

ИССЛЕДОВАНИЕ ВОЗМОЖНОСТЕЙ ПРИМЕНЕНИЯ ЦЕНЗУРНОЙ И БОЛЬШОЙ ЯЗЫКОВОЙ МОДЕЛЕЙ В ОБРАЗОВАТЕЛЬНОМ КОМПЛЕКСЕ ВЫСШЕГО И СРЕДНЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ



Д.А. Бурденко, студент,
e-mail: danilburdenkoken@mail.ru
ФГБОУ ВО «Калининградский государственный
технический университет»

А.А. Вихарев, студент,
e-mail: lesha.vikh@mail.ru
ФГБОУ ВО «Калининградский государственный
технический университет»

Кратко рассматриваются ранее существовавшие чат-боты для КГТУ. Проводится обзор проектов со всего мира, в которых применялись цензурные и большие языковые модели, специально дообученные для конкретной задачи высшего или среднего профессионального образования. Описываются состав датасетов и БЯМ, используемые для генерации данных. Перечисляются модели, взятые за основу в каждом проекте, методы к дообучению. Применяемые метрики для оценки качества и их числовые значения. Сделан вывод о возможности применения цензурной и большой языковой моделей в университетском комплексе «Калининградский Государственный Технический Университет».

Ключевые слова: *большая языковая модель, цензурная модель, образование, дообучение, метрики, нейросеть, университет.*

ВВЕДЕНИЕ

В настоящее время трудно представить нашу жизнь без нейросетей. Их используют как обычные пользователи, так и различные организации. Распознавание речи, генерация картинок, видео, создание презентаций. Всё это уже активно используется в образовательном процессе.

Остановимся на исследовании моделей для работы с текстом, а именно на цензурных и больших языковых моделях. В обучении языковые модели помогают быстро получить необходимую контактную информацию, информацию о поступлении, подсказать о мероприятиях, об истории университета. Они могут также использоваться как эксперт в одной узкой области для помощи студентам в освоении нового материала. А также нужно следить, чтобы пользователь не получал неприемлемых и недопустимых ответов на провокационные вопросы.

Очень часто существующих моделей для этих задач не хватает, поэтому приходится их дообучать.

ОБЪЕКТ ИССЛЕДОВАНИЯ

Объектом исследования являются проекты, использующие цензурные и большие языковые модели в образовательных организациях со всего мира.

ЦЕЛЬ И ЗАДАЧИ ИССЛЕДОВАНИЯ

Цель исследования: изучить существующие проекты и оценить возможность применения дообученных моделей в университетском комплексе «КГТУ».

Задачи исследования:

1. Найти проекты, которые бы дообучали цензурные и большие языковые модели для образовательных организаций.
2. Изучить используемые в качестве основы модели в этих проектах.
3. Изучить методы дообучения, метрики.
4. Сделать выводы о возможности применения дообученных моделей в университетском комплексе КГТУ.

МЕТОДЫ ИССЛЕДОВАНИЯ

Методы исследования:

1. Систематический обзор. Поиск публикаций и репозиторий.
2. Сравнительный анализ. Сопоставление используемых датасетов, базовых моделей, методов дообучения, метрик эффективности.

ИССЛЕДОВАНИЕ СУЩЕСТВУЮЩИХ ПРОЕКТОВ

В первую очередь рассмотрим проекты, которые ранее были реализованы для КГТУ. Это – «ИИ Чат-бот КГТУ» от авторов данной статьи [1] и «AI-помощник для вузов» от Сбер [2].

Что важно узнать из этих проектов:

1. Работали в Telegram.
2. Использовали готовые большие языковые модели + RAG. (добавление информации о КГТУ в каждый запрос к модели).

Интерфейс и ответы от ботов показан на рисунке.

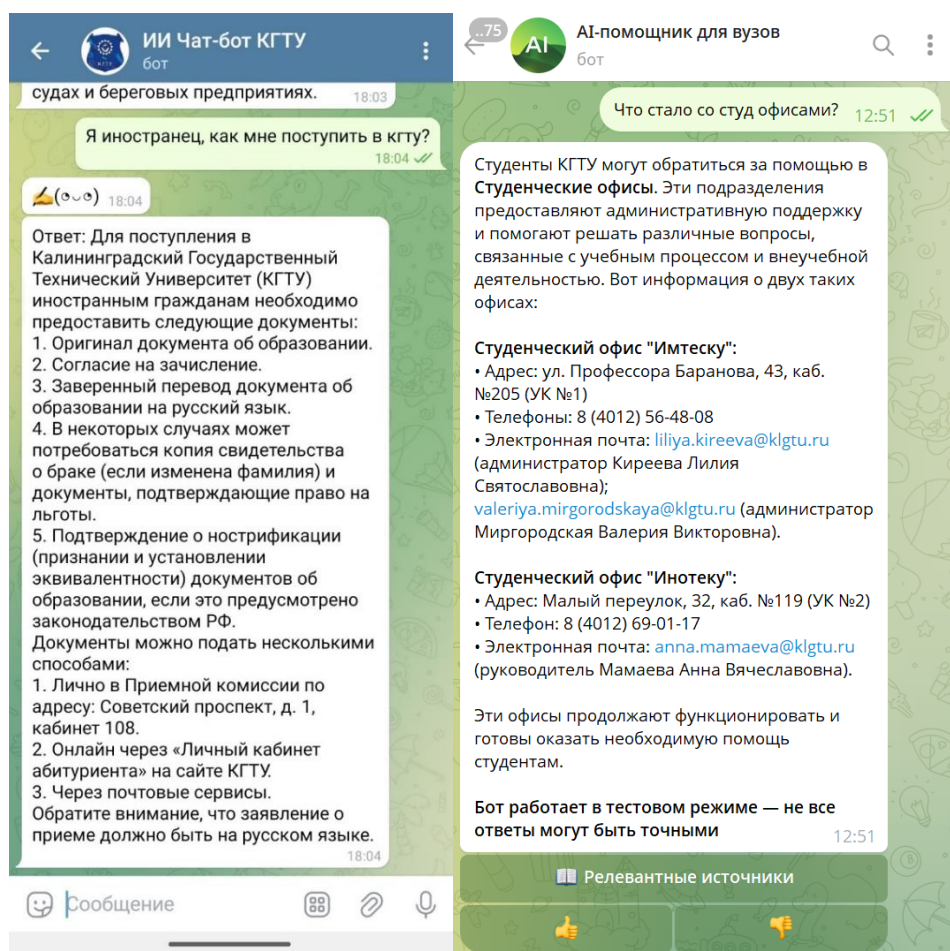


Рисунок – Ранее работавшие чат-боты в Telegram

Ниже рассмотрим проекты из различных стран, в которых дообучали модели для учебных заведений [3-8]. Все они являются либо open-source (проекты с исходным кодом), либо академическими (только статья). В таблице 1 указано, использует ли проект только дообученную модель или ещё генерацию, дополненную поиском (англ. Retrieval Augmented Generation, RAG), а также состав датасетов.

Таблица 1 – Проекты, в которых дообучали БЯМ

Название	Использование RAG	Датасеты
UniFAQ: Fine-Tuned LLM based FAQ Generation for University Admissions	Нет	150 лучших университетов США
StudyAbroadGPT: Fine-tuned LLM for Study Abroad Assistance	Нет	2274 пар диалогов
Fine-Tuning LLM to Answer Basic Questions for Prospective New Students at Syiah Kuala University Using the RAG Method	Да	20 пар вопрос–ответ
Enhancing Admission Inquiry Responses with Fine-Tuned Models and Retrieval-Augmented Generation	Да	8000 пар вопрос–ответ + 210 запросов о поступлении
AI-University: An LLM-based platform for instructional alignment to scientific classrooms (AI University: Finite Element Method)	Да	4648 пар вопрос–ответ
KatzBot: Revolutionizing Academic Chatbot for Enhanced Communication	Да	6280 пар дополнение предложения, 7330 пар вопрос-ответ + 2081 пара вопрос-ответ

В двух проектах известно, какие именно модели использовались для генерации синтетических данных. Так, в «StudyAbroadGPT» датасет состоит из 2274 пар правдоподобных диалогов между студентом и научным руководителем, сгенерированных через API Gemini Pro. А в российском проекте «Enhancing Admission Inquiry Responses» для генерации использовалась китайская модель DeepSeek R1. Для КГТУ помимо обучения на реальных данных, также стоит создать сгенерированный датасет в формате «вопрос-ответ».

В таблице 2 представлены базовые модели и методы их дообучения. В проектах в основном используются модели в диапазоне от 7 до 11 миллиардов параметров. В качестве методов дообучения авторы выбирали LoRA (Low-Rank Adaptation) либо QLoRA (Quantized Low-Rank Adaptation). Их преимуществом является то, что вместо полного переобучения, отдельно обучаются две низкоранговые матрицы. Благодаря этому значительно снижается потребление оперативной памяти (RAM), видеопамяти (VRAM). Также стоит отметить отсутствие в качестве базовых китайских и российских моделей в перечисленных проектах.

Таблица 2 – Используемые базовые модели в проектах

Название	Базовые модели	Метод дообучения
UniFAQ: Fine-Tuned LLM based FAQ Generation for University Admissions	T5-large, BART-large, Mistral 7b, LLaMA-2, LLaMA-3	QLoRA
StudyAbroadGPT: Fine-tuned LLM for Study Abroad Assistance	Mistral-7B-Instruct	LoRA
Fine-Tuning LLM to Answer Basic Questions for Prospective New Students at Syiah Kuala University Using the RAG Method	Mistral 7B	QLoRA
Enhancing Admission Inquiry Responses with Fine-Tuned Models and Retrieval-Augmented Generation	Gemma 9B	LoRA
AI-University: An LLM-based platform for instructional alignment to scientific classrooms (AI University: Finite Element Method)	Llama-3.2-11B	LoRA
KatzBot: Revolutionizing Academic Chatbot for Enhanced Communication	Mistral Instruct, Llama 2, GPT2, Microsoft Phi 1.5	QLoRA

Для КГТУ следует выбирать модели с 7-8 млрд. параметров и поддерживающие русский язык. В качестве метода дообучения лучше выбрать LoRA, для более точных ответов моделей и с возможностью квантизировать их при необходимости.

В таблице 3 представлены метрики качества для каждого проекта и их значения. Только у трёх проектов есть одинаковые метрики - ROUGE-1, ROUGE-2, ROUGE-L. У одного из них значения метрик равно 1.0, поскольку модель дообучалась на слишком малом объёме данных.

Таблица 3 – Метрики качества для каждого проекта

Название	Тип задачи	Метрики	Значения
UniFAQ: Fine-Tuned LLM based FAQ Generation for University Admissions	Генерация часто задаваемых вопросов	ROUGE-1 ROUGE-2 ROUGE-L	0.2418 0.0760 0.2187
StudyAbroadGPT: Fine-tuned LLM for Study Abroad Assistance	Помощь по вопросам в сфере обучения за рубежом	training loss	0.34
Fine-Tuning LLM to Answer Basic Questions for Prospective New Students at Syiah Kuala University Using the RAG Method	Ответы на вопросы о поступлении	ROUGE-1 ROUGE-2 ROUGE-L	1.0 1.0 1.0
Enhancing Admission Inquiry Responses with Fine-Tuned Models and Retrieval-Augmented Generation	Ответы на вопросы о поступлении	Fact Recall Precise Data Recall User Satisfaction	72.7 48.3 7.9
AI-University: An LLM-based platform for instructional alignment to scientific classrooms (AI University: Finite Element Method)	Ответы на вопросы по конкретной учебной теме	Avg. Cos. Sim. Winner Cos. Sim. Judge #1 Judge #2	0.879 86.02% 43.44% 43.23%
KatzBot: Revolutionizing Academic Chatbot for Enhanced Communication	Ответы на вопросы	ROUGE-1 ROUGE-2 ROUGE-L	0.53 0.43 0.51

Для будущей оценки эффективности БЯМ для КГТУ все метрики могут быть полезны, особенно часто применяемые авторами ROUGE.

Цензурная модель необходима для безопасного использования искусственного интеллекта в образовательной среде.

Роль цензурной модели:

1. Предотвращение генерации токсичного контента, дезинформации и вредных советов.
2. Контроль за тем, чтобы модель не выходила за рамки заданных инструкций.
3. Блокировка запросов на создание оружия, вредоносного ПО или химических веществ.

Рассмотрим два проекта, в которых применяются цензурные модели. Это «Unified defense for large language models against jailbreak and fine-tuning attacks in education» [9] и «CoDAE: Adapting Large Language Models for Education via Chain-of-Thought Data Augmentation» [10]. Они направлены на повышение безопасности работы больших языковых моделей и снижение рисков неправильного использования ИИ в образовательной среде. Основной акцент делается на защите от обхода ограничений и повышении надёжности ответов системы.

В качестве датасета в «Unified defense» используется Benchmark EduHarm с парами безопасных и небезопасных инструкций для образовательных сценариев; тесты на разных моделях и видах атак. А в «CoDAE» – реальные учебные диалоги студентов с ИИ и дополнительно сгенерированные примеры с пошаговыми объяснениями.

При сравнении проектов стало ясно, что модели ориентированы на разные цели. Так, «Unified defense» фокусируется на устойчивости модели к jailbreak-атакам и попыткам обхода защиты. Это позволяет сохранить контроль над поведением модели даже при целенаправленных попытках её взлома.

В свою очередь, «CoDAE» больше направлен на улучшение качества образовательного взаимодействия. Он повышает педагогическую полезность ответов, делает объяснения более последовательными и понятными для студентов.

В таблице 4 представлены способы валидации и режимы работы.

Таблица 4 – Сравнение проектов

Название	Способы валидации	Режимы работы системы
Unified defense for large language models against jailbreak and fine-tuning attacks in education	Оценка устойчивости к jailbreak атакам (ASR, compliance rate по benign запросам) и fine-tuning атакам (сниженный ASR при сохранении точности).	Защита выполняется на этапе инференса через трехступенчатый shield-фреймворк (safety-aware attention realignment, layer-wise safety judgment, defense-driven routing) для изучения и фильтрации запросов.
CoDAE: Adapting Large Language Models for Education via Chain-of-Thought Data Augmentation	Проверяется, насколько модель стала лучше объяснять материал, помогать студенту и не выдавать готовый ответ слишком рано. Оценка проводится с помощью автоматических метрик и дополнительной проверки другими моделями.	Сначала модель дообучают на образовательных данных с пошаговыми объяснениями, затем она работает как учебный помощник, который объясняет материал поэтапно и поддерживает диалог со студентом.

Цензурная модель для КГТУ также должна быть устойчива к промпт-инъекциям и различным атакам.

В таблице 5 представлены метрики качества для каждого проекта и их значения. Совпадающих метрик у этих проектов нет, так как они решают разные задачи. «Unified defense» использует метрики безопасности и устойчивости модели к атакам. «CoDAE» оценивается по метрикам качества обучения, таким как Perplexity, Self-BLEU и LLM-as-a-judge.

Таблица 5 – Метрики качества для каждого проекта

Название	Тип задачи	Метрики	Значения
Unified defense for large language models against jailbreak and fine-tuning attacks in education	Защита модели от атак	CR FTA ASR-Jailbreak ASR-Fine-Tuning	93.680 88.500 0.000 14.940
CoDAE: Adapting Large Language Models for Education via Chain-of-Thought Data Augmentation	Адаптация модели для пошаговых объяснений	Self-BLEU LLM-as-judge PPL	Точных значений не предоставлено

Прямое сравнение невозможно, однако обе модели демонстрируют высокую эффективность именно в своей области применения. Совместное использование обеспечит лучшую устойчивость к внешнему воздействию. Описанные метрики помогут оценить эффективность модели для КГТУ.

ЗАКЛЮЧЕНИЕ

В статье были рассмотрены проекты, демонстрирующие активное применение дообученных моделей в образовании. Использование LoRA или QLoRA позволяет получить уникальную модель, специально разработанную для конкретного учебного заведения, на доступном оборудовании.

Применение цензурной и большой языковой моделей в университетском комплексе «Калининградский Государственный Технический Университет» возможно и оправдано. Наиболее предпочтительно было бы выбрать модели с 7-8 млрд. параметров и поддерживающие русский язык. Следует подготовить датасеты, обычно в формате «вопрос-ответ» или аналогичные по структуре, для следования инструкциям. В качестве метода дообучения выбрать LoRA, для более точных ответов моделей и с возможностью квантизировать их при необходимости. Использование дообученной БЯМ + RAG дополнительно повышает точность ответов, за счёт использования свежих данных, которые произошли после обучения. Немаловажно было изучить метрики, которые могут помочь эффективно оценить получившиеся модели.

В рамках продолжения начата реализация проекта по дообучению БЯМ и цензурной модели на данных о КГТУ, для объединения их возможностей в единую систему. Это решение позволит безопасно использовать ИИ в образовании.

СПИСОК ЛИТЕРАТУРЫ

1) Вихарев А. А., Бурденко Д. А. РАЗРАБОТКА ЧАТ-БОТА КГТУ С ИСПОЛЬЗОВАНИЕМ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА // «Дни науки». Материалы Национальной научно-технической конференции студентов и курсантов «Дни науки». «Калининградский государственный технический университет» 14-26 апреля 2025 года [Электронный ресурс] / отв. за вып.: Е. В. Лютова. – Калининград: Издательство БГАРФ ФГБОУ ВО «КГТУ», 2025. С. 305-311.

2) AI-помощник на базе GigaChat для поддержки абитуриентов, студентов и сотрудников вузов // Искусственный интеллект для устойчивого развития. Сборник кейсов. – 2024. – С. 88-89.

3) Github.com: сайт. UniFAQ: Fine-Tuned LLM based FAQ Generation for University Admissions. Режим доступа - свободный. 2026. [Электронный ресурс]. URL: <https://github.com/KayvanShah1/UniFAQ> (дата обращения: 25.04.2026)

4) Github.com: сайт. StudyAbroadGPT: Fine-tuned LLM for Study Abroad Assistance. Режим доступа - свободный. 2026. [Электронный ресурс]. URL: <https://github.com/codermillat/StudyAbroadGPT> (дата обращения: 25.04.2026)

5) Н. Rachmat, Н. Riza, Т. F. Abidin. Fine-Tuning Large Language Model (LLM) to Answer Basic Questions for Prospective New Students at Syiah Kuala University Using the Retrieval-Augmented Generation (RAG) Method. // 2024 Ninth International Conference on Informatics and Computing (ICIC), Medan, Indonesia, 2024 – P. 401-405.

6) А. Virabyan. Enhancing Admission Inquiry Responses with Fine-Tuned Models and Retrieval-Augmented Generation // Параллельные вычислительные технологии – XIX всероссийская научная конференция с международным участием, ПаВТ'2025, г. Москва, 8–10 апреля 2025 г. Короткие статьи и описания плакатов. Челябинск: Издательский центр ЮУрГУ, 2025. С. 99-106.

7) Github.com: сайт. AI University: Finite Element Method. Режим доступа - свободный. 2026. [Электронный ресурс]. URL: <https://github.com/my-ai-university/finite-element-method> (дата обращения: 25.04.2026)

8) Github.com: сайт. KatzBot - Enhancing University Community Communication. Режим доступа - свободный. 2026. [Электронный ресурс]. URL: <https://github.com/AiAI-99/katzbot/> (дата обращения: 25.04.2026)

9) Github.com: сайт. Unified Defense for Large Language Models against Jailbreak and Fine-Tuning Attacks in Education. Режим доступа - свободный. 2026. [Электронный ресурс]. URL: <https://github.com/xinykou/TSSF> (дата обращения: 25.04.2026)

10) Github.com: сайт. CoDAE: Chain-of-Thought Data Augmentation & Evaluation. Режим доступа - свободный. 2026. [Электронный ресурс]. URL: <https://github.com/faerber-lab/CoDAE> (дата обращения: 25.04.2026)

A STUDY OF THE APPLICATION OF CENSORSHIP AND LARGE LANGUAGE MODELS IN THE EDUCATIONAL COMPLEX OF HIGHER AND SECONDARY VOCATIONAL EDUCATION

D.A. Burdenko, student,
e-mail: danilburdenkoken@mail.ru
Kaliningrad State Technical University

A.A. Vikharev, student,
e-mail: lescha.vikh@mail.ru

Previously existing chatbots for KSTU are briefly reviewed. Reviews projects from around the world that utilized censorship and large language models, specifically fine-tuned for a specific task in higher or secondary vocational education. The composition of the datasets and the LLM used to generate the data are described. The models used as the basis for each project, the methods for fine-tuning, and the metrics used for quality assessment and their numerical values are listed. A conclusion was made about the possibility of using censorship and large language models in the university complex "Kaliningrad State Technical University".

Key words: large language model, censorship model, education, fine-tuning, metrics, neural network, university.