Капустин Валерий Викторович
бакалавр, студент
Научный руководитель
Жаркова Оксана Михайловна
канд. физ.-мат. наук, доцент
ФГБОУ ВО «Кубанский государственный университет»
г. Краснодар, Краснодарский край

АДАПТАЦИЯ ЭМБЕДДИНГОВЫХ МОДЕЛЕЙ ПОД УЗКУЮ ПРЕДМЕТНУЮ ОБЛАСТЬ МЕТОДОМ LORA: ОПЫТ ПОВЫШЕНИЯ КАЧЕСТВА СЕМАНТИЧЕСКОГО ПОИСКА В ЦИФРОВОЙ ОБРАЗОВАТЕЛЬНОЙ СРЕДЕ

Аннотация: *статья обобщает практический опыт дообучения эмбеддинговой модели для семантического поиска по учебно-методическим материалам. Показано, почему стандартные поисковые модели неэффективны при работе с профессиональной образовательной лексикой. Описана методика подготовки специализированного датасета из корпоративной базы знаний и процесс низкоранговой адаптации (LoRA) модели T-lite [1]. Приведены конфигурации адаптеров и анализ метрик. Оптимальная конфигурация повышает MRR@10 с 0,35 до 0,53, что критически улучшает контекст для генеративных моделей и снижает уровень галлюцинаций. Сформулированы рекомендации по созданию умных учебных пособий и тьюторов без значительных вычислительных затрат.*

Ключевые слова: *семантический поиск, LoRA, эмбеддинговые модели, цифровая образовательная среда, дообучение, RAG.*

Введение

Цифровая трансформация образования ведёт к накоплению огромных массивов учебно-методического контента: электронных курсов, цифровых библиотек, регламентов и инструкций. Традиционные поисковые системы, основанные на ключевых словах, не улавливают смысловой близости запросов и

документов, особенно при насыщенной профессиональной лексике. Современные архитектуры типа Retrieval-Augmented Generation (RAG) объединяют семантический поиск с генеративными возможностями больших языковых моделей (LLM), обеспечивая точные ответы с опорой на источники [1]. Ключевой компонент такого конвейера – эмбеддинговая модель, преобразующая текст в вектор. Широко распространённые предобученные модели (sentence-transformers) тренированы на общетематических корпусах и резко теряют точность при обработке узкоспециализированной лексики, изобилующей терминами, аббревиатурами и профессиональным сленгом. В образовательной среде этот разрыв критичен: запрос «разобрать примеры решения СЛАУ итерационными методами» может не найти методичку, где используется аббревиатура «СЛАУ», поскольку токенизатор дробит редкие сокращения на случайные субтокены, а модель не обучена устанавливать семантическое тождество разных форм одного понятия [2]. В итоге релевантные документы не попадают в выдачу, и LLM, лишённая корректного контекста, генерирует правдоподобные, но ошибочные ответы.

Полное дообучение модели под конкретный домен – ресурсоёмкая задача с риском катастрофического забывания. Эффективной альтернативой служит Low-Rank Adaptation (LoRA) – метод параметро-эффективной настройки, при котором обучается лишь небольшое число дополнительных параметров, а основные веса остаются замороженными. Это открывает реальную возможность для образовательных организаций создавать собственные специализированные поисковые решения без дорогостоящей инфраструктуры.

Цель настоящей работы – экспериментально проверить, можно ли с помощью LoRA-дообучения на относительно небольшом, но специфичном датасете, созданном на основе материалов цифровой образовательной среды крупного отраслевого учебного центра, существенно повысить качество семантического поиска.

Методология подготовки данных и дообучения

Источником данных послужила закрытая база знаний крупного центра, аккумулирующая технические инструкции, лекционные материалы и внутренние регламенты. Аналогичная задача стоит перед любым вузом, обладающим корпоративным порталом или цифровой библиотекой учебных курсов. Методика подготовки датасета масштабируема и может быть воспроизведена без существенных модификаций.

Процесс состоял из нескольких этапов. Первоначально из системы управления знаниями были экспортированы текстовые документы, прошли фильтрацию от повреждённых и мусорных данных, после чего разбивались на чанки – группы предложений с небольшим перекрытием, что позволяло сохранить семантическую связность каждого фрагмента. Далее для каждого чанка с помощью заранее настроенной LLM генерировалось несколько вопросов и ответов, извлекаемых строго из текста этого же чанка. Такой подход гарантирует, что обучающие пары «вопрос – ответ» являются аутентичными фрагментами изучаемого корпуса, а не внешними знаниями модели.

Ключевым моментом стало формирование Hard-Negative триплетов – троек вида «Вопрос : Правильный ответ : Неправильный, но семантически близкий ответ». Для каждого вопроса подбирался отрицательный пример, который по косинусному сходству находился близко к правильному, но относился к другому контексту. Такой приём заставляет модель учиться более тонко различать смыслы, что особенно важно в условиях высокой терминологической насыщенности учебных текстов.

Итоговый набор данных составил 6000 триплетов, которые были разделены на обучающую (4800 примеров, 80%) и тестовую (1200 примеров, 20%) выборки. Принципиальным требованием являлось отсутствие пересечений между выборками: все вопросы из тестовой части относились к документам, не использовавшимся при обучении. Каждый текстовый элемент токенизировался с ограничением max_seq_length=512 токенов, что соответствует типичной длине методического пояснения или параграфа учебника. График лучшей конфигурации (VER_1) представлен на рисунке 1, а в таблице 1 указаны

параметры конфигураций. Переобучение оправдано, поскольку покрывает полностью необходимую информацию и лишь улучшает несвязанные тематики [3].

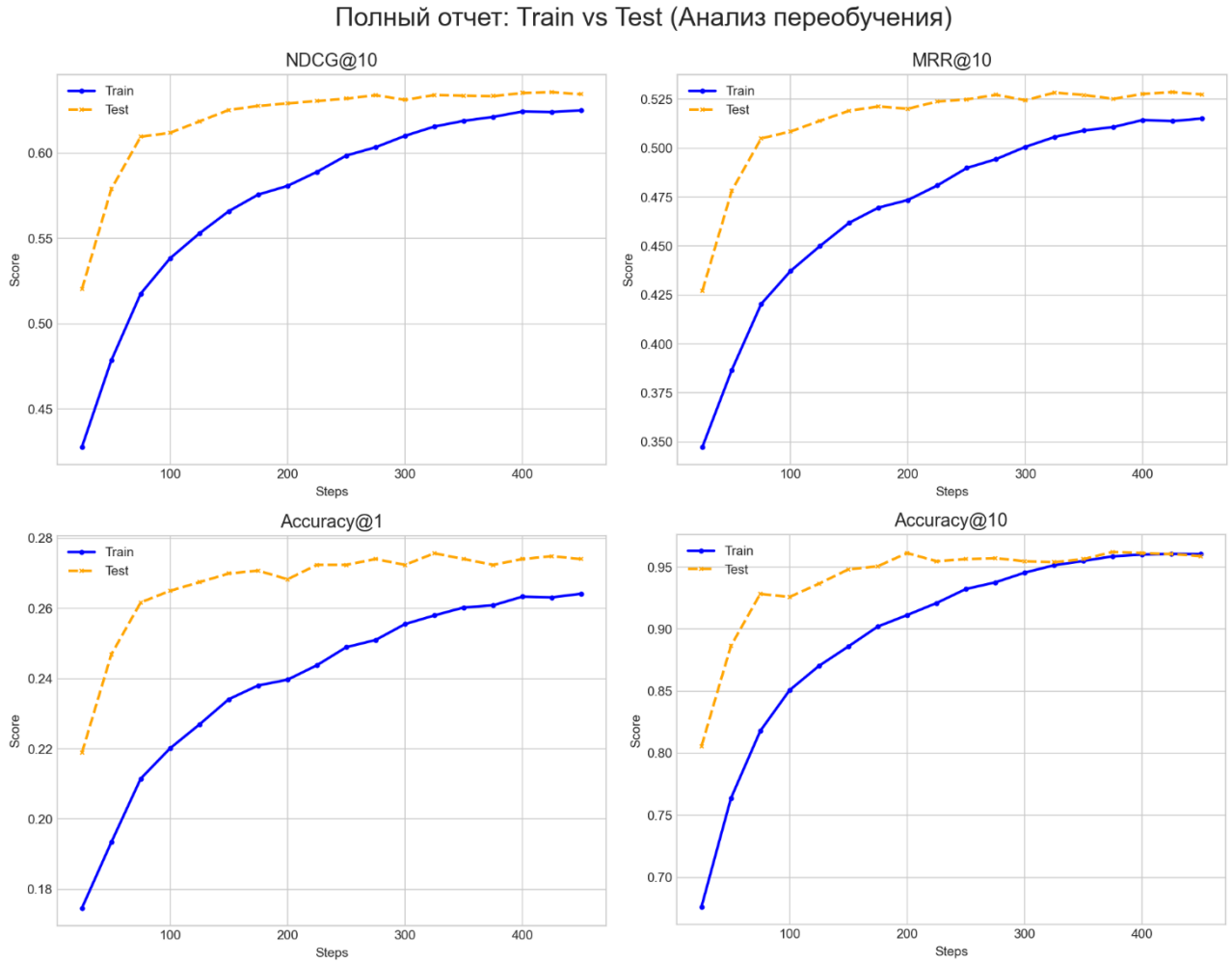

Рис.1. Графики обучения LoRA-адаптера по первой конфигурации (r=32, dropout=0.05)

Таблица 1

Результаты расчета энергетических уровней и сил осциллятора исследуемых конформаций паратерфенила методом TDDFT (Python)

Параметр	VER_1	VER_2	VER_4
Ранг (r)	32	16	64
Альфа (α)	64	32	128
Dropout	0.05	0.10	0.10
Доля обучаемых параметров	2.6%	0.2%	5.2%
Целевые модули	q, v, k, o, wi_0, wi_1, wo	q, v	q, v, k, o, wi_0, wi_1, wo

Заключение

Проведённое исследование наглядно демонстрирует, что использование LoRA-дообучения эмбеддинговых моделей на сравнительно небольших доменных датасетах (6000 триплетов) способно радикально повысить качество семантического поиска в условиях узкой предметной области. Достигнутый прирост метрик подтверждает реальную возможность создания специализированных поисковых ИИ-инструментов без привлечения огромных вычислительных ресурсов. Для образовательных организаций данный опыт открывает следующие перспективы: создание «умных» учебных пособий, виртуальных тьюторов и систем самопроверки знаний. Благодаря малому размеру возможно создание библиотек адаптеров для разных курсов.

Список литературы

1. LoRA: Low-Rank Adaptation of Large Language Models / E.J. Hu, Y. Shen, P. Wallis [и др.] // International Conference on Learning Representations. – 2022.

2. Lost in the Middle: How Language Models Use Long Contexts / N.F. Liu, K. Lin, J. Hewitt [и др.] // Transactions of the Association for Computational Linguistics. – 2024. – Vol. 12. – С. 157–173. DOI 10.1162/tacl_a_00638. EDN KCFYVI

3. Лакшманан В. Машинное обучение. Паттерны проектирования / В. Лакшманан, С. Робинсон, М. Мунн ; пер. с англ. – СПб. : БХВ-Петербург, 2025. – 175 с. – ISBN 978–5-4461–1629–4.