 128K токенов, что является критичным для анализа объемных технических ответов,

2) встроенная оптимизация для задач кодогенерации,

3) совместимость с инструментами HuggingFace.

Так как выбранная базовая модель не предназначена напрямую работать с вопросами и ответами на технические вопросы, то ее производительность и качество ответов не соответствует установленным требованиям. Поэтому, основная задача работы – исследовать, как улучшить способность модели к логическому мышлению в контексте генерации кода и проверки ответов на вопросы по коду.

Были рассмотрены следующие подходы к адаптации модели:

1) full fine-tuning (полная донастройка) дает максимальную гибкость, однако запрашивает высокие требования к GPU/IO;

2) LoRA (адаптеры низкого ранга) дает экономию памяти и времени, также легко подбирать rank/целевые слои; однако такой подход чувствителен к выбору модулей и ранга;

3) IA^3/другие PEFT: самый главный плюс – компактно, но предлагает меньше практики в кодовых задачах.

Обучение БЯМ, содержащих несколько миллиардов параметров, сопряжено со значительными вычислительными затратами, поэтому была выбрана LoRA, позволяющая адаптировать модели к конкретным задачам с минимальными вычислительными затратам [16]. Основная идея заключается в том, чтобы изменять только небольшую часть параметров модели, сохраняя исходные веса «замороженными». Для эффективного дообучения модели через LoRA, включая оптимизацию памяти и ускорение вывода выбрана библиотека Unsloth, которая демонстрирует самую высокую скорость по сравнению с другими библиотеками (рисунок 1).

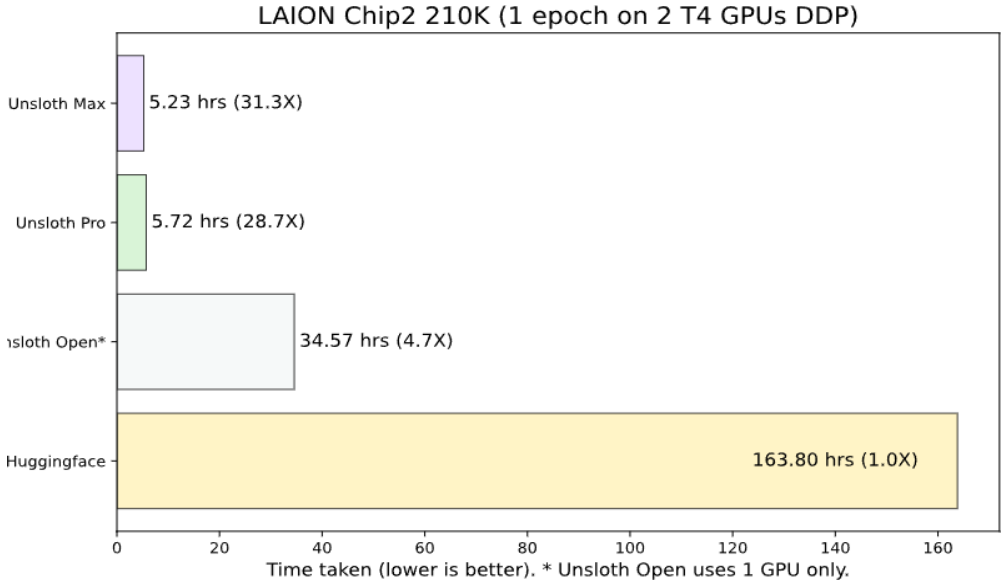

Рисунок 1 – Перформанс тесты на обучение модели с помощью Unsloth

Источник: подготовлен авторами

Даже при использовании оптимальных инструментов эффективность обучения БЯМ в значительной степени зависит от качества и релевантности обучающих данных. На этапе подготовки данных в первую очередь важно отразить в них целевую доменную область – ответы технических специалистов на вопросы тестирования и собеседования.

Для эффективного обучения текстовые данные представлены в токенизируемом формате с четкой структурой: формат «вопрос-ответ», включающий инструкцию, вопрос, ответ и оценку. Согласно рекомендациями из технической документации Unsloth, для достижения приемлемых результатов требуется не менее 100 строк данных, а для максимальной производительности – более 300 строк (увеличение объема данных прямо коррелирует с улучшением итоговых показателей модели).

Для получения объема данных больше 300 строк применялся комбинированный подход, включающий:

– синтетическую генерацию с использованием языковых моделей DeepseekR1 и Qwen2.5max,

– извлечение структурированных данных с платформ проведения технических собеседований,

– контрольные образцы с преднамеренно искаженными паттернами: технические вопросы с нерелевантными ответами и фрагменты поэтических произведений.

Включение заведомо некорректных данных было выполнено для повышения устойчивости модели к шумовым воздействиям и для предотвращения обучения на узких паттернах.

Итоговый датасет по специальности Python-программист (рисунок 2) содержит 2789 строк типа ««вопрос-ответ» + оценка» со следующими характеристиками:

– средняя длина вопроса: 16,1 слов,

– средняя длина ответа: 24,0 слова.

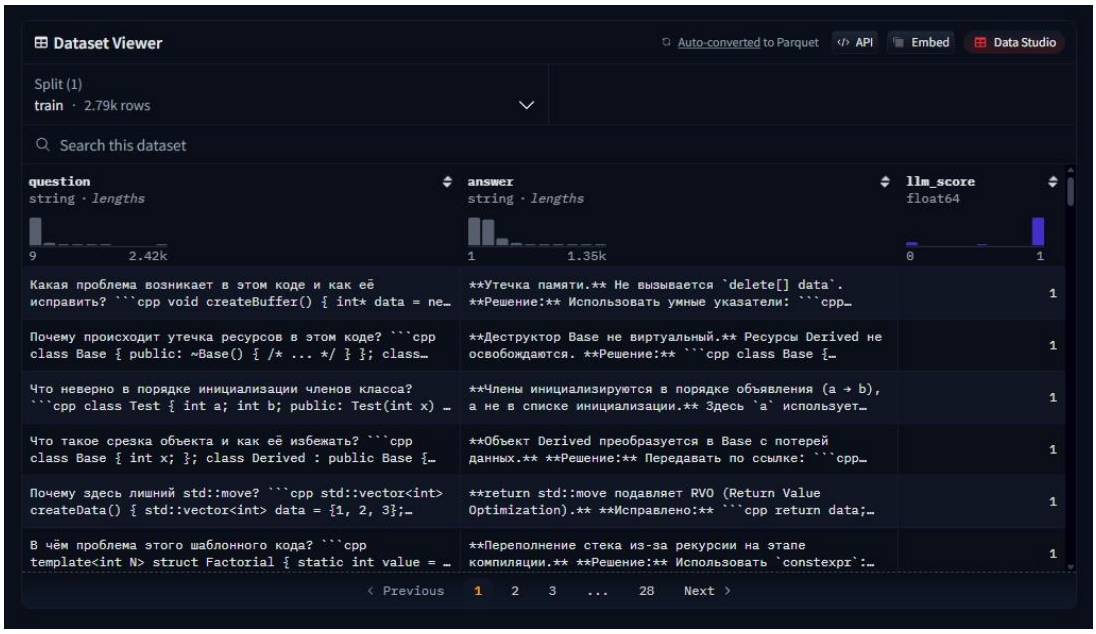

Рисунок 2 – Пример итогового датасета для обучения

Источник: подготовлен авторами

После подготовки датасета был инициирован процесс обучения модели. Для оптимизации использования вычислительных ресурсов, что особо важно для системы с 16 гигабайтами видеопамяти, модель была загружена с 4-битной квантизацией (метод QLoRA), в результате чего требуемый объем памяти сократился в 8 раз, по сравнению с классической 32-битной точностью.

Квантизация весов модели, то есть преобразование параметров из формата с высокой точностью в низкобитовые представления, практически не влияет на качество генерации текста, что подтверждается современными

исследованиями: модель Guanaco, обученная с QLoRA, достигает 99,3% уровня производительности ChatGPT на бенчмарке Vicuna, несмотря на 4-битную квантизацию [17].

Количество обучаемых параметров в LoRA-адаптерах зависит от ранга матрицы разложения: чем выше ранг, тем больше обучаемых параметров, соответственно выше гибкость модели и ее адаптивность к новым задачам [18]. Однако возникают такие ограничения как:

– ресурсные затраты: увеличение объема оперативной памяти примерно на 0,5–1% от общего размера модели для ранга 64, рост вычислительной нагрузки во время обучения в квадратичной зависимости от ранга и увеличение времени сходимости алгоритма обучения;

– проблема переобучения: при рангах, больше 128, наблюдается снижение прироста точности при одновременном увеличении разрыва между метриками на обучающей и валидационной выборках, что указывает на переобучение.

Экспериментальные данные показывают, что ранг 64 является оптимальным и обеспечивает:

1) сохранение 98–99% информативности полной матрицы параметров,

2) ограничение доли обучаемых параметров на уровне 1,44% от общего числа,

3) умеренные требования к памяти - не более 10% от исходного размера модели, и стабильное время обучения, сопоставимое с базовой моделью.

LoRA-адаптеры добавляются к следующим слоям трансформера:

– q_proj, k_proj, v_proj, o_proj (проекционные слои механизма внимания),

– gate_proj, up_proj, down_proj (компоненты FFN-блоков).

Данные слои были выбраны ввиду того, что они содержат более 90% обучаемых параметров модели и играют ключевую роль в адаптации к специфическим задачам.

Помимо QLoRA для минимизации потребления памяти в процессе обучения также применяется градиентный чекпоинт (use_gradient_checkpointing = "unsloth") – это техника, уменьшающая потребление памяти в 3–5 раз за счёт перерасчета промежуточных активаций вместо их хранения [19].

Фиксация случайного начального значения (random_state = 3407) основана на исследовании Дэвида Пикарда о влиянии случайных исходных данных в архитектурах глубокого обучения для компьютерного зрения, где показано, что данное значение обеспечивает стабильность обучения при многократных запусках ($\sigma < 0.01$ для метрик точности) и лучшую сходимость по сравнению с другими сидами (например, 42 или 1337) в контексте БЯМ [20].

Таким образом, в рамках данного подхода обучаются только низкоранговые матрицы, которые впоследствии комбинируются с исходными весами модели.

После применения LoRA все ещё остается ограничение: модель хорошо имитирует стиль и фактологию обучающих примеров, но остаётся чувствительной к «правдоподобно-неверным» ответам и нарушению формата. Поэтому необходима целевая оптимизация под конечную функцию качества – корректность и структурированность ответа на технический вопрос с возможностью автоматической проверки. Для этого используется обучение с подкреплением (RL), где критерий качества задаётся в виде программируемой награды.

Для обучения с подкреплением необходима формализация системы наград, которая обеспечивает обратную связь для оптимизации генерации текста. В данном решении система оценки ответов включает в себя три ключевых этапа:

1) загрузка и предобработка датасета – преобразование исходных данных в диалоговый формат, совместимый с ChatML (стандарт для моделей семейства Qwen2.5),

2) автоматическое выделение метрик качества – llm_score: эталонная оценка, полученная от другой LLM,

3) система наград для обучения модели с подкреплением.

Были рассмотрены следующие RL-подходы:

1) DPO/ORPO (оффлайн предпочтения): простые и устойчивые, но требуют пар «предпочтено/непредпочтено» и хуже используют «градацию» качества нескольких кандидатов для одного запроса;

2) PPO/GRPO (онлайн RL): напрямую оптимизируют заданную награду; PPO оперирует абсолютной наградой, GRPO –

относительным порядком внутри группы ответов, это помогает лучше обобщать и стабилизировать обучение, особенно при шумных или смещенных сигналах награды. В рамках GRPO модель учится не просто выбирать «правильный» ответ, но и ранжировать решения по их относительной логичности и полноте.

Учитывая наличие программируемых проверок и необходимость настроить именно логическую связность рассуждений, был выбран GRPO как основной RL-алгоритм.

GRPO – это алгоритм обучения с подкреплением, специально разработанный для улучшения способности моделей к рассуждениям, что обеспечило актуальность моделей вроде DeepSeek-R1 с высокими показателями в решении задач, требующих логических выводов и анализа. Так, в математических задачах GRPO позволяет напрямую оптимизировать модель на основе правильности ответа, а не субъективных человеческих оценок.

Кроме того, GRPO поощряет модель за генерацию последовательных цепочек рассуждений, а не только за итоговый ответ. Модели без рассуждений могут ошибаться даже на простых вопросах, таких как «сколько будет 9.01 + 9.11» или « сколько букв R в слове strawberry».

Система наград реализует групповую релятивную оптимизацию (GRPO), где параллельно вычисляются несколько компонентов:

– корректность (correctness) – начисление баллов за соответствие эталонному ответу;

– формальная структура (format) – штраф за отклонения от требуемого формата и бессвязный текст вне релевантных тегов <reasoning>...</score>.

Итоговая награда вычисляется как взвешенная сумма компонентов correctness и format.

Параметр в настройке конфига для обучения use_vllm = True позволяет использовать vLLM (Vectorized Large Language Model) на методах эффективного управления кэшированием и параллелизма на GPU для ускорения генерации ответов в 5-10 раз. Об этом свидетельствуют результаты исследования [21], представленные в виде графиков на рисунке 3. Остальные параметры обучения соответствуют рекомендованным из официальной документации от Unsloth.

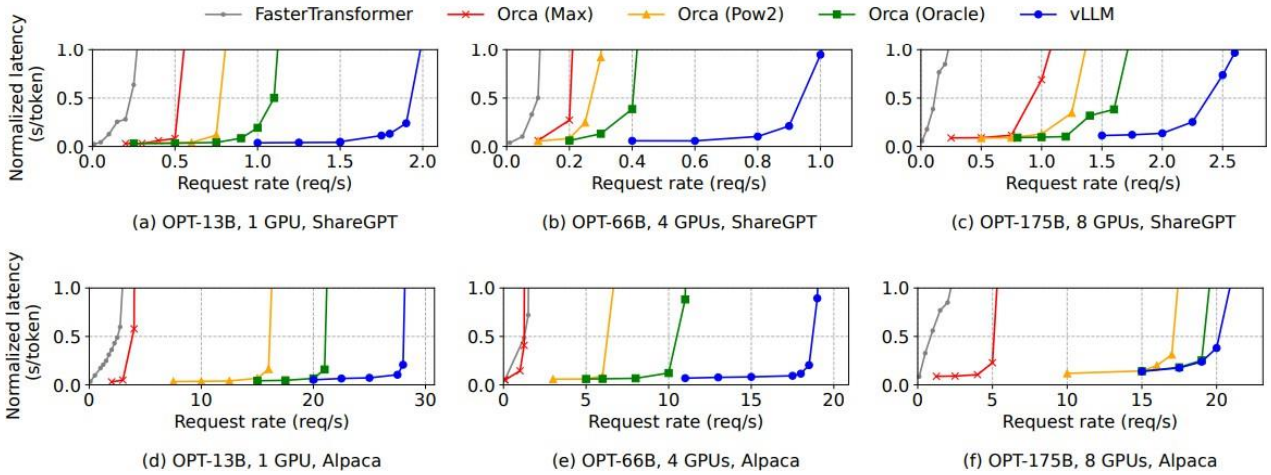

Рисунок 3 – Результаты исследования VLLM

Источник: [21, С. 10]

Обучение завершилось через 2 часа, итоговое потребление памяти составило 12 Гб видеопамяти и 8 Гб оперативной памяти, что соответствует большинству потребительских задач, и, следовательно, модель может быть развернута даже на доступных GPU-ресурсах, таких как T4 в Colab.

Апробация результатов. После реализации тонкой настройки БЯМ, необходимо оценить на реальных данных результаты ее обучения. С этой целью был разработан Telegram-бот для автоматического тестирования технических навыков кандидатов с использование обученной модели, который является прототипом ИИ-ассистента рекрутера.

На платформе Headhunter была размещена тестовая вакансия Python

разработчика с зарплатой от 100 000 рублей до 200 000 рублей с гибридным графиком работы. Такие условия обусловлены популярностью направления и ожидаемым большим потоком кандидатов.

За 72 часа было получено 1102 отклика, из которых 50 человек решили перейти на платформу Telegram и пройти тест в созданном боте. Бот успешно собрал и структурировал данные всех 50 участников за 48 часов, обработав 15270 операций.

Для валидации эффективности дообученной модели с LoRA-адаптером были использованы другие БЯМ с параметрами, в несколько раз превосходящими обученную модель, которые выступили в качестве независимых «экспертов-оценщиков» для верификации оценок от обученной модели. Данный подход соответствует методологии, предложенной в работе «Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena» [22], где предлагается использование БЯМ с большим количеством параметров для обеспечивания более надежной оценки качества ответов и демонстрации высокой корреляции с человеческими суждениями.

Ответы моделей анализируются по нескольким метрикам:

– Accuracy of Analysis (точность анализа) – насколько верны факты и логика в ответе,

– Technical Correctness (техническая корректность) – отсутствие ошибок в технических деталях,

– Depth of Analysis (глубина анализа) – уровень детализации и проработанности ответа,

– Overall Score (общий балл) – средняя оценка за три предыдущие метрики.

Для уменьшения случайности тестирование повторяется 3 раза и берутся средние значения. Для оценки эффективности дообучения модели Qwen2.5-Coder-7B с использованием LoRA-адаптера проведено сравнительное тестирование с базовой версией модели (без адаптера) (рисунок 4). В качестве оценочной модели использована Qwen2.5-Coder-14B, обладающая большим количеством параметров и выступающая в роли эксперта.

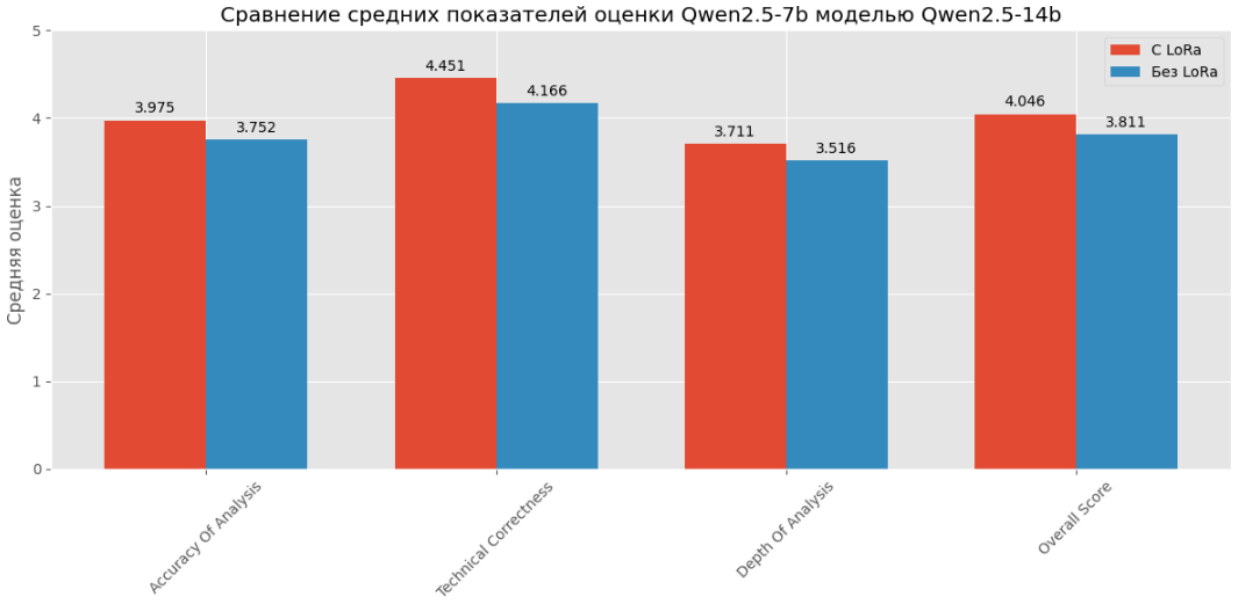

Рисунок 4 – Результаты тестирования модели

Источник: подготовлен авторами

Данный график показывает статистически значимое улучшение модели LoRA-адаптером по всем метрикам:

1) прирост в технической корректности (+6.84%),

2) прирост в глубине анализа (+5.52%).

Также стоит отметить согласованность оценок:

1) все три запуска с LoRA показывают результаты выше среднего, чем без LoRA,

2) максимальный overall_score с LoRA (4.1079) превышает максимальный без LoRA (3.8603) на +6.42%.

График на рисунке 5 демонстрирует сравнение двух моделей (с использованием LoRA и без него). Каждая метрика представлена в виде боксплота, который отражает медиану, квартили и выбросы, а также стандартное отклонение (σ).

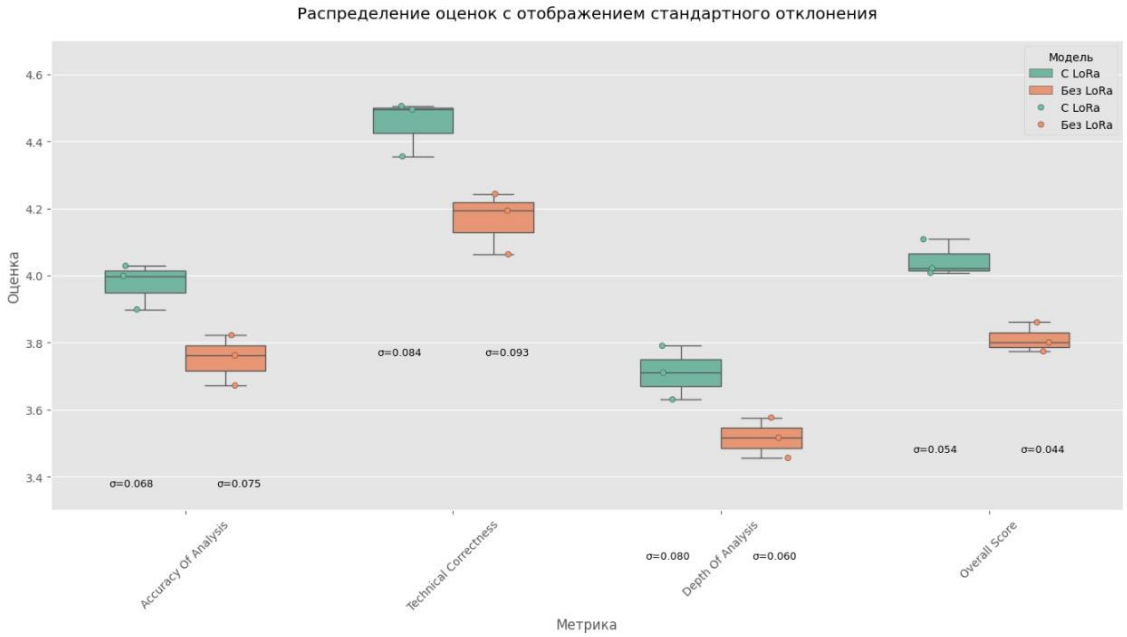

Рисунок 5 – Распределение оценок
Источник: подготовлен авторами

Модель без LoRA-адаптера демонстрирует в среднем меньший разброс показателей стандартного отклонения чем модель с адаптером. Несмотря на то, что в первых двух метриках LoRA уменьшает разброс в среднем примерно на 9,5%, во вторых двух метриках она его значительно увеличивает – на 33% и 23% соответственно. Это свидетельствует о большей нестабильности в результатах модели с адаптером.

Для дополнительной валидации в качестве независимого эксперта была привлечена модель QWQ-32b – это модель среднего размера, которая способна конкурировать по производительности с самыми современными логическими моделями, такими как DeepSeek-R1 и o1-mini. В новое тестирование также была добавлена модель для кросс-проверки Qwen2.5-Coder-14b. Результат представлен на рисунке 6.

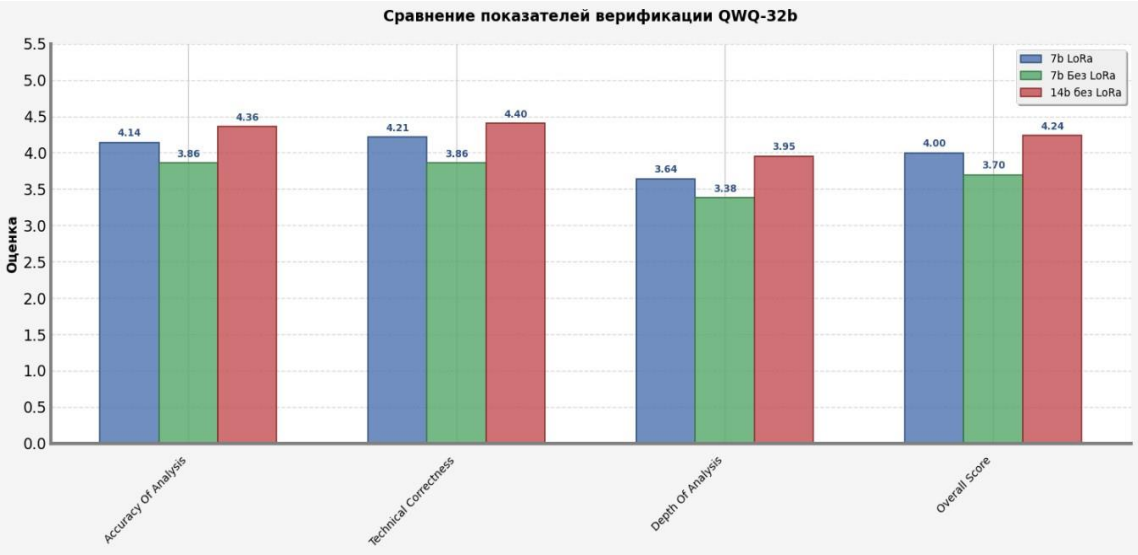

Рисунок 6 – Результаты сравнения верификации QWQ32b
Источник: подготовлен авторами

Qwen2.5-Coder-14b лидирует по всем критериям, что подтверждает преимущество крупных моделей. В то же время LoRA для Qwen2.5-Coder-7B позволяет сократить разрыв с Qwen2.5-Coder-14b, но не закрывает его полностью.

Результаты по Overall Score дополнительно представлены ниже:

1) Qwen2.5-Coder-7b+LoRA – 4.00,
2) Qwen2.5-Coder-7b – 3.70,
3) Qwen2.5-Coder-14b – 4.24,
4) разрыв между Qwen2.5-Coder-7b+LoRA и Qwen2.5-Coder-7b: 8.11%,
5) разрыв между Qwen2.5-Coder-14b и Qwen2.5-Coder-7b+LoRA: 6.00%,
6) разрыв между Qwen2.5-Coder-14b и Qwen2.5-Coder-7b: 14.59%.

Даже с адаптерами малая модель не достигает уровня крупной. Это связано с тем, что LoRA оптимизирует параметры, но не увеличивает их количество. Данное тестирование подчёркивает, что LoRA – не замена крупным моделям, а способ оптимизировать малые.

Выводы и рекомендации. Проведенное исследование демонстрирует эффективность адаптации моделей искусственного интеллекта для рекрутинга в сфере информационных технологий, акцентируя внимание на балансе между качеством результатов и ресурсной эффективностью. Основой работы стала адаптация модели Qwen2.5-7B с использованием LoRA-адаптеров и обучения с подкреплением через алгоритм GRPO. Такой подход позволил не только улучшить ключевые метрики модели, но и сохранить возможность её запуска на потребительском оборудовании (обучение за 2 часа на GPU T4), что критически важно для практического внедрения в системы малого бизнеса или в исследовательские проекты. Внедрение GRPO, в отличие от классического RL, где фокус смещен на максимизацию абсолютной награды, обеспечило ранжирование вариантов ответов по их логичности и полноте, что особенно важно в исследуемой предметной области.

Однако полученные результаты подтверждают, что вышеуказанные методы позволяют малым моделям приближаться по качеству к более крупным аналогам, сохраняя при этом доступность для развёртывания в условиях ограниченных ресурсов. Но следует признать, что даже оптимальная тонкая настройка не компенсирует фундаментального ограничения – объёма параметров.

Проведенный анализ также выявил существенные методологически проблемы. Во-первых, эффективность reward-функций напрямую зависит от качества аннотаций эталонной LLM (llm_score), что требует тщательной валидации самих эталонных данных. Во-вторых, несмотря на использование vLLM для ускорения генерации, обработка больших объемов данных в реальном времени (как в случае с 50 одновременными пользователями бота) потребовала оптимизации конвейера обработки запросов.

Внедрение интеллектуальной системы поддержки кадровых решений позволяет значительно повысить эффективность процессов подбора персонала в условиях высокой загруженности и ограниченных ресурсов. Предложенное решение обеспечивает автоматизированную оценку большого количества кандидатур, что существенно снижает нагрузку на HR-специалистов и исключает необходимость привлечения дорогостоящих технических руководителей на ранних этапах отбора.

Использование моделей искусственного интеллекта делает систему поддержки кадровых решений энергоэффективной и вычислительно недорогой, позволяя задействовать её даже в небольших компаниях и стартапах, не располагающих мощными серверными ресурсами. Таким образом, разработанный подход представляет собой перспективное решение для автоматизации и объективизации процесса оценки кандидатов в сфере ИТ-рекрутинга, способствуя сокращению временных и трудовых затрат, минимизации субъективных факторов и повышению качества найма.

Список источников

1. Васильева Е. В., Каманина А. Н. Дефицит ИТ-кадров в России на современном этапе: причины и пути преодоления // Дискуссия. – 2023. – № 117. – С. 108–118. – DOI: 10.46320/2077-76392023-2-117-108-118. – Текст: электронный.
2. Алексеева Н. В., Сазонов А. А. Анализ степени влияния цифровой экономики на формирование основных трендов на рынке труда и социально-трудовых отношений в Российской Федерации // Вестник Московского государственного областного университета. Серия: Экономика. – 2019. – № 2. – С. 28–36. – DOI: 10.18384/2310-6646-2019-2-28-36. – Текст: электронный.
3. Лукичев П. М. Человеческий капитал в эпоху искусственного интеллекта // Экономика труда. – 2025. – Том 12. – № 5. – С. 683–702. – DOI: 10.18334/et.12.5.123167. – Текст: электронный.
4. Stack Overflow. 2025 Stack Overflow Developer Survey: AI. 2025 [Электронный ресурс]. – URL: https://survey.stackoverflow.co/2025/ai (дата обращения: 06.07.2025). – Текст: электронный. (In Eng.).
5. Bedrina O. The State of Developer Ecosystem 2025: Coding in the Age of AI, New Productivity Metrics, and Changing Realities // The Research Blog (JetBrains). 2025 [Электронный ресурс]. – URL: https://blog.jetbrains.com/research/2025/10/state-of-developer-ecosystem-2025/ (дата обращения: 06.07.2025). – Текст: электронный. (In Eng.).
6. Stack Overflow. Learning #1. AI adoption is accelerating, but trust is declining – 2025 // Stack Overflow Developer Survey [Электронный ресурс]. – URL: https://stackoverflow.co/internal/resources/2025-stack-overflow-developer-survey-for-leaders/ai-adoption/ (дата обращения: 06.07.2025). – Текст: электронный. (In Eng.).
7. Perry N., Srivastava M., Kumar D., Boneh D. Do Users Write More Insecure Code with AI Assistants? // arXiv. 2022 [Электронный ресурс]. – URL: https://arxiv.org/pdf/2211.03622 (дата обращения: 06.07.2025). – Текст: электронный. (In Eng.).
8. Максимов М. И., Роговая А. Е. Развитие системы управления персоналом в условиях оттока ИТ-специалистов и высокой текучести кадров // Региональная и отраслевая экономика. – 2024. – № 3. – С. 131–140. – DOI: 10.47576/2949-1916. 2024.3.3.018. – Текст: электронный.
9. Шляпов И. В., Титовнина Е. И., Гурушкин П. Ю. Технологии искусственного интеллекта в цифровых HR- коммуникациях: перспективы и риски // Litera. – 2025.– № 2. – DOI: 10.25136/2409-8698.2025.2.73198. – Текст: электронный.
10. Ребрикова Н. В. Особенности использования рекрутинга в системе управления персоналом организации // Инновации и инвестиции. – 2021. – № 5. – С. 79–82.

References

1. Vasil'eva E. V., Kamanina A. N. The Shortage of Professional Staff in Russia at the Present Stage: Causes and Ways to Overcome. *Diskussiya*. 2023. No. 117. pp. 108–118. (In Russ.). DOI: 10.46320/2077-76392023-2-117-108-118.
2. Alekseeva N. V., Sazonov A. A. Analysis of the Degree of Influence of the Digital Economy on the Formation of the Main Trends in the Labor Market and Social and Labor Relations in the Russian Federation. *Vestnik Moskovskogo gosudarstvennogo oblastnogo universiteta. Seriya: Ekonomika*. 2019. No. 2. pp. 28–36. (In Russ.). DOI: 10.18384/2310-6646-2019-2-28-36.
3. Lukichev P. M. Human Capital in the Era of Artificial Intelligence. *Ekonomika truda*. 2025. No. 12. (5). pp. 683–702. (In Russ.). DOI: 10.18334/et.12.5.123167.
4. Stack Overflow. 2025 Stack Overflow Developer Survey: AI. 2025. Available at: https://survey.stackoverflow.co/2025/ai (Accessed: 06.07.2025).
5. Bedrina O. The State of Developer Ecosystem 2025: Coding in the Age of AI, New Productivity Metrics, and Changing Realities. *The Research Blog (JetBrains)*. 2025. Available at: https://blog.jetbrains.com/research/2025/10/state-of-developer-ecosystem-2025/ (Accessed: 06.07.2025).
6. Stack Overflow. Learning #1. AI adoption is accelerating, but trust is declining – 2025. *Stack Overflow Developer Survey*. Available at: https://stackoverflow.co/internal/resources/2025-stack-overflow-developer-survey-for-leaders/ai-adoption/ (Accessed: 06.07.2025).
7. Perry N., Srivastava M., Kumar D., Boneh D. Do Users Write More Insecure Code with AI Assistants? *arXiv*. 2022. Available at: https://arxiv.org/pdf/2211.03622 (Accessed: 06.07.2025).
8. Maksimov M. I., Rogovaya A. E. Development of HR Management System in the Conditions of Outflow of IT Specialists and High Staff Turnover. *Regional'naya i otraslevaya ekonomika*. 2024. No. 3. pp. 131–140. (In Russ). DOI: 10.47576/2949-1916. 2024.3.3. 018.
9. Shlyapov I. V., Titovnina E. I., Gurushkin P. Yu. Artificial Intelligence Technologies in Digital HR Communications: Prospects and Risks. *Litera*. 2025. No. 2. (In Russ). DOI: 10.25136/2409-8698. 2025.2.73198.
10. Rebrikova N. V. Features of Using Recruitment in the Organization's Personnel Management System. *Innovatsii i investitsii*. 2021. No. 5. pp. 79–82. (In Russ.).

11. Новиков Д. А. Использование искусственного интеллекта при найме работников: проблемы и перспективы правового регулирования // Journal of Digital Technologies and Law. – 2024. – № 2 (3). – С. 611–635.– DOI: 10.21202/jdtl.2024.31. – Текст: электронный.
12. El Ouakili, O. The Impact of Artificial Intelligence (AI) on Recruitment Process // Open Journal of Business and Management. – 2025. – Vol. 13. – P. 749–762. – DOI:10.4236/ojbm.2025.132039. – Текст: электронный. (In Eng.).
13. Slaykovskiy V., Zvegintsev M., Sakhonchyk Y., Ajamian H. Evaluating AI Recruitment Sourcing Tools by Human Preference // arXiv. 2024 [Электронный ресурс]. – URL: https://arxiv.org/html/2504.02463v1 (дата обращения: 06.07.2025). – Текст: электронный. (In Eng.).
14. Hadzhikoleva S., Rachovski T., Ivanov I., Hadzhikolev E., Dimitrov G. Automated Test Creation Using Large Language Models: A Practical Application // Applied Sciences. – 2024. – No. 14 (19). – P. 9125. – DOI:10.3390/app14199125. – Текст: электронный. (In Eng.).
15. Hui B., Yang J., Cui Z., Yang J. et al. Qwen2.5-Coder Technical Report // arXiv. 2024 [Электронный ресурс]. – URL: https://arxiv.org/html/2409.12186v1 (дата обращения: 06.07.2025). – Текст: электронный. (In Eng.).
16. Edward J. Hu, Yelong Shen, Phillip Wallis et al. LoRA: Low-Rank Adaptation of Large Language Models // arXiv. 2021 [Электронный ресурс]. – URL: https://arxiv.org/abs/2106.09685 (дата обращения: 06.07.2025). – Текст: электронный. (In Eng.).
17. Dettmers T., Pagnoni A., Holtzman A., Zettlemoyer L. QLORA: Efficient Finetuning of Quantized LLMs // arXiv. 2023 [Электронный ресурс]. – URL: https://arxiv.org/pdf/2305.14314 (дата обращения: 06.07.2025). – Текст: электронный. (In Eng.).
18. Wang Z., Liang J., He R. et al. LoRA-Pro: Are Low-Rank Adapters Properly Optimized? // arXiv. 2025 [Электронный ресурс]. – URL: https://arxiv.org/abs/2407.18242 (дата обращения: 06.07.2025). – Текст: электронный. (In Eng.).
19. Chen T., Xu B., Zhang C., Guestrin C. Training Deep Nets with Sublinear Memory Cost // arXiv. 2016 [Электронный ресурс]. – URL: https://arxiv.org/abs/1604.06174 (дата обращения: 06.07.2025). – Текст: электронный. (In Eng.).
20. Picard D. Torch.manual_seed (3407) is All You Need: On the Influence of Random Seeds in Deep Learning Architectures for Computer Vision. arXiv. 2021 [Электронный ресурс]. – URL: https://arxiv.org/abs/2109.08203 (дата обращения: 06.07.2025). – Текст: электронный. (In Eng.).

11. Novikov D.A. The Use of Artificial Intelligence in Hiring Employees: Problems and Prospects of Legal Regulation. *Journal Of Digital Technologies and Law*. 2024. No. 2 (3). pp. 611–635. (In Russ.). DOI: 10.21202/jdtl.2024.31.

12. El Ouakili, O. The Impact of Artificial Intelligence (AI) on Recruitment Process. *Open Journal of Business and Management.* 2025. Vol. 13. pp. 749–762. DOI:10.4236/ojbm.2025.132039.

13. Slaykovskiy V., Zvegintsev M., Sakhonchyk Y., Ajamian H. Evaluating AI Recruitment Sourcing Tools by Human Preference. *arXiv. 2024*. Available at: https://arxiv.org/html/2504.02463v1 (Accessed: 06.07.2025).

14. Hadzhikoleva S., Rachovski T., Ivanov I., Hadzhikolev E., Dimitrov G. Automated Test Creation Using Large Language Models: A Practical Application. *Applied Sciences*. 2024. No. 14 (19). P. 9125. DOI:10.3390/app14199125.

15. Hui B., Yang J., Cui Z., Yang J. et al. Qwen2.5-Coder Technical Report. *arXiv. 2024.* Available at: https://arxiv.org/html/2409.12186v1 (Accessed: 06.07.2025).

16. Edward J. Hu, Yelong Shen, Phillip Wallis, et al. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv. 2021*. Available at: https://arxiv.org/abs/2106.09685 (Accessed: 06.07.2025).

17. Dettmers T., Pagnoni A., Holtzman A., Zettlemoyer L. QLORA: Efficient Finetuning of Quantized LLMs. *arXiv. 2023*. Available at: https://arxiv.org/pdf/2305.14314 (Accessed: 06.07.2025).
18. Wang Z., Liang J., He R. et al. LoRA-Pro: Are Low-Rank Adapters Properly Optimized? *arXiv. 2025*. Available at: https://arxiv.org/abs/2407.18242 (Accessed: 06.07.2025).

19. Chen T., Xu B., Zhang C., Guestrin C. Training Deep Nets with Sublinear Memory Cost. *arXiv. 2016* Available at: https://arxiv.org/abs/1604.06174 (Accessed: 06.07.2025).

20. Picard D. Torch.manual_seed (3407) is All You Need: On the Influence of Random Seeds in Deep Learning Architectures for Computer Vision. *arXiv. 2021*. Available at: https://arxiv.org/abs/2109.08203 (Accessed: 06.07.2025).

21. Kwon W., Li Z., Zhuang S. et al. Efficient Memory Management for Large Language Model Serving with Paged Attention. arXiv. 2023 [Электронный ресурс]. – URL: https://arxiv.org/abs/2309.06180 (дата обращения: 06.07.2025). – Текст: электронный. (In Eng.).
22. Zheng L., Chiang W.-L., Sheng Y., et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv. 2023 [Электронный ресурс]. – URL: https://arxiv.org/abs/2306.05685 (дата обращения: 06.07.2025). – Текст: электронный. (In Eng.).

21. Kwon W., Li Z., Zhuang S. et al. Efficient Memory Management for Large Language Model Serving with Paged Attention. *arXiv. 2023*. Available at: https://arxiv.org/abs/2309.06180 (Accessed: 06.07.2025).

22. Zheng L., Chiang W.-L., Sheng Y., et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv. 2023*. Available at: https://arxiv.org/abs/2306.05685 (Accessed: 06.07.2025).

Статья поступила в редакцию 11.07.2025; одобрена после рецензирования 05.12.2025; принята к публикации 12.12.2025.
The article was submitted 11.07.2025; approved after reviewing 05.12.2025; accepted for publication 12.12.2025.