Денисова Полина Андреевна
Email: polina.denisova.2013@gmail.com
Муромский институт (филиал) ФГБОУ ВПО «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых»

Методы использования больших языковых моделей с собственной базой знаний для обработки запросов пользователей

Большинство промышленных предприятий долгое время пребывали в застое, опираясь на устаревшие технологии и методы работы. Политические и экономические изменения стали катализатором их трансформации. В качестве иллюстрации проблемы можно вспомнить холдинг РЖД с коллективом в 1 млн человек. В компании так много бумажной документации, что еще в советский период там создали Центр научно-технической информации и библиотек, где хранятся тонны макулатуры, анализ которых крайне затруднен. По оценкам зарубежных аналитиков, в организациях, где отсутствует база знаний, работники тратят на написание уже существующих заметок и статей больше времени чем на создание новых ресурсов**.** В связи с этим продвижение и реализация технологических инноваций, а также внедрение средств и инструментов цифровизации становятся главным фактором конкурентоспособности промышленных предприятий. Основным элементом такой трансформации должно стать создание собственной базы знаний, а для эффективной работы с ней – это использование ИИ-ассистента.

Создание ИИ-ассистента на собственной базе знаний включает в себя два направления: методы обработки базы знаний и интерпретация запросов. В работе рассматриваются методы использования больших языковых моделей с собственной базой знаний для обработки запросов пользователей.

Существует два главных метода работы больших языковых моделей с собственной базой знаний для обработки запросов:

1. Retrieval-Augmented Generation (RAG) – метод, где подключаются внешние данные во время запроса, то есть модель извлекает актуальные сведения из базы знаний, не меняя своих параметров.
2. Fine-tuning LLM (дообучение) – метод, где встраиваются знания в сами параметры модели, то есть модель обновляет веса на новом датасете.

Для промышленного предприятия оптимальным решением является RAG. Данный подход проще и дешевле реализовать, также модель всегда использует актуальные данные из базы знаний. Принцип работы ИИ-ассистента на основе RAG представлен на рисунке 1.

Рисунок 1. - Принцип работы ИИ-ассистента на основе RAG

Эффективность работы RAG-системы напрямую зависит от выбора большой языковой модели, именно от неё зависит качество, точность итогового ответа для пользователя. Далее представлено сравнение пяти наиболее распространенных LLM в таблице 1.

Таблица 1. – Сравнение больших языковых моделей

Критерий	DeepSeek	Llama	Mistal	GigaChat	GPT-4
Разработчик	DeepSeek AI (Китай)	Meta (США)	Mistral AI (Франция)	Сбер (Россия)	OpenAI (США)
Модель развертывания	Локально, облако	Локально	Локально, облако	Облако	Облако
Поддержка языков	Более 20 языков	8 языков	Более 12 языков	Русский язык, частично поддержка английского	Более 50 языков
Преимущества	- Контекстное окно до 128к токенов - Открытые веса - Низкая цена АР	- Полностью открытый код, - Большое сообщество и разнообразные инструменты для обучения	- Открытые веса - Хорошее соотношение скорость/ качество	- Отличная поддержка русского - Интеграция с экосистемой Сбера	-Мультимодальность - Лучшая точность и стабильность - Контекстное окно до 128к
Ограничения	- Требует мощных GPU - Не все версии стабильны	- Нет официальной поддержки русского, требуется доп. обучение - Требует много ресурсов	- Для больших версий нужны мощный GPU - Качество на русском ниже чем на английском без доп. обучения	- Закрытый исходный код - Контекстное окно до 32к токенов -Зависимость от внутренних API Сбера	- Высокая стоимость - Ограничения по конфиденциальным данным
Стоимость	Веб и мобильные приложения — бесплатные. API — $0.07–1.68 за 1 млн токенов	Модель бесплатная, затраты на инфраструктуру (железо, GPU)	Бесплатный тариф с лимитами. Платные Тарифы: Pro $14.99/мес, и Team - $24.99/мес	1 млн токенов бесплатно ежемесячно. После платные пакеты (от 200 ₽ до 1500 ₽ за 1млн токенов). Требуется авторизация через «Сбер ID»	Бесплатная версия с лимитами, Pro-план — около $200/мес.

Выбор модели для промышленных предприятий определяется балансом между требованиями к приватности, точностью обработки данных и стоимостью внедрения. Локальные модели (DeepSeek, Llama, Mistral) подходят для предприятий с высокими требованиями к информационной безопасности. Облачные решения (GigaChat, GPT-4) целесообразны при необходимости быстрого внедрения и масштабирования.

На основе проведённого исследования был разработан прототип ИИ-ассистента на основе базы знаний Linux Wiki. Прототип реализован на основе архитектуры RAG с использованием

языковой модели GigaChat, что позволяет ему использовать актуальную базу знаний для формирования точных и релевантных ответов. Предварительная оценка качества работы прототипа показала хорошие результаты: точность и релевантность ответов составляет более 85%.

Таким образом, для промышленных предприятий наиболее эффективным и экономически целесообразным подходом создания ИИ-ассистента на основе собственной базы знаний является использование архитектуры RAG. Среди рассмотренных больших языковых моделей оптимальный выбор зависит от требований к языковой поддержке, стоимости, доступности и уровню конфиденциальности данных.

Выражение благодарности:

Соавторы работы: к.т.н.,доцент Астафьев Александр Владимирович, МИВлГУ