Применение LLM в банковском КЦ: Киберсуфлер и AI-контролер. Часть 1

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

Большое количество специфической информации
(сборники тарифов, инструкции и тд.)

Частые обновления

Информация сложно структурирована

Ответ, найденный оператором, односложный или написан бюрократическим языком

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

AI-контролер

Сотни тысяч диалогов в день:
- ~ 200 000 звонков и чатов с операторами
- ~ 150 000 диалогов с голосовыми помощниками

Ручная выборочная проверка требует колоссальных ресурсов или покрывает ~ 1-2% коммуникаций

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

Цели проекта

Этап 1 Январь – Май 2024

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

Этап 2 Август 2024 – Сентябрь 2025

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

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

Что умеют технологии?

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 дообучения 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 для генераций резюме

>250 000
резюме и суммаризаций

///
Ручная и автоматизированная оценка работ по поиску, генерации, суммаризации

Доведение моделей до точности **>85%**

Сложность реализации

01

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

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

02

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.)

Дробление проекта на этапы. Контроль генераций 1 этапа сотрудниками

03

Отсутствие знаний на стороне LLM о внутренних процедурах КЦ конкретного банка. Необходимость подготовки данных для дообучения моделей

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

Какие работы провели на этапе пилотов и проверки гипотез?

Киберсуфлер

1. 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
2. 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 для уточнения запроса
   - Суммаризация вопроса и дополнительных ответов клиента для генерации клиентского запроса
3. 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

1. Анализирует диалог клиента и оператора
   - Перевод диалогов в текстовый формат с указанием стороны говорящего
   - Очистка диалогов от клиентских данных
2. 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
3. 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

01

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

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

02

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

- Создаем сложные рабочие комбинации взаимодействия различных AI инструментов для комплексного решения вопроса клиента и автоматизации КЦ
- 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	▪ Перевод голоса в текст ▪ Анализ запроса ▪ Маршрутизация клиента на основании запроса ▪ Запуск нужного сценария	▪ ASR ▪ NN, ML, LLM, кластеризация ▪ NN, ML, классическая разработка ▪ NN, ML, классическая разработка
(02) Уточнение запроса	▪ Определение дополнительных вопросов клиенту ▪ Суммаризация ответов и формирование финального запроса на поиск	▪ Классическая разработка, LLM, RAG ▪ Классическая разработка, LLM, RAG
(03) Поиск информации для ответа на вопрос	▪ Данные в Автоматизированных банковских системах ▪ Данные в базе знаний	▪ Классическая разработка, ML ▪ RAG, LLM
(04) Генерация ответа	▪ По активным действиям с продуктами клиента ▪ По консультационным вопросам	▪ Классическая разработка, ML, LLM ▪ RAG, LLM
(05) Предварительная проверка ответа оператором	▪ Временный шаг для консультационных вопросов и контроля работы LLM	
(06) Автоклассификация запросов клиента	▪ Определение и фиксация тематик обращения клиента в рамках одного звонка (4 уровня)	▪ NN, ML
(07) Контроль качества	▪ Определение решен или не решен вопрос клиента ▪ Классификация диалогов по корректности, тональности, удовлетворенности, следование скрипту	▪ LLM + классическая разработка + (временно) специалист ▪ LLM + классическая разработка + (временно) специалист

Результаты 1 этапа и перспективы тиража

	Масштабы 1 этапа	Перспективы тиража
Найдено ответов на консультационные вопросы клиентов	>100 000	>1 500 000 в месяц
Сэкономлено времени сотрудников на поиск в инструкциях	850 человекочасов	12 500 человекочасов в месяц
Оценено качество диалогов	>35 000	>5 000 000 в месяц
Сэкономлено времени сотрудников на ручную проверку диалогов	90 человекочасов*	800 человекочасов в месяц *
Найдено клиентских благодарностей	750	>10 000 как минимум :)

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

Польза всем участникам

Клиенту

- 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

Сотруднику

- 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 облуживания и индивидуальные рекомендации по улучшению диалога в реальном времени
- Статус. Из простого оператора в инженера по качеству AI суфлера

Банку

- 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.

Применение LLM в банковском КЦ: Киберсуфлер и AI-контролер. Часть 1.

Спасибо за внимание