



Использование GenAI для аналитики в КЦ

Методика Call Drivers (CISCO)

Олег Зельдин

Алекс Берг

Типовые области аналитики КЦ

Авто-резюме
разговоров

Эмоциональный
анализ и поведение

Тематическая
классификация

Генерация FAQAuto

QA (качество
обслуживания)

Прогнозирование
намерений клиента

Структура презентации

Общее описание подхода

Генерация Call drivers

Автоматизация поиска инсайтов на базе Call drivers

- Кластеризация (поиск тематик)

- Классификация и самообучение модели

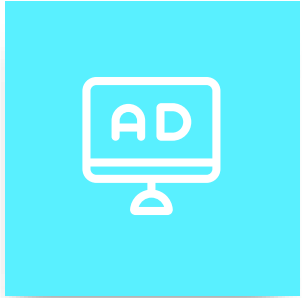
- Определение тенденций по частоте

- Формирование FAQ

Экономический эффект

Общее описание подхода

Обзор

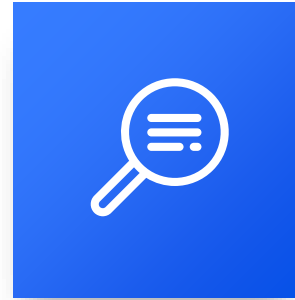


Аннотация

CISCO представили систему, которая автоматически определяет причины обращений (Call Drivers), используемые в дальнейшем для решения таких задач, как:

- тематическое моделирование
- классификация поступающих звонков
- выявление тенденций
- создание часто задаваемых вопросов (FAQ)

Показан экономически эффективный подход к реализации



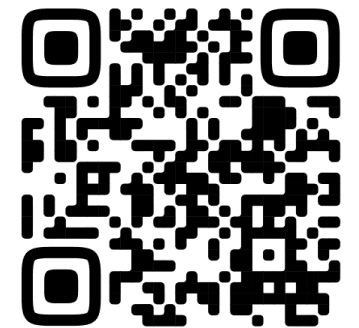
Источник

Название статьи и дата публикации

- LLM-Based Insight Extraction for Contact Center Analytics and Cost-Efficient Deployment. **24.03.2025**

Ссылка

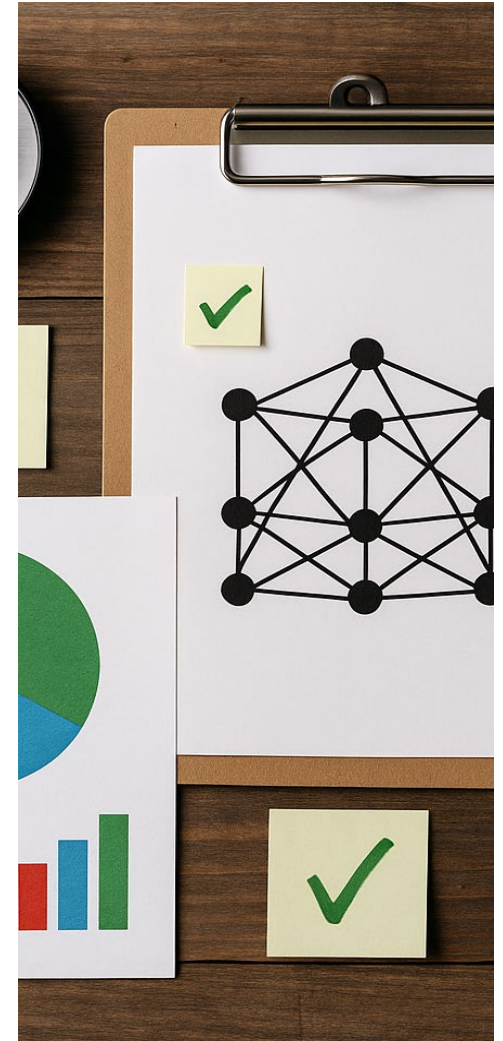
- <https://clck.ru/3Mkd7M>



Почему нужно понимать тематику обращений?

Чтобы анализ обращаемости был эффективным, система должна **распознавать, по каким темам чаще всего звонят клиенты.** Обычные системы тематического анализа (topic modeling):

- ✓ Дают полезные выводы
- ✓ Зависят от текста транскрипции и словаря,
- ✓ Часто **перекрываются по темам** (один звонок попадает в несколько категорий)
- ✓ Требуют **ручной доработки** от аналитиков КЦ.



✓ Какой подход предлагает Cisco?

Переход от тематик к ключевым причинам звонка (call drivers). Это **основная цель звонка**, выраженная кратко и понятно. Преимущества:

- ✓ меньше неоднозначности
- ✓ понятнее для аналитики
- ✓ можно использовать в автоматических системах

Для этого они используются **LLM, дообученные на инструкциях (instruction-tuned)** — это позволяет работать даже с небольшим объёмом данных и без необходимости вручную размечать миллионы звонков.



⚠ Проблемы сторонних моделей

Некоторые поставщики (например, OpenAI, Google и др.) предлагают готовые LLM, которые тоже умеют выделять call drivers. Но у такого подхода есть **минусы**:

- ✓ **Высокие затраты**, особенно при большом количестве звонков
- ✓ **Риски конфиденциальности** — чувствительные данные уходят за пределы компании
- ✓ **Ограниченная кастомизация** — если нужно дообучение под свою специфику, это стоит очень дорого





Что включает предложенная система?

Основные модули:

Модуль

Что делает

1. Генерация call drivers	Извлекает краткую суть звонка (главную причину обращения) из расшифровки. Это основа для решения всех задач в других модулях
2. Тематическое моделирование	На основе этих причин формирует группы тем с понятными метками.
3. Обнаружение трендов	Выявляет новые и растущие по частоте темы по подкатегориям
4. Генерация FAQ	Находит повторяющиеся темы и создаёт из них Часто Задаваемые Вопросы.

Предшествующие исследования

- Появление архитектур на базе трансформеров *"Attention is All You Need"* (2017) привело к существенному прогрессу в индустрии Контактных Центров
- Анализ упомянутых сущностей и настроения клиента. «*Entity-level sentiment analysis in contact center telephone conversations*», «*Improving named entity recognition in telephone conversations via effective active learning with human in the loop*» (2022)
- Генерация резюме после звонков. «*Topic-oriented spoken dialogue summarization for customer service with saliency-aware topic modeling*» (2021)
- Виртуальные агенты, которые не придумывают ответ, а ищут в документации. Методика RAG «*Retrieval-augmented transformer-xl for close-domain dialog generation*» (2021)
- Определение тренинговых ресурсов для операторов на базе причин обращений. «*Developing a production system for purpose of call detection in business phone conversations*» (2022)
- Автоклассификация звонков, чатов. «*Call-type classification and unsupervised training for the call center domain*» (2003). «*Topic modeling for automatic analysis of natural language: A case study in an Italian customer support center*» (2022), «*Topic modeling for customer service chats*» (2021)

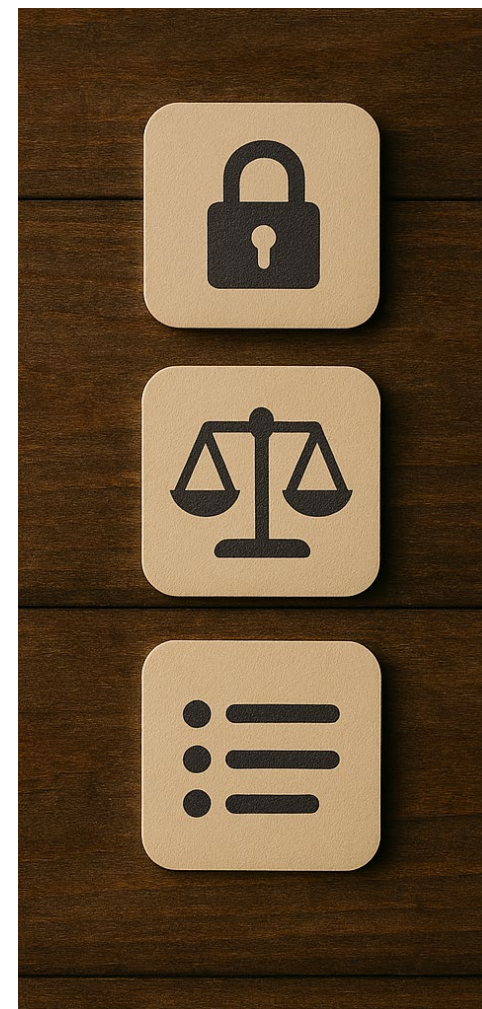
Генерация Call drivers

Call driver — это краткое описание (15–20 слов), которое подводит итог основной причины звонка клиента. Чем короче Call driver, тем лучше

Цель создания системы генерации Call drivers

Создать модель, готовую к использованию в реальной среде, которая СМОЖЕТ:

- автоматически **извлекать основную причину обращения** из расшифровки звонка;
- учитывать при этом:
 - ✓ **конфиденциальность данных,**
 - ✓ **соответствие юридическим требованиям,**
 - ✓ **оптимизацию затрат,**
 - ✓ **минимальные задержки обработки.**



Call driver. Пример

Расшифровка звонка (Transcript)

Клиент: Здравствуйте, не могли бы вы помочь мне получить доступ к моему аккаунту? Я не могу войти.

Оператор: Конечно. Вы знаете email или номер телефона, с которым вы регистрировались?

Клиент: Да, моя почта — <email>, а номер — <номер телефона>.

Оператор: Хм, похоже, ваш аккаунт неактивен. Когда вы в последний раз пытались войти?

Клиент: Я подписан на рассылку, но никогда не заходил в портал.

Оператор: Понятно. Похоже, из-за длительного бездействия ваш аккаунт был заблокирован. Я могу помочь вам его восстановить. Хотите, чтобы я это сделал?

Клиент: Да, конечно. Давайте сделаем это.

Call driver: *Получение помощи с доступом к заблокированному аккаунту*



Почему длинные Call drivers — это плохо?

- ✓ Длинные фразы мешают системам анализа понять суть обращения.
 - ✓ Они могут включать сразу несколько причин обращения, что затрудняет автоматическую классификацию и портит точность аналитики.
 - ✓ Модели начинают **ошибочно считать**, что такие фразы "нейтральны", т.е. не уверены, что они вообще правильно описывают звонок
- 💡 Поэтому команда специально ввела **штраф за длину** (length penalty), чтобы обученная модель генерировала **короткие и точные** описания обращения клиента.



Использование моделей для сравнения

ChatGPT-3.5 Turbo

Коммерческая модель от OpenAI (очень мощная, но платная и не открытая)

Mistral-7B-Instruct-v0.2

Открытая модель, которую можно запускать на своих серверах бесплатно

LoRA FT 4-bit Mistral

Модель Mistral дополнительно обучили на сгенерированных звонках через LoRA



Что такое LoRA?

LoRA (Low-Rank Adaptation) – метод, который **значительно снижает количество обучаемых параметров** при адаптации к новым задачам, но демонстрирует такую же или **лучшую точность**, чем полная настройка

Первоисточник

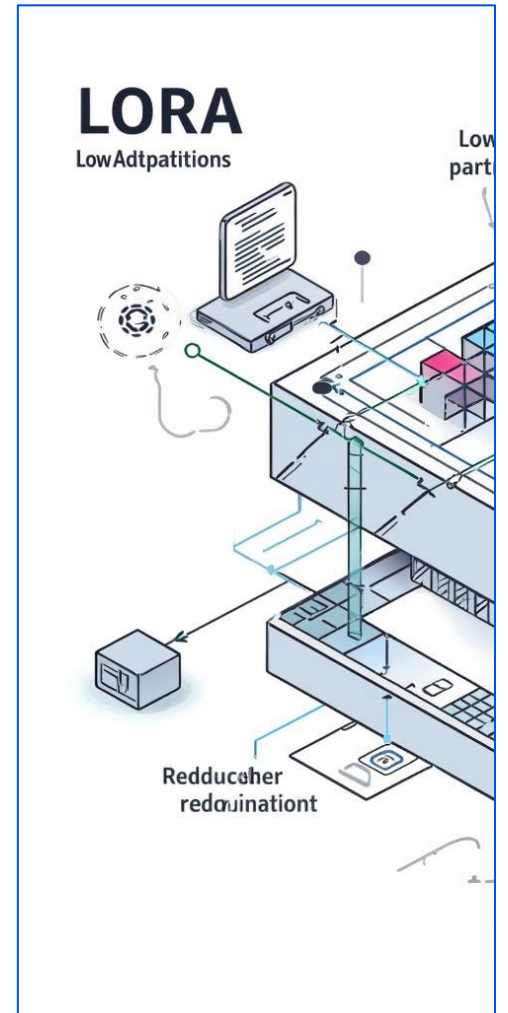
- LoRA: Low-Rank Adaptation of Large Language Models. **17.06.2021**

Ссылка

- <https://clck.ru/3Mkgdq>

Исходный код на github

- 👉 <https://github.com/microsoft/LoRA>



Как обучали и оценивали модель?

- Тестовый набор
- Обучающий набор
- Оценка результатов

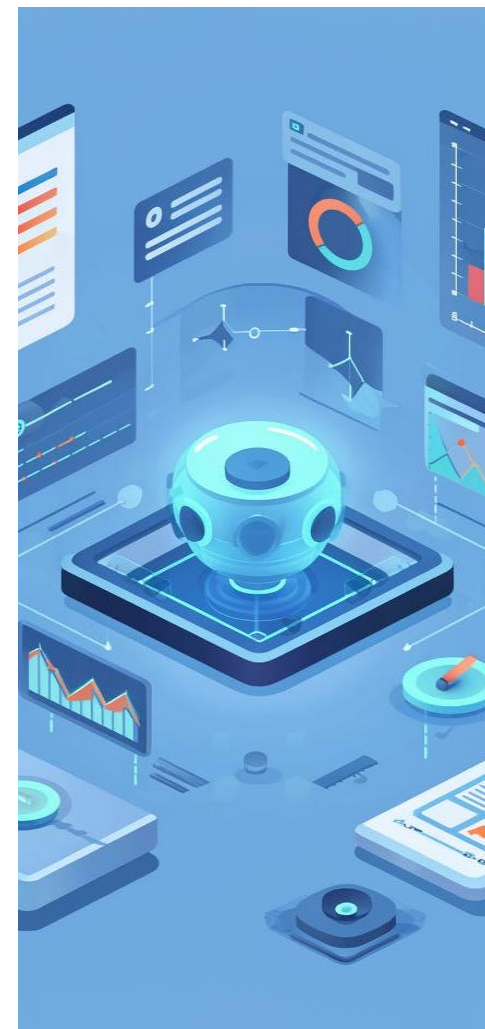
Тестовый набор (для проверки модели)

Чтобы проверить, насколько хорошо модель создаёт описания причин звонков (call drivers), авторы использовали **две настоящие базы звонков** из разных сфер:

- Служба доставки — 2000 звонков
- IT-служба поддержки — 3000 звонков

Звонки были **распознаны из аудио в текст** с помощью сервиса **Azure Speech-to-Text** (то есть, сначала перевели аудио в текст, где видно, кто говорит — клиент или оператор).

После этого **человек вручную описывал основную причину каждого звонка** — чтобы была "правильная" версия, с которой можно сравнивать работу модели.





Обучающий набор (для обучения модели)

Из-за ограничений по конфиденциальности данных клиентов, настоящие звонки нельзя было использовать для обучения. Поэтому они сами сгенерировали 750 "придуманных" звонков, где:

- два человека разыгрывали разговор — один был клиентом, другой оператором;
- в сценарии заранее задавались: тема, проблема и возможное решение;
- затем другие люди описывали основную причину обращения по этим "псевдо-звонкам".
- Таким образом, обучение шло на искусственно созданных, но реалистичных звонках.





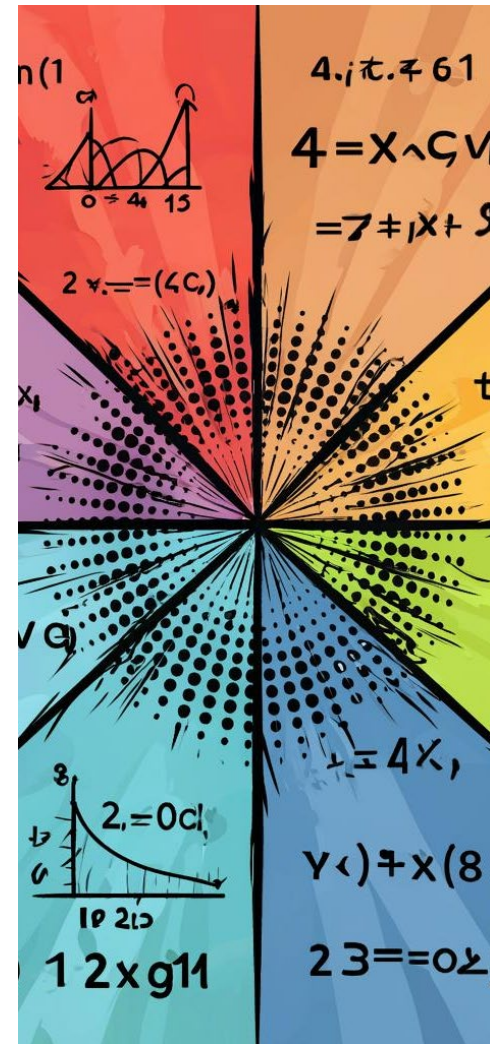
Как оценивали результат

Использовалась модель, которая умеет проверять, **имеет ли одна фраза тот же смысл, что и другая** — называется это **entailment** (включение смысла).

👉 Например: "Аккаунт заблокирован" и "Нужна помощь со входом в аккаунт" может означать одно и то же, и система должна это понять.

Ввели **штраф за длину** (так называемый *length penalty*), чтобы модель не «размазывала» смысл.

Использовались **специальные математические формулы оценки качества**



Результаты работы моделей по созданию Call drivers

ChatGPT-3.5 Turbo

Коммерческая модель от OpenAI
(очень мощная, но не открытая)

Mistral-7B-Instruct-v0.2

Открытая модель, которую можно
запускать на своих серверах

LoRA FT 4-bit Mistral

Модель Mistral дополнительно
обучили на сгенерированных
звонках через LoRA



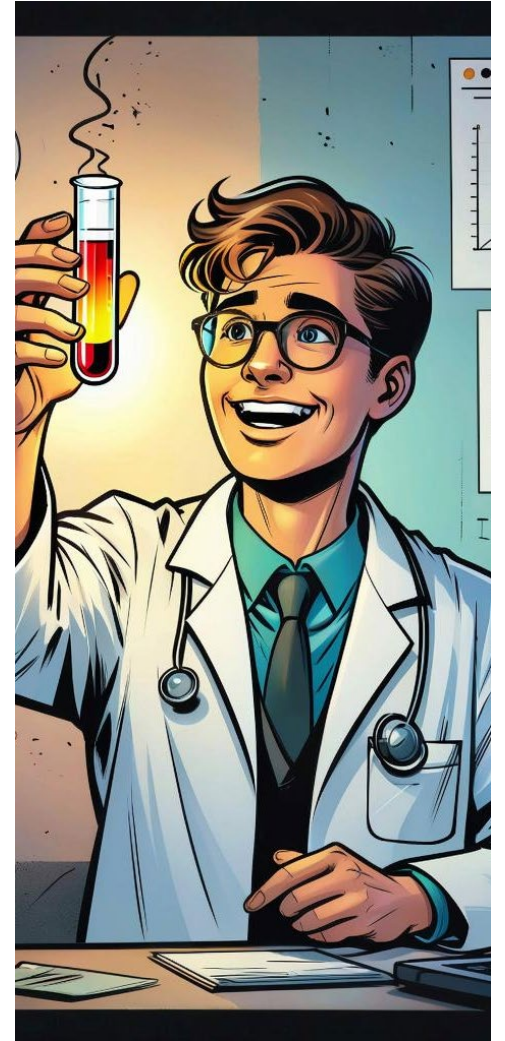
Что показали результаты?

Все модели справились по-разному. Ниже примеры, как одна и та же ситуация была описана тремя моделями:

👉 **Ситуация:** клиент просит ноутбук на замену.

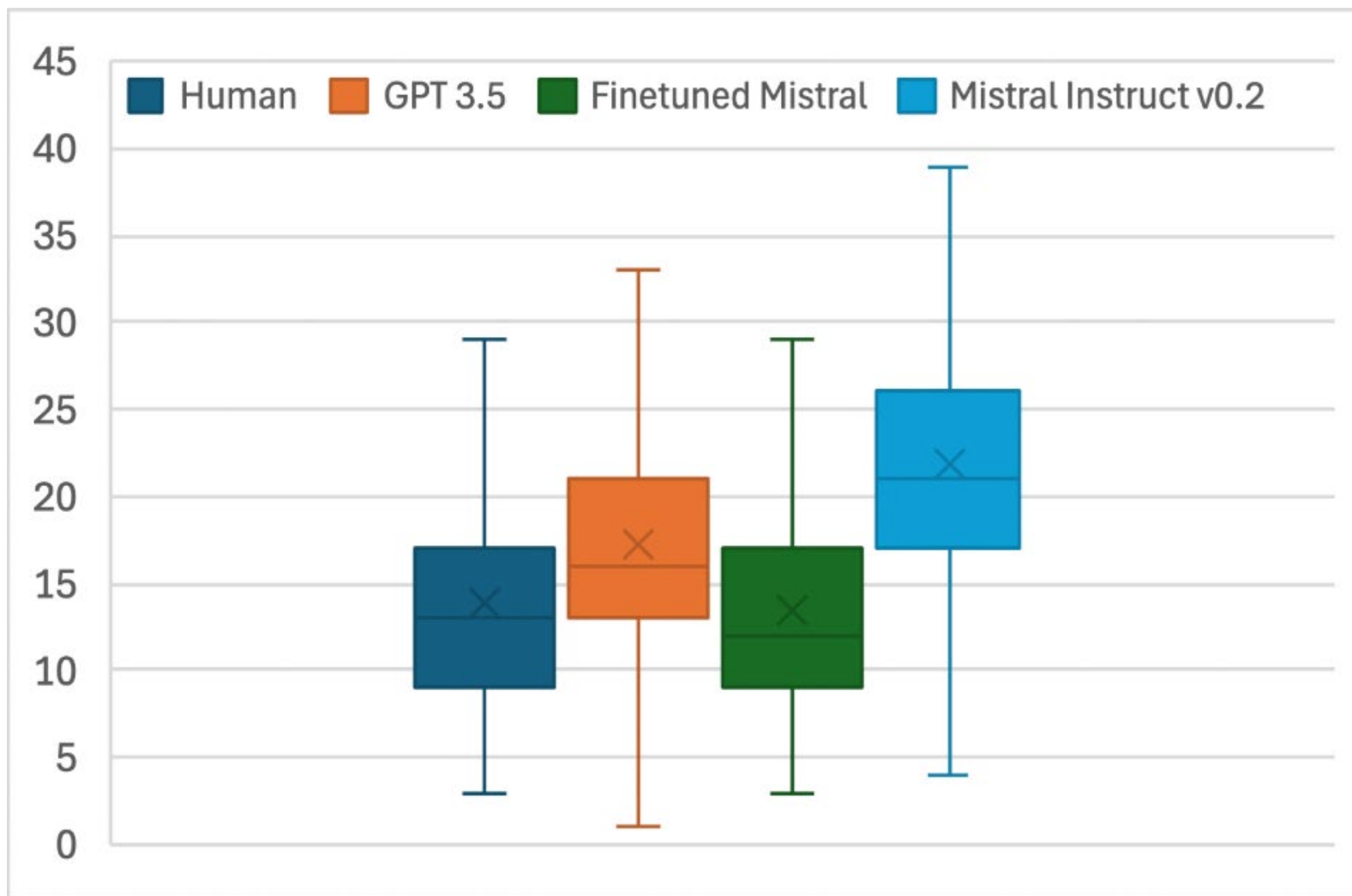
Ответы моделей:

- ♦ **LoRA FT 4-bit Mistral (дообученная модель):**
«Запрос на предоставление временного ноутбука»
(коротко, по сути, как человек-оператор написал бы в карточке обращения)
- ♦ **ChatGPT-3.5 Turbo (без обучения):**
«Клиент запросил дополнительный ноутбук и хотел узнать процедуру оформления»
(уже длиннее, два действия в одном предложении)
- ♦ **Mistral-7B-Instruct-v0.2 (без обучения):**
«Клиент интересовался, как запросить ноутбук с определёнными характеристиками, а также спрашивал об одобрении и доставке»
(ещё длиннее, слишком подробно)





Что показали результаты?



GPT-3.5 и обычная Mistral генерируют слишком длинные описания.

А дообученная Mistral создаёт короткие, точные описания, похожие на то, как это делает человек.

Автоматизация поиска инсайтов на базе Call drivers

- Кластеризация (группировка обращений по темам)
- Классификация (автоопределение темы новых обращений)
- Обнаружение новых тенденций по частотности
- Создание FAQ

Кластеризация (группировка)

Кластеризация. Как это работает — пошагово

Шаг 1. Используем
«call drivers»

Шаг 2. Превращаем
тексты в цифры

Шаг 3. Группируем
похожие тексты

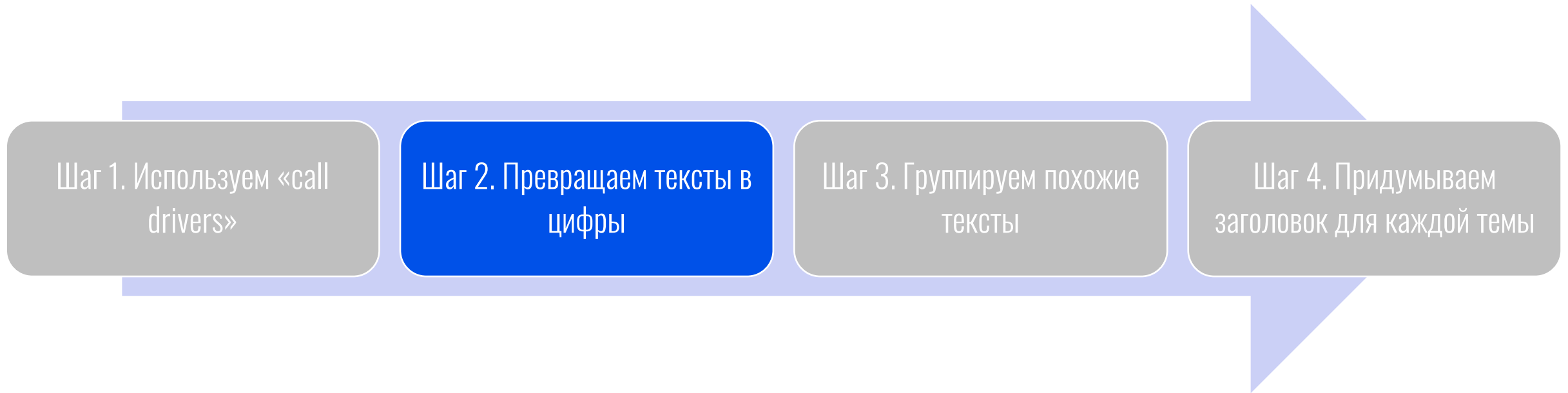
Шаг 4. Придумываем
заголовок для
каждой темы

Кластеризация. Как это работает — пошагово



Call drivers — это короткие описания, **почему клиент позвонил** (пример: "не удаётся войти в аккаунт"). Такие описания были сгенерированы моделью. Теперь они используются как «зерно» для поиска тем.

Кластеризация. Как это работает — пошагово



Все описания преобразуются в специальные числовые представления, с помощью модели **MiniLM**.

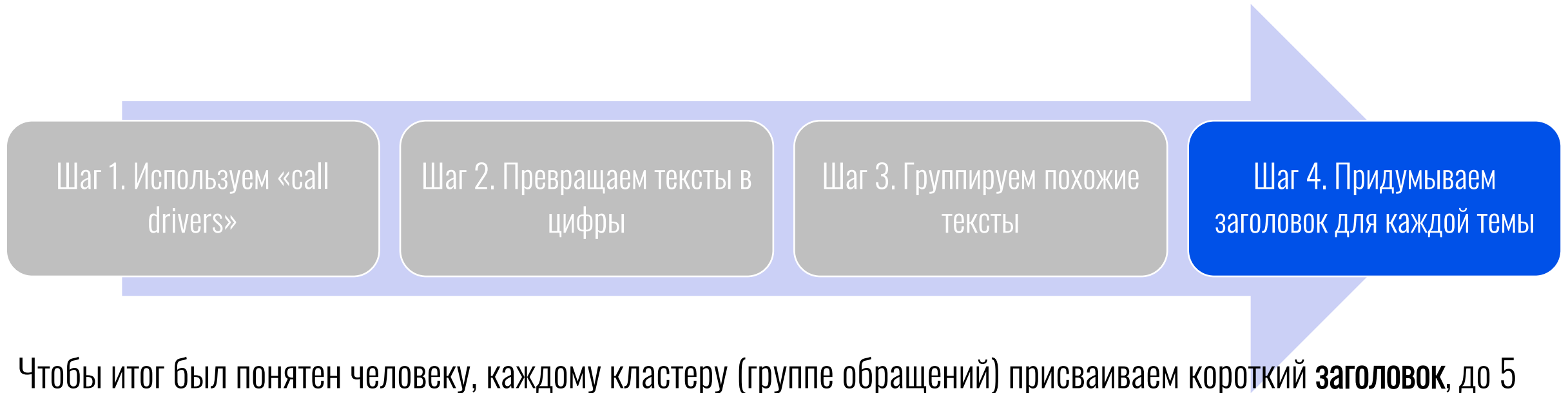
Это как если бы каждый текст стал точкой в многомерном пространстве.

Кластеризация. Как это работает — пошагово



- Используется алгоритм **HDBSCAN** (он умный и не требует заранее указывать, сколько будет тем). Он сам объединяет похожие точки в кластеры — то есть темы.
-  Почему HDBSCAN, а не привычный K-means?
Потому что он **сам определяет количество тем**, а не требует заранее вручную вводить их.

Кластеризация. Как это работает — пошагово



Чтобы итог был понятен человеку, каждому кластеру (группе обращений) присваиваем короткий **заголовок**, до 5 слов. Например: «Проблемы со входом в аккаунт»

Для этого используется та же языковая модель (Mistral), но в обычной версии, **без дообучения**. Модель получает:

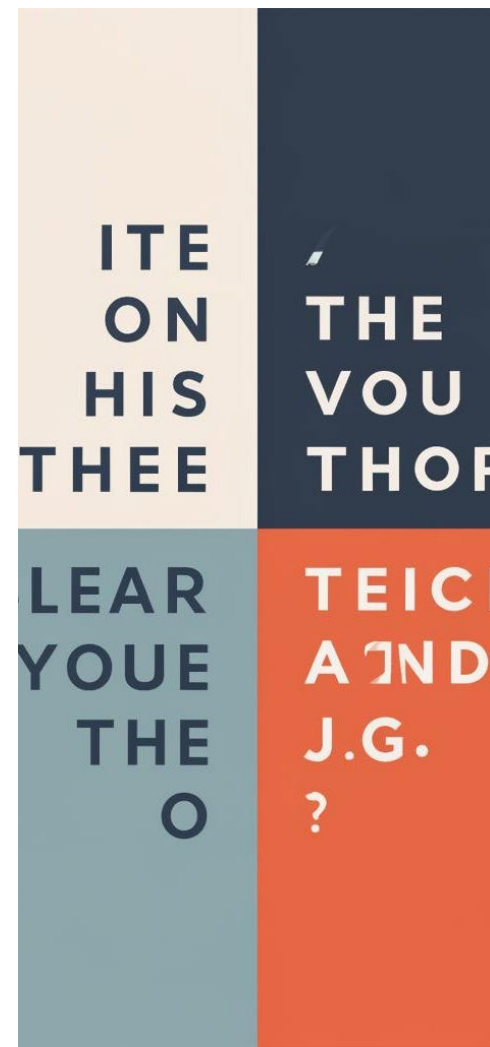
- Несколько фраз из группы
- Три самых популярных слова в этой группе

И на этой основе создает понятный заголовок

⚙️ Как выбираются фразы и слова для генерации темы?

Для каждой темы:

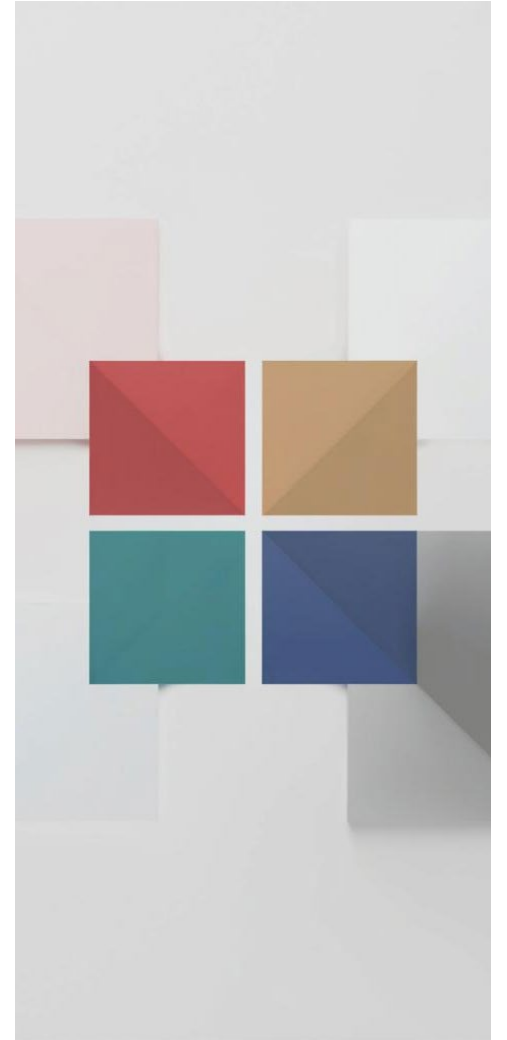
- удаляются **стоп-слова** (например, "и", "на", "с");
- слова приводятся к основе ("запрашивать", "запросил" → "запрос");
- выбираются **25 самых повторяющихся описаний** и **3 самых популярных слова**.





Как проверяется качество кластеров (тем)?

- **Коэффициент сплочённости тем (DBCV)**
→ показывает, насколько хорошо фразы внутри темы похожи друг на друга.
- **Косинусное сходство**
→ насколько тема и каждый call driver по смыслу совпадают.
- **Логическая связь (entailment)**
→ действительно ли тема **логически охватывает** каждую фразу в группе





Что это даёт менеджеру контакт-центра?

Благодаря **call drivers**, система может автоматически:

- понять, какие темы волнуют клиентов,
- оценить, какие темы требуют больше операторов,
- построить аналитику и прогнозы без ручной классификации.

Хорошие **короткие** описания звонков позволяют:

- точно группировать обращения,
- автоматически подписывать темы в отчётах,
- уменьшать нагрузку на аналитиков и супервизоров.



Классификация



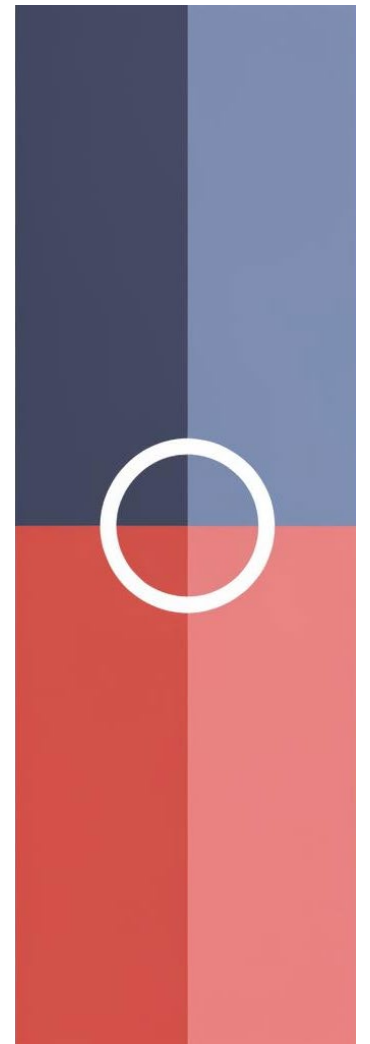
Как это делается без доп. обучения модели?

Раньше, чтобы классифицировать звонки, нужно было обучать модель на **новых данных**. Это долго и дорого.

Теперь — всё проще. Используется уже готовая **карта тем** (тематическая **модель**), созданная на предыдущем шаге.

 Что происходит, когда поступает **новый звонок**?

- Модель создаёт *call driver* (например: "Клиент хочет заказать оборудование").
- Этот call driver **сравнивается со всеми уже известными темами** (кластерами), чтобы найти **самую похожую**.
- Звонок автоматически помечается **меткой подходящей темы**, например: "Заказ оборудования".





Как система обучается на ходу?

Если call driver:

- немного отличается по словам от других в группе,
- но при этом **по смыслу подходит к теме** —
→ он просто **добавляется в эту тему** и помогает **обновить её** содержание.

В противном случае формируется новая тема и актуализируется в случае достаточной частоты

То есть система **постепенно становится умнее** — без ручной настройки.





Что это даёт менеджеру контакт-центра?

- 📌 Каждый звонок моментально получает тему, без участия человека.
- 📊 Можно видеть, какие темы растут по частоте
- 🔄 Темы автоматически обновляются, когда появляются новые формулировки.
- 📄 Можно в любой момент выгрузить отчёт: «Сколько было звонков по теме X за неделю?»



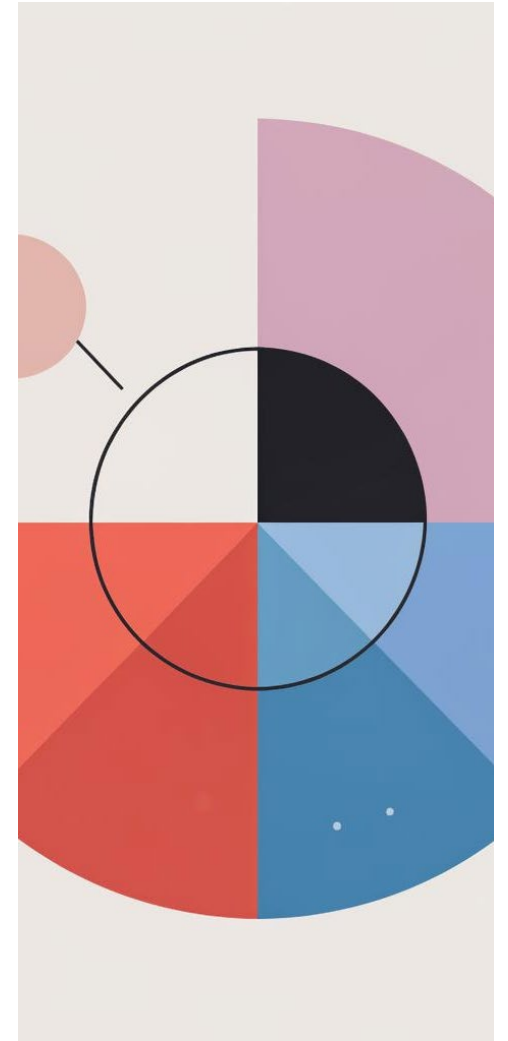
Обнаружение новых тенденций

В чём суть?

Контактные центры получают тысячи звонков. Система уже группирует звонки по темам. Но мир меняется — и клиенты начинают звонить по новым поводам, **которых раньше не было.**

 Задача — **заметить появление новой темы как можно раньше,** чтобы:

- подготовить агентов,
- обновить чат-ботов и скрипты,
- включить тему в отчёты и мониторинг.





Как система это делает?

Когда приходит новый "call driver", он может попасть:

- В **крупную тему**, которая уже давно есть (например, "Доставка").
- В **маленькую тему**, которая пока не очень частая (например, "Возврат через приложение").
- В **"мусорную группу"** — где собраны редкие и непонятные темы.

Если маленькая группа начинает быстро расти — это знак, что появляется новая волна обращений. 👉 Значит, мы выявили новый тренд!

Если растёт только "мусорная группа", без чёткой темы — это пока не считается новым трендом.

При накоплении похожих случаев в "мусорной группе" — создаётся **новая тема автоматически**





Что это даёт руководителю КЦ?



Вы можете видеть: “Начинает появляться новая тема, по которой растёт поток”.



Можно заранее:

- обновить FAQ,
- усилить чат-бота,
- сделать алерты для супервайзеров.



И всё это — без переобучения модели. Система адаптируется сама.



Часто задаваемые вопросы (FAQ)

Подход к формированию FAQ

- 1 Берём темы звонков и для каждой темы — находим фразы клиента
 - 📌 Например: Call driver: "Хочу отменить заказ". Фраза клиента в звонке: "Здравствуйте, мне нужно отменить заказ, который я сделал вчера"
- 2 Выбираем 5–20 таких фраз (случайный отбор)
- 3 Отправляем фразы в GPT-3.5

Модель GPT **сама определяет, какие из них звучат как часто задаваемые вопросы, и переформулирует** их в красивый формат FAQ:

📌 Пример: Вход: 10 фраз клиентов вроде "Где мой заказ?", "Когда доставят?", "Прошло уже 5 дней!"

Результат: **Вопрос:** «Когда будет доставлен мой заказ?»

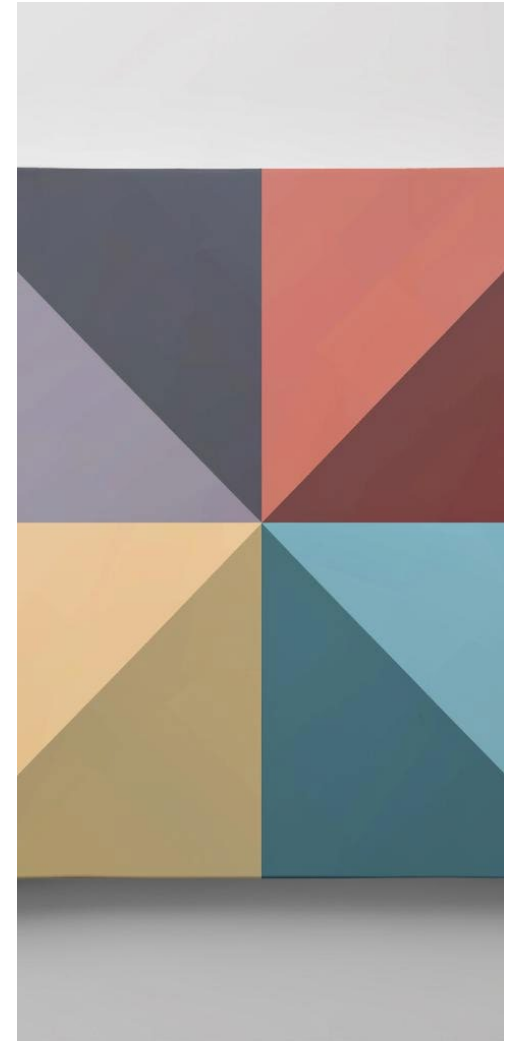
- 4 Группируем похожие вопросы





Как это применить в КЦ?

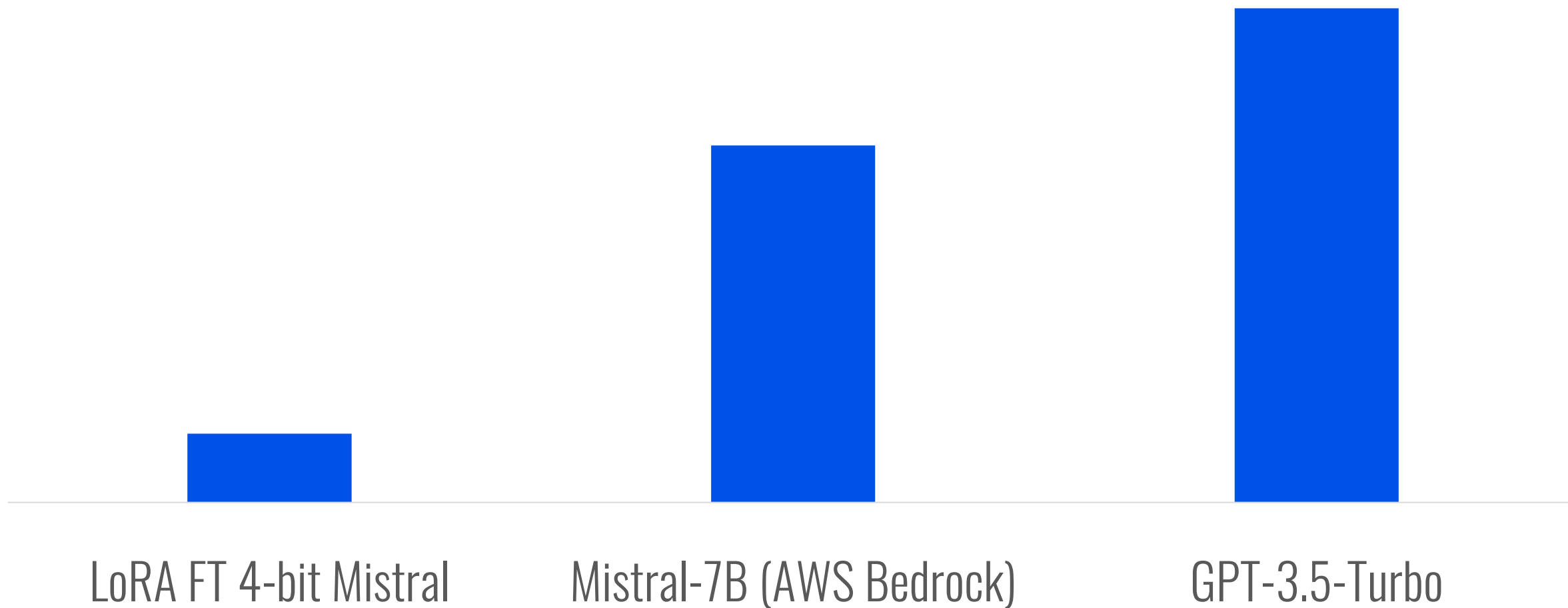
- ➡ Обновить голосовое меню (IVR): вставить туда самые частые вопросы.
- 💬 Подключить к чат-боту: он будет отвечать ещё до перехода к агенту.
- 📄 Сделать памятки агентам: «На этот вопрос отвечаем так...»
- 📌 **Итог:**
 - 📉 Меньше нагрузки на агентов.
 - 📈 Выше удовлетворённость клиентов.
 - 🧠 Возможность **видеть реальные болевые точки клиентов** и решать их проактивно.



финансовый
эффект



Сравнение затрат



СПАСИБО ЗА ВНИМАНИЕ!

apexberg.ru

ТГ-КАНАЛ:
Клиентский сервис –
искусство служить людям

