Дмитрий Пром

AI Engineer · production LLM-системы и аналитика на разговорных данных

ОТКРЫТ К SENIOR / LEAD APPLIED AI · 5 ЛЕТ В РАЗРАБОТКЕ, 2 В AI · УДАЛЁННО · UTC+3

xpromx94@yandex.ru
github.com/94spec
94spec.github.io
leadchat.online

КРАТКО

Строю production LLM-системы, которые закрывают ручную работу коммерческого блока: разговор с клиентом превращается в проверяемые данные. И сам собираю на этих данных аналитику, по которой в компании принимают решения — по продажам, маркетингу, продукту и тарифам.

Отвечаю за весь путь: бизнес-регламент → поведение модели → архитектура → реализация → оценка качества → внедрение в CRM и BI → стоимость и latency → отчёт, по которому действуют. Людьми не руководжу: работаю в связке с разработчиками и сам закрываю AI-контур.

КЛЮЧЕВЫЕ РЕЗУЛЬТАТЫ · КАЖДАЯ ЦИФРА СО ЗНАМЕНАТЕЛЕМ

100% / ≤20%
подходящих диалогов оценивается вместо ручной выборки
знаменатель — диалоги, прошедшие фильтр длительности и содержания

98%
совпадение решений AI и экспертов
знаменатель — 500 сделок замороженного эталонного набора (golden set); пропуски критических нарушений проверяются отдельным порогом

−88%
стоимость ключевого LLM-сценария на 1 000 сделок
при неизменных схеме ответа, бенчмарке и гейтах качества

45–50% · ×1,2
типовых консультаций без оператора; рост квалифицированных лидов
знаменатель — подходящие типовые консультации; лиды — первый месяц против прежнего процесса

≥ 1 млн ₽
экономии на контакт-центре в месяц
стоимость обслуживания того же потока консультаций до и после автоматизации

90 / 464 / 246
production-пайплайнов, критериев чек-листов, полей и детекторов саммари
плюс 55 активных чек-листов и 32 конфигурации саммари

ОПЫТ · PRODUCTION-СИСТЕМЫ

AI-контроль качества всех коммуникаций с клиентом

КРУПНАЯ EDTECH-ПЛАТФОРМА (РФ) · ВЕСЬ КОММЕРЧЕСКИЙ БЛОК · СОБРАЛ КОНТУР ЦЕЛИКОМ

Раньше отдел контроля качества слушал выборку разговоров руками. Сейчас поток коммуникаций транскрибируется, фильтруется, оценивается по чек-листам и превращается в структурированные данные для руководителей, CRM и BI — по всем сотрудникам, которые говорят и пишут с клиентом. За этим стоят 90 production-пайплайнов, 55 активных чек-листов и 464 критерия оценки. Моё: логика платформы, системные промты, выбор моделей, пайплайны, правила оценки, выгрузка и калибровка на живом потоке.

Оценка перестала быть выборочной: 100% подходящих диалогов вместо ≤20%.

Аналитика, по которой принимают решения

ПРОДАЖИ · МАРКЕТИНГ · ПРОДУКТ И ТАРИФЫ · КОНКУРЕНТЫ

Задачи приходят от бизнеса как заявки с целью и экономическим обоснованием. Приношу ответ, пригодный для решения: что происходит, чем это подтверждается и чего эти данные не показывают. До 3 000 сделок и до полугода горизонта, по звонкам и переписке.

- **Почему просела конверсия за квартал.** Помесячное сравнение показало, что дело не в отделе продаж: трафик остывал, доля готовых к покупке падала, денежные возражения росли. → решения по закупке трафика, скрипту и презентации тарифов.
- **«Дорого» — не всегда про цену.** Двухуровневая таксономия причин: наверху «стоимость», а внутри группы преобладает недоступность рассрочки. Это работа не с ценой, а с платёжными инструментами.
- **Часть работы отдела продаж — не продажи.** Классификация входящих обращений по смыслу и по адресату вскрыла сервисную нагрузку в воронке продаж и неверный знаменатель конверсии отдела.

Каждое из этих исследований закончилось изменением процесса: закупка трафика, платёжные инструменты в оффере, пересчёт знаменателя конверсии отдела.

Voice AI

СИПУНИ · 10 ЛИНИЙ · ВМЕСТО ВНЕШНЕГО КОНТАКТ-ЦЕНТРА

Перевёл регламент контакт-центра в живой диалог: ветвление по скрипту, уточнения, перебивания, ответы из базы знаний. Телефония Сипуни, 10 линий одновременно, круглосуточно. Бюджет задержки — 3 секунды до первой реплики и до 10 секунд между репликами; сокращается не выбором модели, а тем, как рано начинается распознавание и сколько контекста уходит в запрос. Перебивания и ветка диалога заданы системным промтом, а не отдельной машиной состояний: правки регламента выходят без релиза, но политика диалога проходит evaluation так же, как код.

45–50% типовых консультаций без оператора, ×1,2 квалифицированных лидов в первый месяц, экономия от 1 млн ₽ в месяц.

Корпоративный RAG

DIFY · 37 ДОКУМЕНТОВ · 100+ ПОЛЬЗОВАТЕЛЕЙ В ДЕНЬ

Два ассистента на Dify — по продажам и по корпоративной документации, 37 документов в базе, поиск чисто векторный: ключевые слова дают выигрыш на точных артикулах и редких терминах, на корпусе такого размера гибридный поиск добавил бы сложность без результата. Отвечал за нарезку документов, сборку контекста, опору на источники и качество ответа. Основная работа — не в модели, а в подготовке документов и в том, чтобы ассистент честно говорил «в документах этого нет».

Ответ со ссылкой на документ меньше чем за 10 секунд вместо примерно пяти минут ручного поиска.

Оценка LLM и экономика моделей

4 СПЕЦИАЛИЗИРОВАННЫХ GOLDEN SET · РЕГРЕССИОННЫЕ ГЕЙТЫ

Отдельные наборы и метрики под балльную оценку, саммари, смысловую аналитику и нарушения. Экспертная калибровка, регрессии и разбор ошибок после каждой смены модели. Перенёс тот же валидированный результат с дорогой модели на компактную: при переезде первыми ломаются длинный контекст и стабильность формата, поэтому релиз гейтится по валидности схемы и полноте на критических нарушениях, а не по средней точности.

98% совпадения решений на замороженном наборе из 500 сделок; -88% стоимости ключевого сценария.

Lead Chat — собственный SaaS

LEADCHAT.ONLINE · ОПЛАТА ЗА РЕЗУЛЬТАТ: 50 ₽ ЗА СОЗДАННЫЙ ЛИД

Работающий продукт, а не демо: агент ведёт диалог по бизнес-цели, квалифицирует клиента, отрабатывает возражения, отвечает по базе знаний и совершает действия — записывает, собирает контакт, отдаёт лид в CRM. Моё: продуктовая концепция, AI-логика и ключевая архитектура. С технической стороны со мной работает партнёр-разработчик. Next.js, Node.js, tRPC, Prisma, PostgreSQL / pgvector, RAG, tool calling, фоновые задачи, биллинг.

Продукт работает публично: живой биллинг, оплата за созданный лид, а не за подписку.

ИНЖЕНЕРНЫЕ РЕШЕНИЯ И ИХ ЦЕНА

Транскрибация	Готовый сервис распознавания речи вместо своего Whisper: качество на спонтанной телефонной речи и диаризация из коробки, предсказуемая цена, ноль эксплуатации. Пересмотр — когда объём сделает разницу в цене больше стоимости эксплуатации.
Retrieval	pgvector в том же PostgreSQL вместо отдельной векторной БД: поиск рядом с бизнес-фильтрами, один запрос вместо двух хранилищ и синхронизации. Порог перехода — когда время поиска станет заметным на фоне времени модели.
Оценка качества	LLM-судья против замороженной экспертной разметки вместо обучения классификатора под каждый критерий: критерии меняются вместе с регламентом, а судья объясняет решение. Плата — дрейф при смене версии модели, поэтому набор заморожен.
Человек в петле	Там, где цена ошибки высокая, решение остаётся за человеком. Это осознанное ограничение, а не недоделка.

СТАНДАРТ, ПО КОТОРОМУ СДАЮТСЯ ИССЛЕДОВАНИЯ

- Контроль выборки: сколько записей вошло, сколько потеряно и почему — до всех процентов.
- Знаменатель у каждой доли, и при сравнении периодов он один и тот же.
- Наблюдение и объяснение разведены: гипотеза называется гипотезой и сопровождается способом проверки.
- Каждая категория подтверждается цитатой клиента, а заказчик может проверить разметку руками.
- Раздел ограничений: чего эти данные не показывают и какой вывод из них делать нельзя.

ЧТО ЗАКРЫВАЮ С НУЛЯ

- **Контур оценки качества LLM** — эталонный набор, метрики под задачу, регрессионные гейты, переезд между моделями без потери качества.
- **Голосовой агент на телефонии** — политика диалога, перебивания, бюджет задержки по этапам, расчёт ёмкости линий.
- **Корпоративный RAG** — подготовка документов, нарезка, сборка контекста, опора на источники и честный отказ.
- **Аналитическая функция на разговорных данных** — от заявки бизнеса с обоснованием до решения, которое приняли.

СТЕК

В PRODUCTION, ПОД МОЕЙ ОТВЕТСТВЕННОСТЬЮ

Python · TypeScript · FastAPI · Node.js · PostgreSQL · pgvector · SQL · Docker · OpenAI API · Anthropic API · GigaChat · YandexGPT · Dify · RAG · AI Agents · Structured Outputs · Tool Calling · LLM-as-a-Judge · Voice AI · CRM / BI · Git

ПИЛОТЫ И ЛАБОРАТОРНЫЕ ЗАДАЧИ

LangChain · LangGraph · Qdrant · vLLM · Hugging Face · Fine-tuning / LoRA · RAGAS · A/B Testing · Guardrails · Prompt Injection · NLP

ПУБЛИЧНЫЙ КОД · GITHUB.COM/94SPEC · 7 200 СТРОК ИСХОДНИКОВ, 393 ТЕСТА, CI ЗЕЛЁНЫЙ

llm-evaluation-harness	Фреймворк оценки LLM без внешних зависимостей: четыре класса задач, типизированная таксономия ошибок, bootstrap-интервалы, релизные гейты. · 2 300 строк, 125 тестов
golden-set-builder	Построение эталонных наборов: стратифицированный отбор с отчётом о недоборе, согласованность разметчиков до метрик модели, разрешение споров без голосования большинством, заморозка с хешем на запись. · 1 500 строк, 63 теста
prompt-contracts	Промты как версионируемые контракты: схема ответа, правила отказа, бюджет размера, golden set. Оценка влияния изменения на релиз вместо текстового дифа. · 850 строк, 58 тестов
voice-agent-runtime	Голосовой рантайм: политика диалога таблицей переходов, перебивание без доигрывания фразы, бюджет задержки по этапам против 3 и 10 секунд, ёмкость линий. · 1 100 строк, 73 теста
rag-grounding-lab	Цена нарезки, сравнение стратегий поиска и проверка ответа по предложениям: выдуманное число ловится отдельно, честный отказ засчитывается как обоснованный. · 1 200 строк, 54 теста
conversation-intelligence-lab	Та же методика аналитики на синтетических данных: строгий контракт вывода, трассируемые агрегаты, воспроизводимый отчёт. · 300 строк, 20 тестов

Каждая цифра здесь приводится с выборкой, знаменателем и правилом сравнения — детали разберу на созвоне. Актуальная версия резюме: 94spec.github.io/cv/