СЕКЦИЯ 3
ИИ-ТЕХНОЛОГИИ В РАЗРАБОТКЕ ПРИКЛАДНЫХ ПРОГРАММНЫХ ПРОДУКТОВ

М. М. Ершов, С. А. Жилин, М. Н. Перевалова
Тюменский государственный университет, г. Тюмень
УДК 004.94

РАЗРАБОТКА ИНТЕЛЛЕКТУАЛЬНОГО ПОМОЩНИКА ПО АНАЛИЗУ ДАННЫХ ДЛЯ ЦИФРОВОЙ КАФЕДРЫ ТюмГУ

Аннотация. Статья описывает разработку Telegram-бота для автоматизации ответов на вопросы студентов в рамках самостоятельного обучения по учебным материалам. Решение интегрирует OCR и LLM в чат-бот систему, сокращая время ответов до 5 минут, снижая нагрузку на тьюторов и повышая эффективность самостоятельного обучения.

Ключевые слова: интеллектуальный помощник, анализ данных, обработка данных, LLM, самостоятельное обучение.

Введение. В настоящее время образовательные услуги активно трансформируются благодаря цифровым технологиям, открывая новые возможности для самостоятельного обучения и персонализированного консультирования. Однако рост онлайн-форматов в процессе обучения выявил ряд проблем: продолжительное время ответа на вопрос студента, отсутствия пояснений к ответам.

Это приводит к снижению эффективности учебного процесса, увеличению нагрузки на преподавателей и ограничивает возможности своевременного освоения материала, особенно в дисциплинах, требующих оперативного взаимодействия.

Для решения обозначенных проблем необходимо разработать информационную систему, обеспечивающую:

Оперативное получение ответов за счет интеграции оптического распознавания текста (OCR) [1] и больших языковых моделей (LLM) [2, 3];

Генерацию пошаговых решений на основе анализа контекста запросов с использованием LLM;

Сбор аналитических данных через автоматизированную базу данных, фиксирующую метрики качества ответов и ошибки системы.

Оптимизация управления памятью [4] и алгоритмическая обработка данных [5] позволяют сократить время ответов до 5 минут, снизить нагрузку на преподавателей и обеспечить соответствие современным стандартам цифровизации образования.

Актуальность. Разработка интеллектуального помощника является актуальной задачей для улучшения качества самостоятельного обучения.

Опрос студентов в социальной сети с помощью Google Forms показал, что идея интеллектуального помощника востребована. Среди проходящих курс по анализу данных от ТюмГУ студентов положительно ответили 86,7% опрошенных.

Цель. Разработать информационную систему в виде Telegram-бота, направленную на устранение задержек обратной связи при самостоятельном обучении. Решение обеспечивает:

1. Автоматическую генерацию ответов на вопросы студентов с задержкой менее 5 минут;
2. Распознавание текста из изображений с использованием OCR для обработки графических запросов;
3. Сбор аналитических данных через автоматизированную базу данных.

Задачи. Для достижения цели проект реализуется в несколько этапов:

1. Предобработка данных.
2. Разработка графического модуля для распознавания текста с изображений на основе выявленных шаблонов и разметки вопроса.
3. Проектирование базы данных для автоматизации обработки запросов.
4. Сравнение и обучение наилучших моделей больших языковых моделей (LLM) на основе определенных метрик.
5. Разработка пользовательских интерфейсов.

Этап 1. Предобработка данных. На начальном этапе обработаны данные, включающие текстовые вопросы, варианты ответов и ответы студентов, представленные в формате .xlsx.

Для обеспечения качества данных выполнены очистка от пустых значений, корректировка грамматических ошибок, валидация структуры вопросов, проверка логичности финальных ответов.

Результатом этапа стал структурированный датасет, используемый для обучения LLM.

Этап 2. Графический модуль. Данный этап направлен на автоматизацию обработки визуальных запросов. На вход модуля поступают скриншоты вопросов (Рис. 1), структура которых стандартизирована:

- Верхняя часть содержит текстовый вопрос;
- Варианты ответов сопровождаются графическими маркерами:
  - *Круги* — для единственного выбора;
  - *Квадраты* — для множественного выбора.
- Текст представлен в упрощенном формате для минимизации ошибок распознавания.

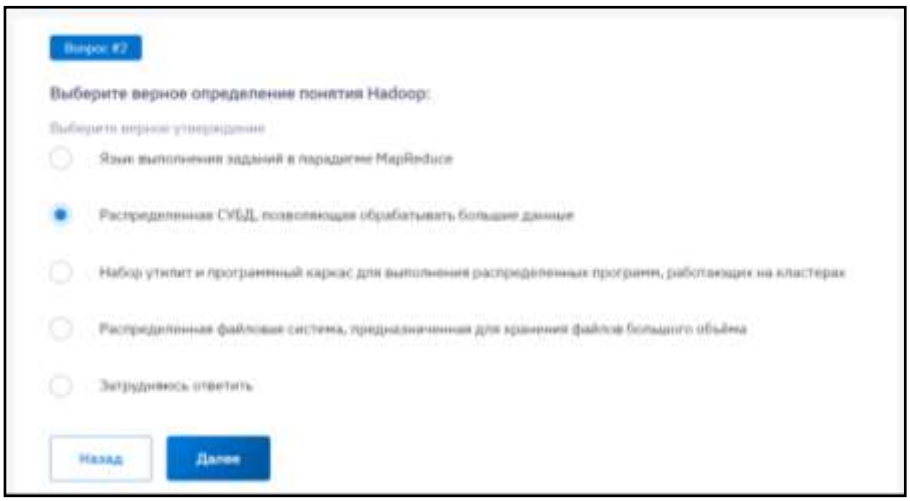

Рис. 1. Пример изображения, для шаблона которого разрабатывался графический модуль

Модуль графики выполняет задачу обработки изображения поступающего и разметки вопроса, если вопрос относится к типам вопросов единственного или множественного выбора ответа.

Классификация происходит на основе графического поиска соответствующих атрибутов для каждого класса напротив вариантов ответов. Полный алгоритм модуля изображен на рис. 2 и 3.

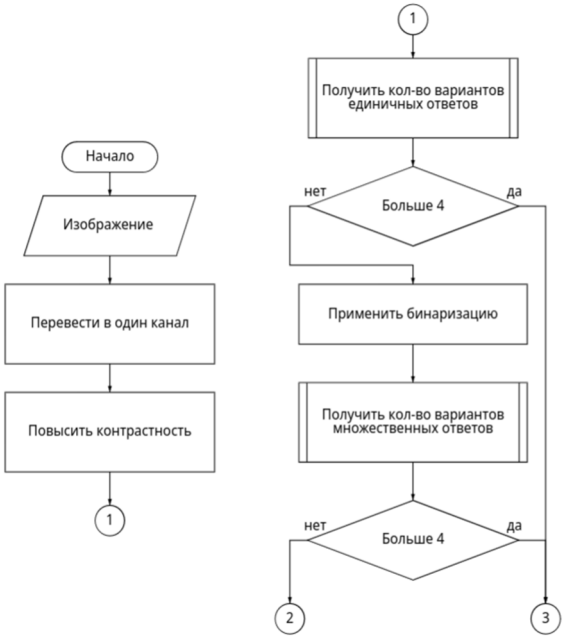

Рис. 2. Первая часть алгоритма работы графического модуля

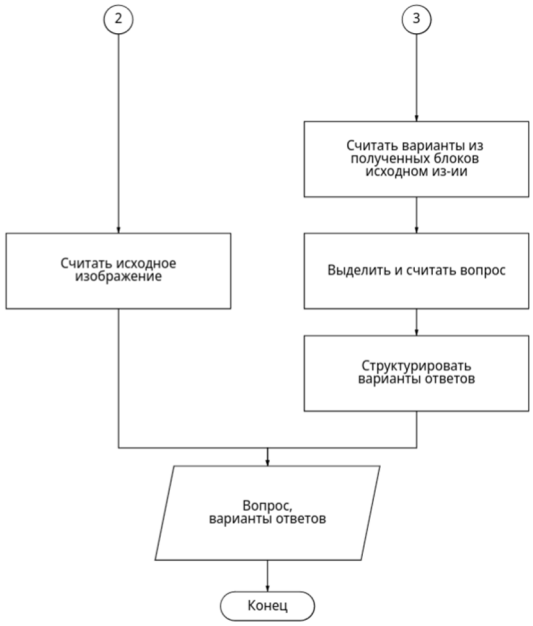

Рис. 3. Вторая часть алгоритма работы графического модуля

Другой задачей было проектирование базы данных для информационной системы.

Этап 3. Проектирование базы данных. В информационной системе существует всего 2 объекта: запрос и фотография. К одному запросу относятся от 0 до нескольких фотографий.

Логическая модель базы данных показана на рис. 4.

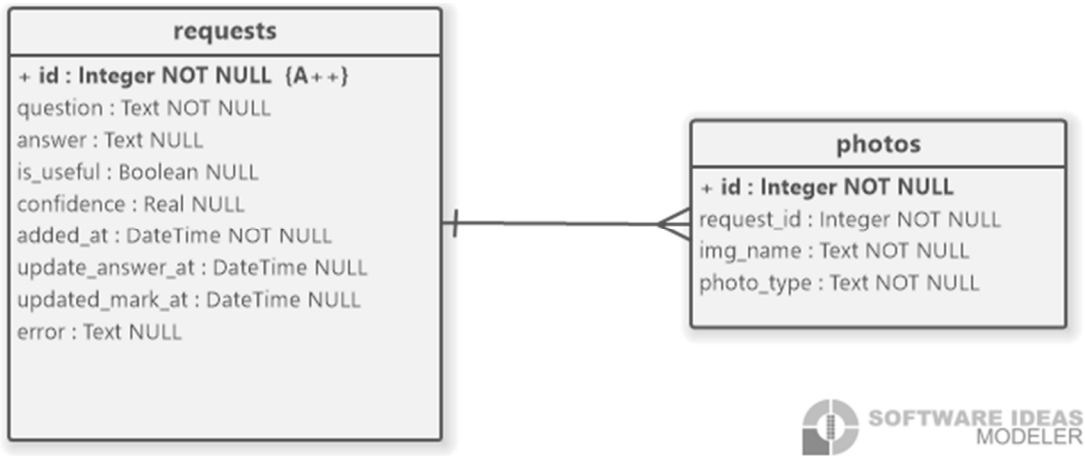

Рис. 4. Логическая модель базы данных

Некоторые поля requests:

- question — весь запрос с считанным с изображений текстом;
- asnwer — ответ, данный LLM;
- is_useful — оценка студента “полезно”/ “не полезно”;
- confidence — BERTScore-оценка, данная LLM;
- error — зафиксированная ошибка и прочие.

Этап 4. Тестирование и выбор LLM. Тестирование производилось с помощью обращения к серверу llama.cpp скриптом, перечисляющим выделенные вопросы доменной зоны или логические. Запросы проводились без ограничения количества токенов в ответе, с температурой генерации 0.5.

Проведено сравнение нескольких моделей LLM (табл. 1) на основе метрик, таких как использование VRAM, точность ответов и точность речи на русском языке. Все показатели кроме оценки BERTScore приведены в количестве сообщений с ошибкой из 20 запросов.

Таблица 1

Результаты сравнения LLM

gguf	*VRAM, GB*	*Кол-во речевых ошибок, шт*	*Кол-во логичных ответов в 1 тесте, шт*	*Кол-во нелогичных ответов в 1 тесте, шт*	*F1 (BERTScore)*
Meta-Llama-3-70B-Instruct-Q4_K_M	43	7	20	0	0,806
vikhr-7b-instruct-0.4.Q6_	6,3	2	12	8	0,788
saiga-llama3-8b.Q5_0	6	5	17	3	0.845
mixtral-8x7b-instruct-v0.1.Q4_K_M	26,3	9	8	12	0.865
t-lite-it-1.0-q8_0.gguf	10	2	16	4	0.915

Была взята для сравнения лучшая по результатам BERTScore T-lite Q8 и T-lite Q4. Результаты скорости генерации приведены в таблице 2.

Таблица 2

Сравнение 2 степеней квантования одной LLM

№ Запроса	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
8 бит, с	11,57	13,46	14,06	15,52	16,77	18,01	19,26	20,51
4 бит, с	15,86	15,50	19,41	20,48	22,25	24,03	25,80	27,58
Разница, %	37	15	38	31	32	33	33	34

Модель T-lite 1.0 продемонстрировала в среднем 32% разницы в производительности токенов в секунду при квантовании.

Этап 5. Разработка пользовательских интерфейсов. Для разработки чат-бота была выбрана платформа Telegram из-за ее популярности мессенджера и наличия удобных инструментов разработки ботов. Для пользователей бот поддерживает текстовые запросы (до 4096

симв.) и изображения (JPG/PNG, 1920x1080 ppx) в базовом интерфейсе Telegram. При ошибках распознавания графики бот отправляет уведомление, а система сохраняет логи для анализа и улучшения работы.

Результаты. На текущем этапе проекта достигнуты следующие результаты:

- Разработан модуль OCR, включая классификацию типов вопросов на основе скриншотов. Тестирование на корректность чтения скриншота на датасетах по 10 изображений каждого класса показало следующие результаты (табл. 3).

Результаты тестирования показывают, что модуль распознает корректно 100% изображений с единственным классом и 80% изображений с множественным классом.

Таблица 3

Результаты тестирования модуля графики

Корректность чтения	*Полностью верное*	*Частично верное*	*Неверное*
Единственный класс	10	0	0
Множественный класс	8	1	1

- Обучена LLM T-lite-0.1 на подготовленных данных.
- Реализован пользовательский интерфейс для студентов на платформе Telegram, поддерживающий текстовые и графические запросы.

Выводы. Разработка интеллектуального помощника для курсов по анализу данных в ТюмГУ представляет собой шаг к улучшению качества самостоятельного обучения. Автоматизация процесса ответов на вопросы студентов позволяет сократить время ожидания, повысить удовлетворенность студентов и снизить нагрузку на преподавателей.

Следующим улучшением чат-бота является внедрение модуля аналитики, который на основе собранных данных (оценка полезности ответов, BERTScore, частота ошибок) улучшать как сам продукт, так и качество исходного курса.

СПИСОК ЛИТЕРАТУРЫ

1. Доронин Ю. Д. Методы обработки изображений и использование компьютерного зрения в OCR // StudNet. — 2021. — № 5. — URL: https://cyberleninka.ru/article/n/metody-obrabotki-izobrazheniy-i-ispolzovanie-kompyuternogo-zreniya-v-ocr (дата обращения: 02.05.2025).
2. Борисов А. Н. Исследование влияния оптимизации инференса LLM на качество LLM // Выпускная квалификационная работа. Национальный исследовательский университет «Высшая школа экономики». — 2025. — URL: https://www.hse.ru/edu/vkr/927268269 (дата обращения: 03.05.2025).
3. Фрицлер А. А. Квантизация нейронных сетей для повышения эффективности расчета прогноза // Выпускная квалификационная работа. Национальный исследовательский университет «Высшая школа экономики». — 2025. — URL: https://www.hse.ru/edu/vkr/471638544 (дата обращения: 03.05.2025).
4. Старушенкова Е. Е. Основные этапы проектирования баз данных // E-Scio. — 2021. — № 11 (62). — URL: https://cyberleninka.ru/article/n/osnovnye-etapy-proektirovaniya-baz-dannyh (дата обращения: 02.05.2025).
5. Kwon W. et al. Efficient Memory Management for Large Language Model Serving with PagedAttention // arXiv. 2023. arXiv:2309.06180 [cs.LG]. — URL: https://arxiv.org/abs/2309.06180 (дата обращения: 03.05.2025).