

Часть 1

Применение LLM в банковском КЦ: Киберсуфлер и AI-контролер

В каких кейсах используем LLM?

Киберсуфлер

Отвечает на вопросы клиентов по тарифам и услугам банка

AI-контролер

Оценивает качество консультации и формирует резюме диалога

Почему LLM?

Киберсуфлер

- (01) Большое количество специфической информации (сборники тарифов, инструкции и тд.)
- (02) Частые обновления
- (03) Информация сложно структурирована
- (04) Ответ, найденный оператором, односложный или написан бюрократическим языком

Решение задачи методами классической разработки трудоемко в части ресурсов тестирования и поддержки. Требуется постоянных обновлений, чтобы не проседало качество консультаций.

AI-контролер

- (01) Сотни тысяч диалогов в день:
 - ~ 200 000 звонков и чатов с операторами
 - ~ 150 000 диалогов с голосовыми помощниками
- (02) Ручная выборочная проверка требует колоссальных ресурсов или покрывает ~ 1-2% коммуникаций

При таких объемах, анализ отклонений и формирование рекомендаций носят в основном реактивный характер, в ответ на резонансные кейсы и негативные оценки клиентов

Цели проекта

Этап 1:
январь 24 – май 24

- (01) Проведение исследований и пилотов «полного цикла» (консультация клиентов + оценка качества обслуживания) по выбранным тематикам
- (02) Подтверждение технологической возможности интеграции LLM в ИТ ландшафт банковского КЦ
- (03) Определение возможностей LLM и объемов/направлений дообучения моделей
- (04) Выбор оптимального промышленного решения по результатам проведенных пилотов
- (05) Доработка бизнес-процессов на стороне банка под выбранное решение
- (06) Доработка RAG под текущие возможности LLM и требования Банка

Этап 2:
август 24 – сентябрь 25

- (01) Интеграция выбранного решения в контур банка
- (02) Дообучение моделей внутренним регламентам и процедурам банка.
- (03) Проработка серых зон с точки зрения правовой базы и информационной безопасности
- (04) Внедрение промышленного решения на ограниченном круге пользователей
- (05) Тираж решения на операторов (суфлер)
- (06) Тираж решения на клиентов банка

Границы пилотов на первом этапе

Киберсуфлер

Отвечает на вопросы клиентов по тарифам и услугам банка

- «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»

AI-контролер

Оценивает качество консультации и формирует резюме диалога

- По диалогам клиент-оператор в голосовом канале и чате
- По завершенным и незавершенным диалогам
- В кейсах, когда вопрос клиента решен полностью или частично
- В кейсах, где у клиента один или несколько вопросов к оператору

Что умеют технологии?

Киберсуфлер

Отвечает на вопросы клиентов по тарифам и услугам банка

- Перерабатывает инструкции и тарифы в понятный для LLM формат
- Получает вопрос клиента на естественном языке
- Ищет компоненты для ответа на вопрос клиента
- Генерирует ответ

При оценке диалогов Киберсуфлера AI-контролером получаем полный цикл AI обработки

AI-контролер

Оценивает качество консультации и формирует резюме диалога

- Анализирует диалог клиента и оператора
- Фиксирует резюме решен или не решен вопрос клиента
- Классифицирует решенные и не решенные диалоги по группам

Объемы исследований

15

LLM (Российские и русскоязычные open-source решения)

4

месяца исследований, доработок и дообучений моделей

>140

тарифов и справочников из открытого доступа (с сайта банка)

>100 000

тестовых генераций

>5 000

данных с ручной разметкой для проверки работ LLM

2

фреймворка для создания приложений с LLM

97

гипотез и их комбинаций

>35 000

обезличенных диалогов

>250 000

резюме и суммаризаций



Ручная и автоматизированная оценка работ по поиску, генерации, суммаризации

Доведение моделей до точности

>85%

Сложности реализации

Отсутствие (на момент пилотов 1го этапа) необходимой инфраструктуры для размещения LLM в контуре Банка

Отсутствие четкого правового регулирования применения LLM в финансовой сфере (принадлежность прав на запросы и сгенерированные ответы, ответственность сторон в случае угроз информационной безопасности и тд.)

Отсутствие знаний на стороне LLM о внутренних процедурах КЦ конкретного банка, и необходимость подготовки данных для дообучения моделей

Дополнительные работы по обезличиванию клиентских данных и существенное ограничение в выборе пилотных тематик

Дробление проекта на этапы. Контроль генераций 1 этапа сотрудниками

Ручная разметка и подготовка дата-сетов для обучения моделей по выбранным направлениям

Какие работы провели на этапе пилотов и проверки гипотез?

Киберсуфлер

- (01) Перерабатывает инструкции и тарифы в понятный для LLM формат
 - Перевод табличных документов текстовый формат Json
 - Нарезка на чанки «с сохранением вложенности» контента
 - Формирование «псевдо-Json», без синтаксиса
 - Генерация парафраз с помощью LLM
 - Генерация синтетических вопросов
- (02) Ищет компоненты для ответа на вопрос клиента
 - Поиск чанков, содержащих ответ
 - При отсутствии чанка с ответом или нахождении > 10 чанков на общие вопросы - генерация дополнительных вопросов клиенту
 - Суммаризация вопроса и дополнительных ответов клиента для генерации клиентского запроса
- (03) Генерирует ответ
 - Генерация ответа клиенту на основании запроса и найденных чанков с информацией
 - При односложном «сухом» ответе – повторный запрос LLM и расширение фразы с учетом контекста и Tone of Voice
 - Анализ оптимального размера контекста для генерации ответа клиенту

AI-контролер

- (01) Анализирует диалог клиента и оператора
 - Перевод диалогов в текстовый формат с указанием стороны говорящего
 - Очистка диалогов от клиентских данных
- (02) Фиксирует резюме решен или не решен вопрос клиента
 - Ручная разметка диалогов для дообучения модели
 - Ручная разметка диалогов для тестирования результатов работы LLM
 - Few shot классификация vs Zero shot классификация
 - Few shot map vs Few shot map reduce классификация
- (03) Классифицирует решенные и не решенные диалоги по группам
 - Ручная разметка диалогов для дообучения модели
 - Ручная разметка диалогов для тестирования результатов
 - Проверка корректности действий оператора
 - Настроение клиента (благодарность, нейтрально, недовольство)
 - Технические причины (обрыв звонка, перевод, проблемы со связью)

Уникальность проекта

Проект в тренде на исследования и практики применения LLM в банковской сфере

- (01) Проведено масштабное исследование рынка Российских и русскоязычных LLM
- (02) Подготовлены практические инструменты и сервисы для работы
- (03) Выбрана модель для дальнейшего использования
- (04) Протестировано обслуживание «полного цикла» (консультация клиентов + оценка качества обслуживания) с помощью AI инструментов, без участия оператора

Встройка LLM и RAG в текущую комбинацию AI инструментов в КЦ для получения дополнительных эффектов

- (01) Создаем сложные рабочие комбинации взаимодействия различных AI инструментов для комплексного решения вопроса клиента и автоматизации КЦ
- (02) Подготавливаем базу для реализации цифрового финансового советника для внутренних и внешних клиентов банка

Встройка LLM и RAG в текущую комбинацию AI инструментов в КЦ

Этап обслуживания	Действия системы	Набор AI инструментов
(01) Первичный запрос клиента	<ul style="list-style-type: none"> ▪ Перевод голоса в текст ▪ Анализ запроса ▪ Маршрутизация клиента на основании запроса ▪ Запуск нужного сценария 	<ul style="list-style-type: none"> ▪ ASR ▪ NN, ML, LLM, кластеризация ▪ NN, ML, классическая разработка ▪ NN, ML, классическая разработка
(02) Уточнение запроса	<ul style="list-style-type: none"> ▪ Определение дополнительных вопросов клиенту ▪ Суммаризация ответов и формирование финального запроса на поиск 	<ul style="list-style-type: none"> ▪ Классическая разработка, LLM, RAG ▪ Классическая разработка, LLM, RAG
(03) Поиск информации для ответа на вопрос	<ul style="list-style-type: none"> ▪ Данные в Автоматизированных банковских системах ▪ Данные в базе знаний 	<ul style="list-style-type: none"> ▪ Классическая разработка, ML ▪ RAG, LLM
(04) Генерация ответа	<ul style="list-style-type: none"> ▪ По активным действиям с продуктами клиента ▪ По консультационным вопросам 	<ul style="list-style-type: none"> ▪ Классическая разработка, ML, LLM ▪ RAG, LLM
(05) Предварительная проверка ответа оператором	<ul style="list-style-type: none"> ▪ Временный шаг для консультационных вопросов и контроля работы LLM 	
(06) Автоклассификация запросов клиента	<ul style="list-style-type: none"> ▪ Определение и фиксация тематик обращения клиента в рамках одного звонка (4 уровня) 	<ul style="list-style-type: none"> ▪ NN, ML
(07) Контроль качества	<ul style="list-style-type: none"> ▪ Определение решен или не решен вопрос клиента ▪ Классификация диалогов по корректности, тональности, удовлетворенности, следование скрипту 	<ul style="list-style-type: none"> ▪ LLM + классическая разработка + (временно) специалист ▪ LLM + классическая разработка + (временно) специалист

Результаты 1 этапа и перспективы тиража



* Экономия времени сотрудников контроля качества с учетом текущего количества сотрудников. При тираже методики можно будет оценивать 100% диалогов.

Польза для всех участников

Для клиента

- Эффективность и скорость ответов, актуальность информации
- Адаптация ответов под естественную речь клиента (без бюрократии и формализма)
- Точность маршрутизации при переключении из цифровых каналов на сотрудника
- Реальная омниканальность: подготовка ответов с учетом истории цифровых взаимодействий с Банком

Для сотрудника

- Передача рутины на системный уровень, акцент на сложные запросы и профессиональное развитие
- Уверенность и скорость адаптации для новичков с подсказками от LLM
- Обратная связь по качеству обслуживания и индивидуальные рекомендации по улучшению диалога в реальном времени
- Статус. Из простого оператора в инженера по качеству AI суфлера

Для банка

- Рост производительности и сокращение затрат на ФОТ
- Точность аналитики и обработка обратной связи от клиентов в режиме онлайн
- Отслеживание трендов в поведении клиентов и их ожиданий, своевременная реакция на изменения рынка
- Снижение рисков влияния человеческого фактора и фрода со стороны сотрудников за счет автоматизации процессов.

Киберсуфлер и AI- контролер. Часть 1. Спасибо за внимание