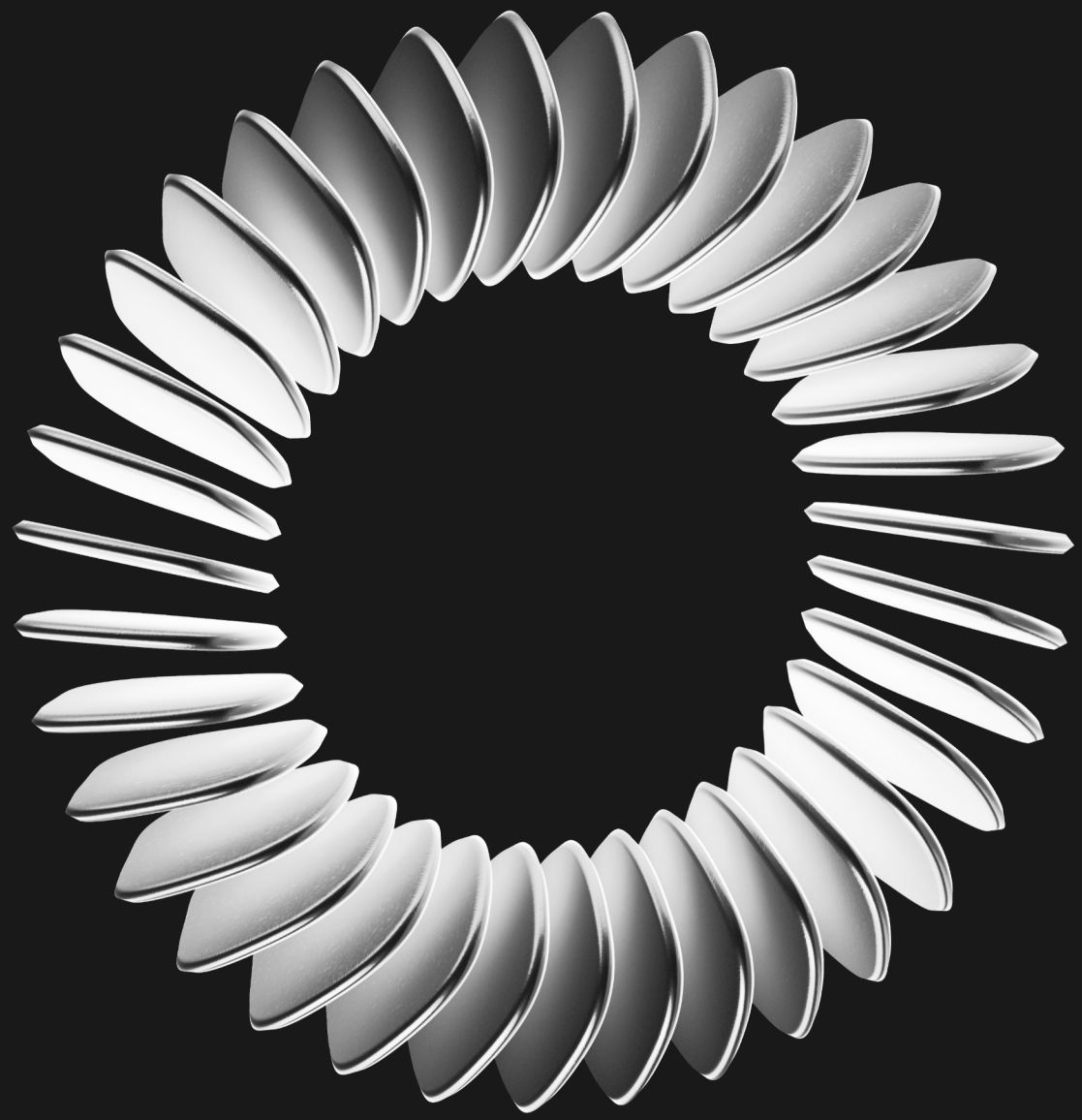
ПРИМЕНЕНИЕ LLM НА ПРИМЕРЕ БАНКОВСКОГО КЦ

Что такое LLM?

Большая языковая модель Large Language Model

— это нейронная сеть, созданная для обработки естественного языка.

Модель обучается на обширных текстовых данных и может выполнять разнообразные языковые задачи, такие как: перевод, суммаризация, ответ на вопросы и создание текста на основе заданного контекста.



На какие условные категории можно разделить LLM модели

Наличие исходного кода
- Открытый код
- Закрытый код

Языковая поддержка
- Модели, ориентированные на один язык
- Модели, поддерживающие несколько языков

Способ обучения
- Обученные с нуля
- Использующие дообучение или fine-tuning на специальных задачах

Объем обучающего корпуса
- Модели, обученные на большом количестве данных
- Модели, обученные на специализированных или ограниченных данных

Наиболее известные модели в РФ и за рубежом

CHATGPT | ЯНДЕКС GPT | MISTRAL | GIGACHAT | GEMMA | FALCON | GEMINI | LLAMA | CLAUDE | COMMAND | GROK | ORCA | PHI-3

Область применения
- Чат-боты
- Генерация текста
- Анализ текста и NLP-задачи

Географическая принадлежность
- Российские
- Зарубежные

Целевая аудитория
- Общее назначение
- Специализированные (например, модели для медицины, права и т. д.)

Тип лицензии
- Свободные
- Коммерческие

Доступность для пользователей
- Облачные решения
- Локальные решения (инструменты, которые можно установить и использовать на своих серверах)

Ожидания от LLM (в банковском КЦ)

Оптимизация

- Правильно определит потребность клиента
- Уточнит недостающие данные (в банковских системах или у клиента)
- Быстро подготовит правильный и понятный ответ на языке клиента
- Поймет любую речь клиента на любом языке

Улучшение качества

- Поймет что нужно для решения клиентского запроса
- Синтезирует голосовой ответ с заботой и эмпатией
- Удовлетворит все запросы быстро и безопасно, в соответствии с политикой Банка и законами РФ.
- Поймет все ли проблемы клиентов мы решили

Сделает работу за нас?

- Построит оптимальный клиентский путь
- В моменте сформирует оптимальный дизайн для клиента и оператора, с учетом устройства, с которого обращается человек
- Напишет код для отображения этого дизайна на устройстве

Какими пилотами мы проверяли возможности LLM?

Создание сервиса консультации клиентов по тарифам продуктов и услуг Банка

(QA сервис по сборникам тарифов на основе RAG)

Источником данных являлись сборники тарифов с сайта ВТБ:

- «По пакетам банковских услуг»
- «По депозитам и накопительным счетам»
- «По банковским картам»
- «По расчетно-кассовому обслуживанию»
- «По ипотеке»
- «По автокредитам»
- «По сейфовым ячейкам»
- «По ОМС и слиткам драгоценных металлов»
- «По аккредитивам»

Интеллектуальный анализ звонков клиентов

Суммаризация расшифровок звонков и генерация резюме звонка

Источником данных стали:

>35 тыс.

обезличенных диалогов

С ручной разметкой:

>2 тыс.

диалогов

Разметка содержала:

- Резюме звонка - решен или не решен вопрос клиента
- Краткое саммари диалога клиента и оператора

По каким критериям выбирали LLM для пилотов?

ruLLM_1

- Работа на русском языке
- Возможность дообучения модели на данных банка

ruLLM_2

- Потенциальная возможность установки решения в контур банка
- Наличие команды поддержки пилота

*В итоге остановились на 2х известных Российских моделях и 2х не менее известных фреймворках для создания приложений с помощью LLM

Какие дополнительные работы пришлось провести для старта пилотов?

Для сервиса консультации клиентов по тарифам продуктов и услуг Банка

- Конвертация данных из таблиц в текстовый формат
- Разделение данных на чанки (текстовые блоки до 2000 символов)
- Разработка RAG сервиса, для поиска информации и генерации ответа клиенту
- Подготовка тестовых данных для проверки точности поиска RAG (разметка)
- Выбор методики проверки (ручная/автоматическая) и подготовка данных для проверки полученных от LLM ответов (разметка/промпт)

Для интеллектуального анализа звонков клиентов

- Обезличивание клиентских диалогов, очистка текстовых данных от персональной информации
- Ручная разметка части диалогов (формирование краткого саммари и резюме звонка)
- Подготовка промпта для суммаризации
- Выбор методики проверки полученных результатов для суммаризации и генерации резюме. Подготовка данных для проверки

Какие исследования и работы появились уже во время пилота?

Для сервиса консультации клиентов по тарифам продуктов и услуг Банка

- Очистка текстовых данных от лишней информации (перевод json в псевдо-json)
- Формирование с помощью LLM парафраз и синтетических вопросов к каждому чанку для улучшения результатов ранжирования.
- Доработка RAG сервиса для генерации читабельного ответа клиенту
- Ручная оценка сгенерированного ответа в зависимости от контекста, подаваемого на вход LLM. Определение оптимального размера контекста для генерации.

Для интеллектуального анализа звонков клиентов

- Применение Map-reduce суммаризации. (Деление длинных диалогов на части, генерация саммари по каждой части, генерация финального саммари из промежуточных)
- Дополнительная разметка для оценки окончания диалога с помощью LLM (завершен/перевод/обрыв/благодарность)
- Сравнение Few-shot классификаций резюме звонка (проставление резюме решен/не решен вопрос клиента на основе разметки, с map-reduce и без)
- Сравнение Few-shot и Zero-shot классификации резюме звонка (проставление резюме с размеченными примерами и без них)

Итоги пилотов

Для сервиса консультации клиентов по тарифам продуктов и услуг Банка

Модель применима и имеет промышленный потенциал. При этом предъявляются повышенные требования к RAG сервису.

Необходимо дальнейшее пилотирование на внутренних базах знаний Банка:

- Проверка на естественных вопросах клиента
- Проработка стратегии ответов на общие вопросы
- Дополнительный контроль для оценки работ (ручной или автоматический), исключающий генерацию неверных/фантазийных данных.
- Изменение структуры хранения данных для поиска.

Для интеллектуального анализа звонков клиентов

В части суммаризации диалогов модели могут использоваться в промышленных задачах.

В части генерации резюме звонка необходимы дополнительные работы и тестирования:

- Дообучение моделей стандартам качества обслуживания
- Проработка стандартов классификации резюме в неоднозначных диалогах (например тех, где проблема решена частично, или разное видение решения проблемы со стороны клиента и оператора)
- Определение методик контроля за работой LLM

Дальнейшие шаги, чтобы приблизить

(01) Внесение LLM в контур банка

(02) Разработка и поддержка дополнительных сервисов для поиска, генерации, очистки данных и т.д.

(03) Изменение структуры хранения данных: текстовый формат, возможность дробления, минимизация дублей и т.д.

(04) Проработка стратегии ответов на общие и/или неоднозначные вопросы

(05) Дообучение моделей стандартам качества обслуживания банка

(06) Разработка стандартов классификации резюме

(07) Формирование системы контроля за работой LLM, для исключения неверных ответов

СПАСИБО ЗА ВНИМАНИЕ

Контакты

@NIKITAZHV

Жадаев Никита
Руководитель группы разработки
КЦ ВТБ
nzhadaev@vtb.ru

@LARSKIKH

Ларских Алексей
Замиститель лидера стрима
КЦ ВТБ
larskih@vtb.ru