

БАКАЛАВРСКАЯ РАБОТА · 09.03.04 ПРОГРАММНАЯ ИНЖЕНЕРИЯ

Мультиагентная система на базе генеративных языковых моделей

для персонализированного
сопровождения процесса
бронирования гостиничных услуг

ВЫПОЛНИЛ

Мухин Дмитрий Игоревич

Направление 09.03.04 «Программная инженерия»

НАУЧНЫЙ РУКОВОДИТЕЛЬ

Бородин А. В.

к. э. н., доцент

2026

г. Йошкар-Ола

Диалог — критический барьер в воронке бронирования

Цифровые каналы — основное пространство взаимодействия гостя со средством размещения.

- 01** Гость, не получив быстрого ответа на уточняющий вопрос, прерывает бронирование и уходит на сторонние ресурсы.
- 02** У независимого отеля нет круглосуточной мгновенной поддержки по всем каналам.
- 03** Сценарные чат-боты не справляются со свободной речью пользователя и спецификой конкретного средства размещения.

ПИЛОТ · ПЛАТФОРМА TRAVELLINE

ВОРОНКА БРОНИРОВАНИЯ

Посещение канала отеля

Интерес к размещению

Уточнение условий

← потеря

Переход к оплате

Завершённое
бронирование

Этап уточнения условий — узкое место, где гость без мгновенного ответа покидает воронку.

Цель и задачи работы

ЦЕЛЬ

Мультиагентная LLM-система персонализированного сопровождения процесса бронирования.

ЗАДАЧИ

01 Анализ предметной области
и формирование требований

02 Архитектура
шлюз сообщений, мосты,
переиспользуемый runtime

03 Персонализация
промпт-цепочки отелей + RAG +
контекст

04 Реализация и развёртывание
опытно-промышленного контура

05 Экспериментальная оценка
контур качества, сценарии и события мониторинга

Диалог как транзакционный интерфейс управления бизнес-логикой продукта

Перенос парадигмы агентных систем из области автоматизации разработки ПО в сферу эксплуатации прикладных бизнес-решений: совмещение процессов **консультирования и транзакционного управления** в едином сессионном контексте.

01

Консультирование + управление

Интеграция механизмов ответов по базе знаний (RAG) и выполнение операций модификации данных в рамках одной диалоговой сессии.

02

Human-in-the-loop

Обязательное двухэтапное подтверждение транзакций оператором гарантирует безопасность и консистентность данных в целевой системе.

03

Замкнутый операционный контур

Реальные диалоги → динамическая таксономия проблем → приоритизированный бэклог правок БЗ, промптов, инструментов.

В отличие от классических информационных чат-ботов и изолированных спецификаций ADK / MCP / LiteLLM — предложенный подход впервые объединяет указанные свойства в единый замкнутый контур эксплуатации.

Сравнительный анализ с существующими решениями

	СЦЕНАРНЫЙ БОТ	ОДНОАГЕНТНЫЙ LLM	РАЗРАБОТАННАЯ СИСТЕМА
Свободная речь	нет, жёсткое дерево диалога	да, один промпт	да, специализация агентов
Адаптация под регламент отеля	ручная	деградирует с ростом функций	промпт-цепочки + RAG
Транзакционные операции	отсутствуют	риск ошибочных вызовов и утечки	разделение на безопасные/ небезопасные и верификация изменений

Рассмотренные аналоги не обеспечивают **одновременную поддержку** распознавания естественной речи, глубокой адаптации контекста и безопасной модификации данных в СУБД.

06 АРХИТЕКТУРА И ПОТОКИ ДАННЫХ

СЛОИ

- 01 Клиентские интеграции
- 02 Шлюз сообщений (Message Gateway)
- 03 Среда выполнения (Agent Runtime)
- 04 Хранилище данных и СУБД (RAG)

СТЕК

GOOGLE ADK

LITELLM

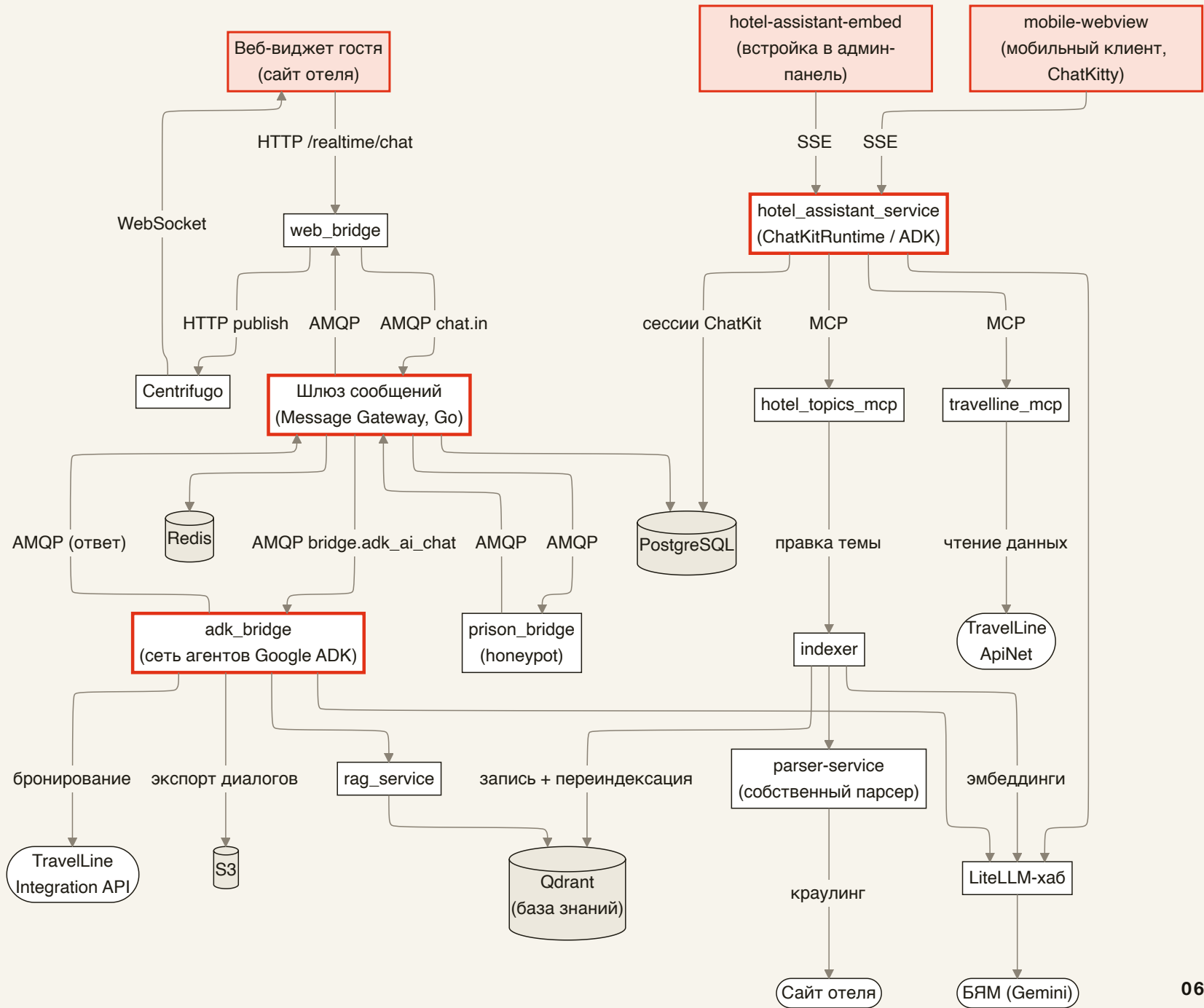
MCP

GO

FASTAPI

QDRANT

KUBERNETES



Переиспользуемые компоненты: runtime и шлюз

КОМПОНЕНТ 1 · СРЕДА ВЫПОЛНЕНИЯ АГЕНТОВ — ОТДЕЛЬНАЯ БИБЛИОТЕКА

Потоковая передача ответов

по мере генерации, без накопления полного текста

Агрегация инструментов MCP

несколько серверов → одно пространство имён · отложенная загрузка схем

Постоянные сессии · субагенты

состояние сессии сохраняется при перезапуске системы

Human-in-the-loop

блокировка выполнения сценария до валидации оператором (HITL)

КОМПОНЕНТ 2 · ШЛЮЗ НА GO

Управление сессиями и историей контекста

источник контекста для агентов

Абстрагирование транспортных интерфейсов

каналы и оркестрация заменяемы независимо

Восстановление диалога

автоматическое возобновление сессии при обрыве соединения

Оба компонента переиспользуются сервисами платформы и не зависят от конкретного канала или среды оркестрации.

РЕАЛИЗОВАННЫЕ ИНЖЕНЕРНЫЕ РЕШЕНИЯ

01

Оптимизация передачи SSE обратными прокси

потокковое вещание данных клиенту в обход промежуточных буферов

02

Восстановление сессии

диалог продолжается после разрыва соединения

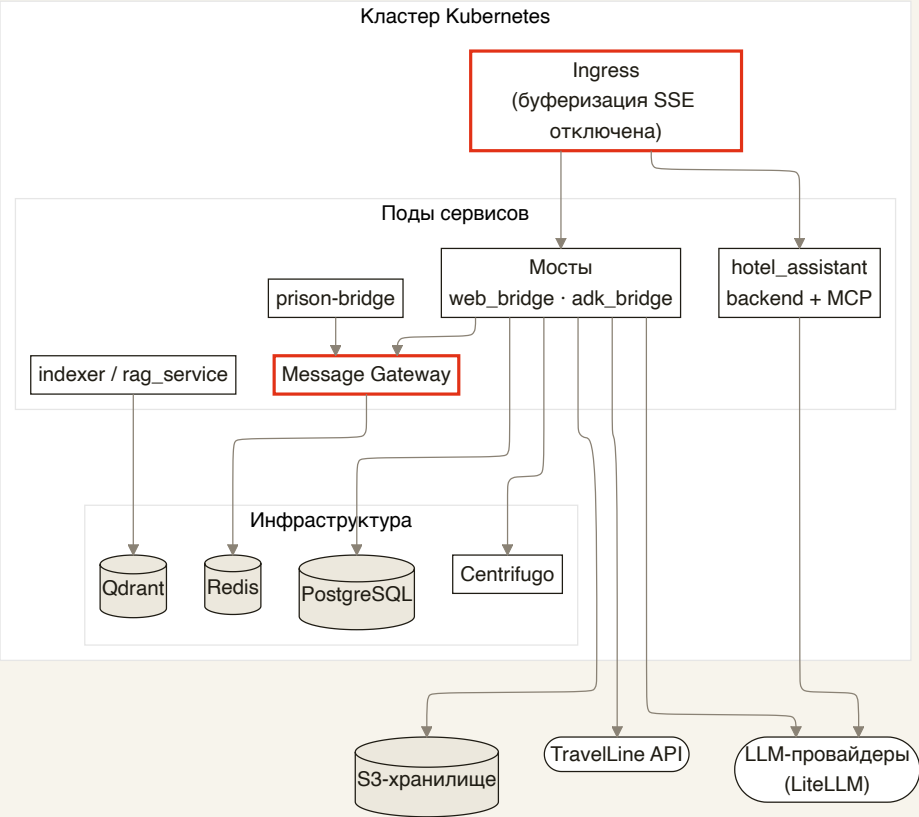
03

Параллельные вызовы инструментов

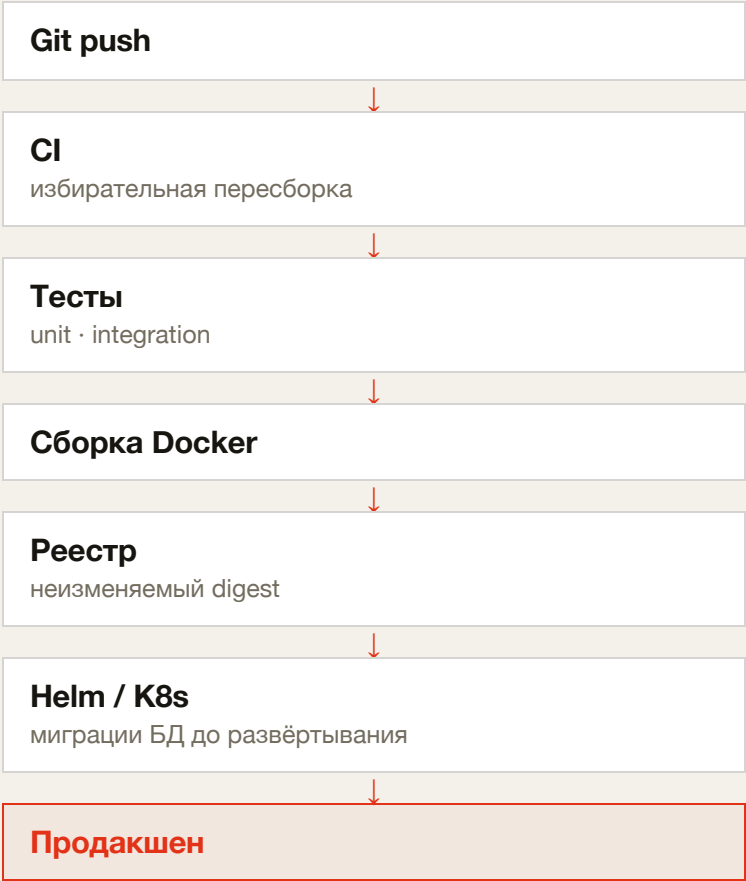
независимые инструменты исполняются одновременно

От Git до продакшена + развёртывание в K8s

РАЗВЁРТЫВАНИЕ В KUBERNETES

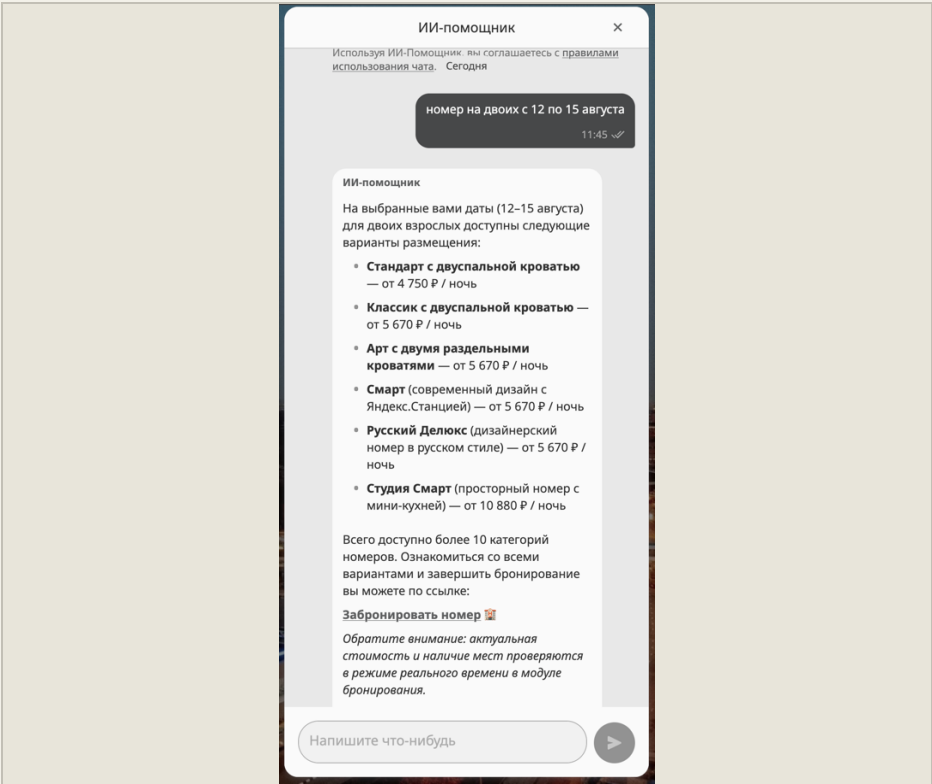


КОНВЕЙЕР ДОСТАВКИ

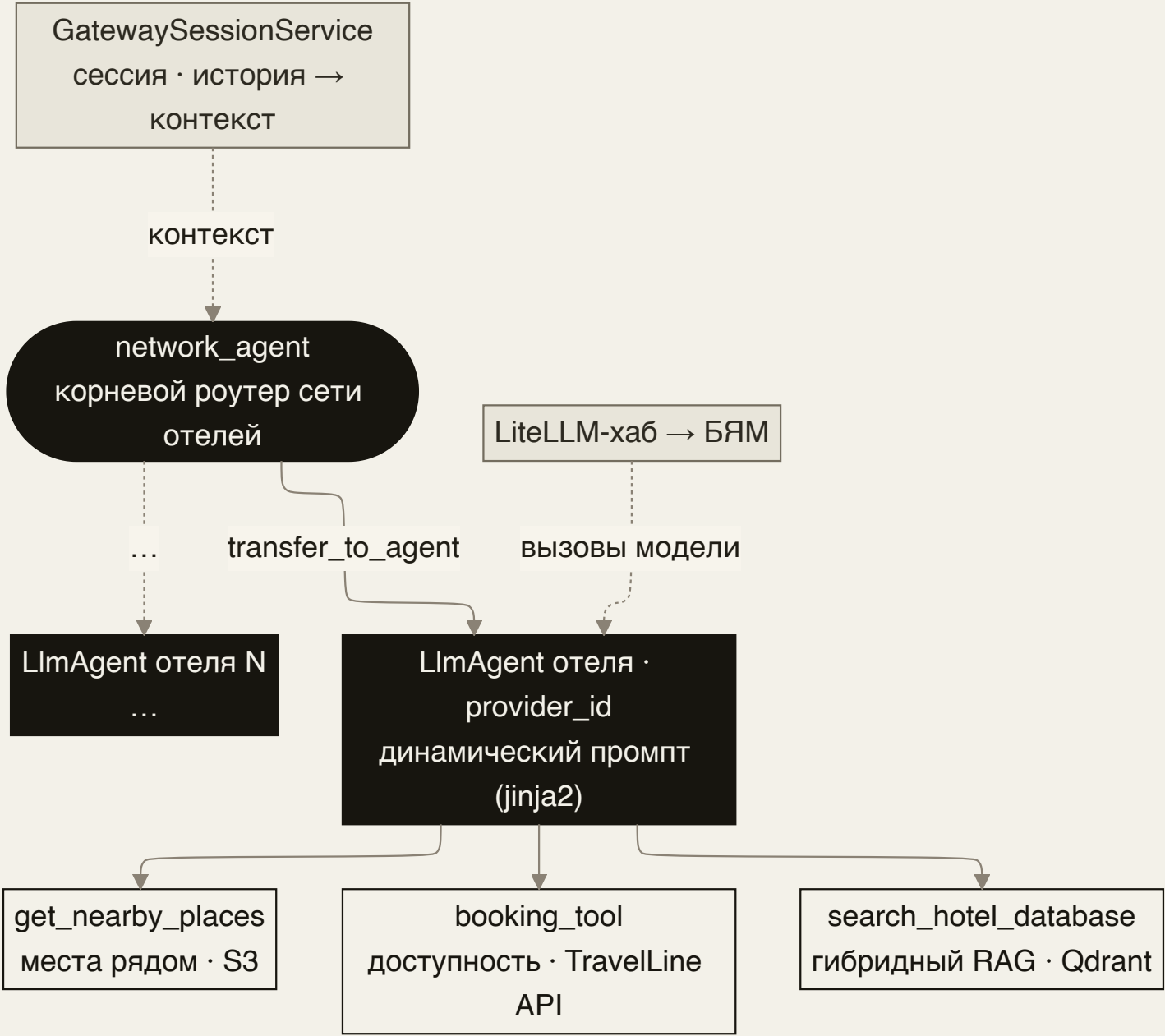


Гость: консультирование и бронирование

Сеть отелей → агент отеля → инструменты. Только чтение.

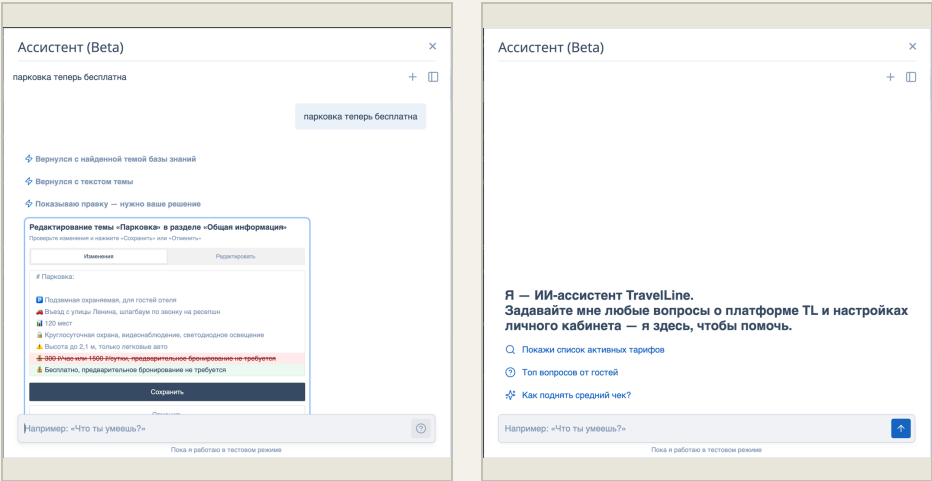


ВИДЖЕТ АССИСТЕНТА ГОСТЯ



Отельер: управление базой знаний с HITL

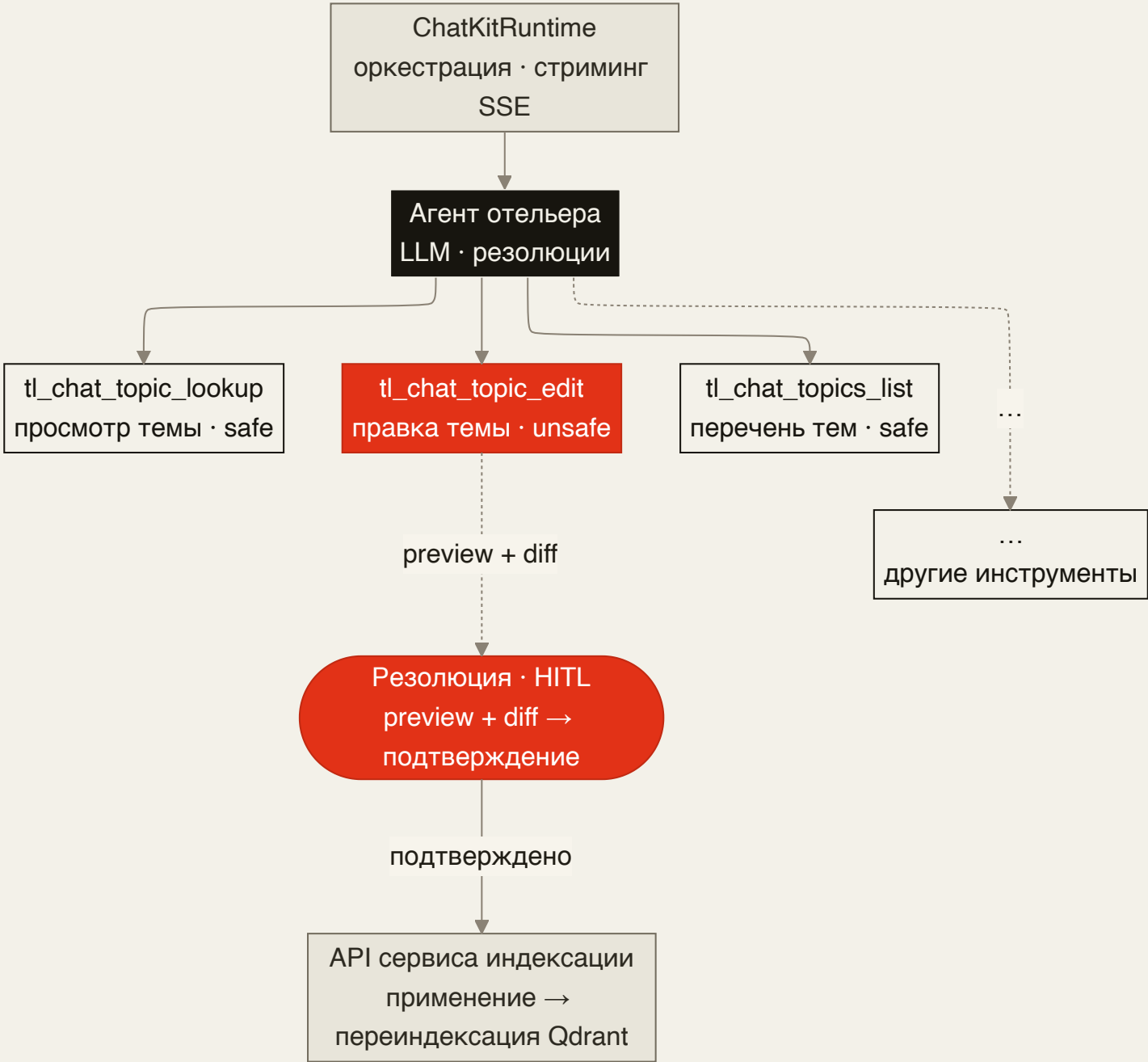
Runtime → агент → MCP-инструменты; **unsafe-правка** через HITL.



HITL · DIFF

ИНТЕРФЕЙС ОТЕЛЬНОГО

☐ чтение ☒ запись



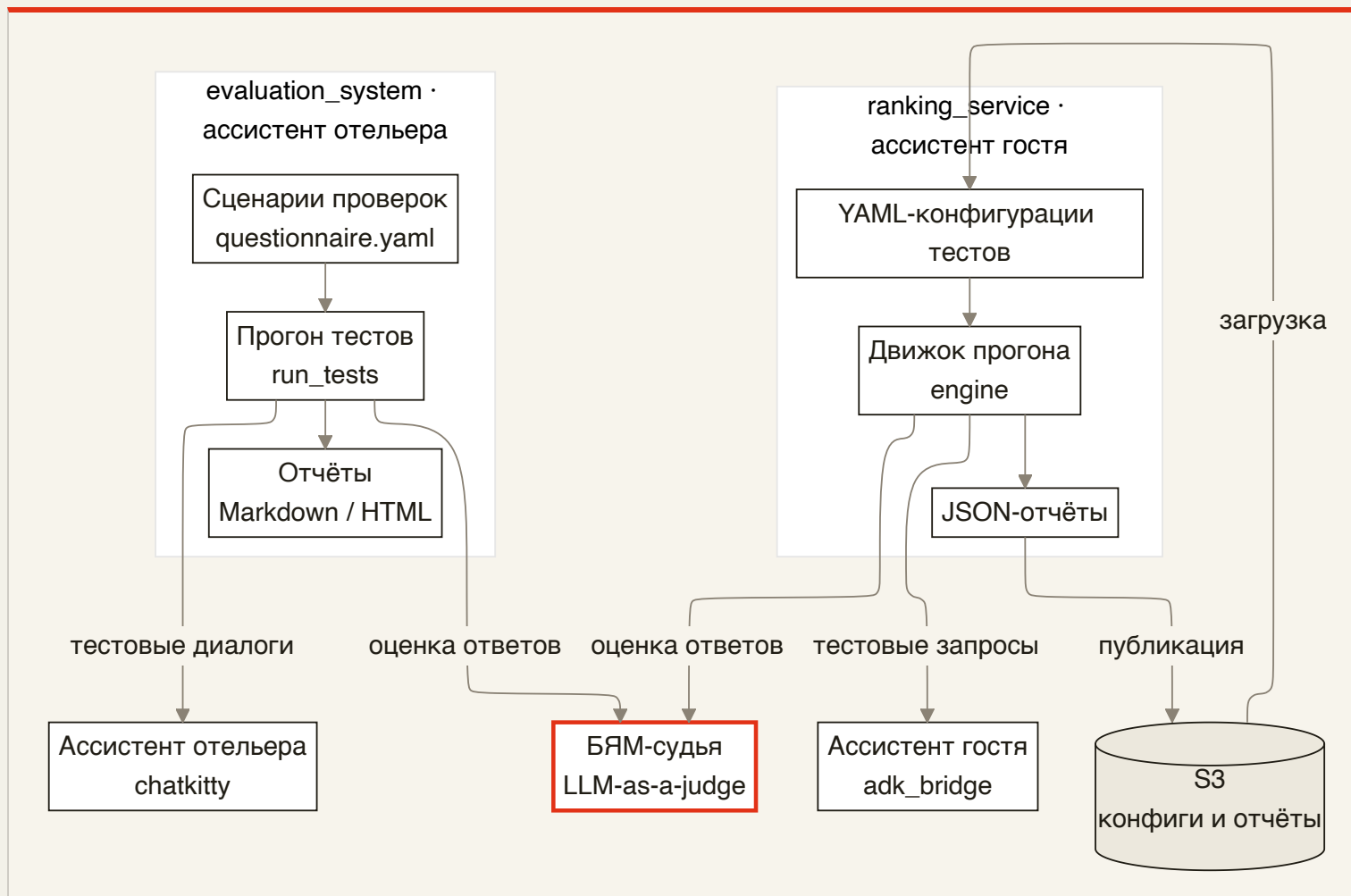
Трёхуровневый контур контроля качества

ТРИ УРОВНЯ КОНТРОЛЯ

- 01 Детерминированные проверки** ·
Утверждения (assertions): проверка точных совпадений, подстрок, запрещённых фрагментов и трасс вызова инструментов. Объективны, но не покрывают вариативность формулировок.
- 02 LLM-судья** — гибкая оценка перефразирований при узкой инструкции и T=0. Гибок, но подвержен ошибкам.
- 03 Непрерывный мониторинг эксплуатации** — реальный пользовательский опыт, невозпроизводимый в искусственной тестовой среде.

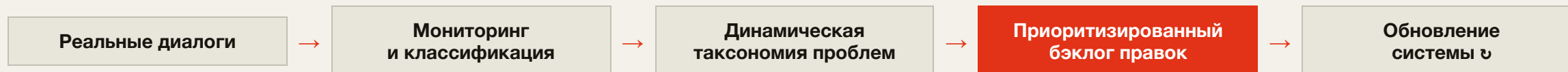
RANKING_SERVICE · АССИСТЕНТ ГОСТЯ

Отдельный сервис ранжирования: взвешенные критерии и версионизируемые отчёты. Регрессия **128 тестовых сценариев** в составе 32 тест-кейсов — полное прохождение в эталонном окружении.



КОНТУР ОЦЕНКИ КАЧЕСТВА · EVALUATION_SYSTEM + RANKING_SERVICE

Замкнутый контур качества: от диалогов к бэклогу



128

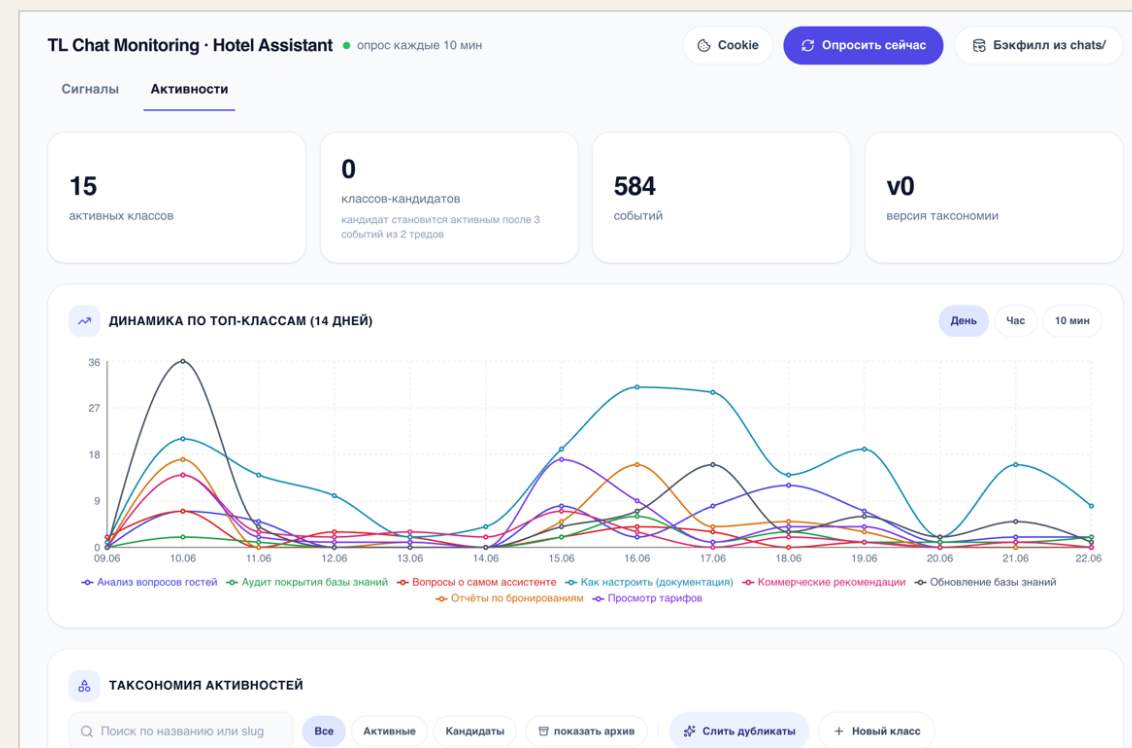
тестовых сценариев в составе 32 тест-кейсов · полное прохождение в эталонном окружении

176

событий мониторинга на 107 диалогах · 15 классов

УРОВНИ КОНТРОЛЯ

- Детерминированные проверки · LLM-судья · непрерывный мониторинг
- Регрессионные тесты верифицируют трассировку выполнения внешних инструментов
- Дедупликация инцидентов: точное совпадение сигнатур и семантический анализ (оценка близости эмбедингов; спорные случаи разрешаются LLM-судьёй)
- Классификация дефектов масштабируется в автоматическом режиме



ДАШБОРД МОНИТОРИНГА ДИАЛОГОВ · СИГНАЛЫ И АКТИВНОСТИ

Практическая апробация и результаты опытно-промышленной эксплуатации

600+

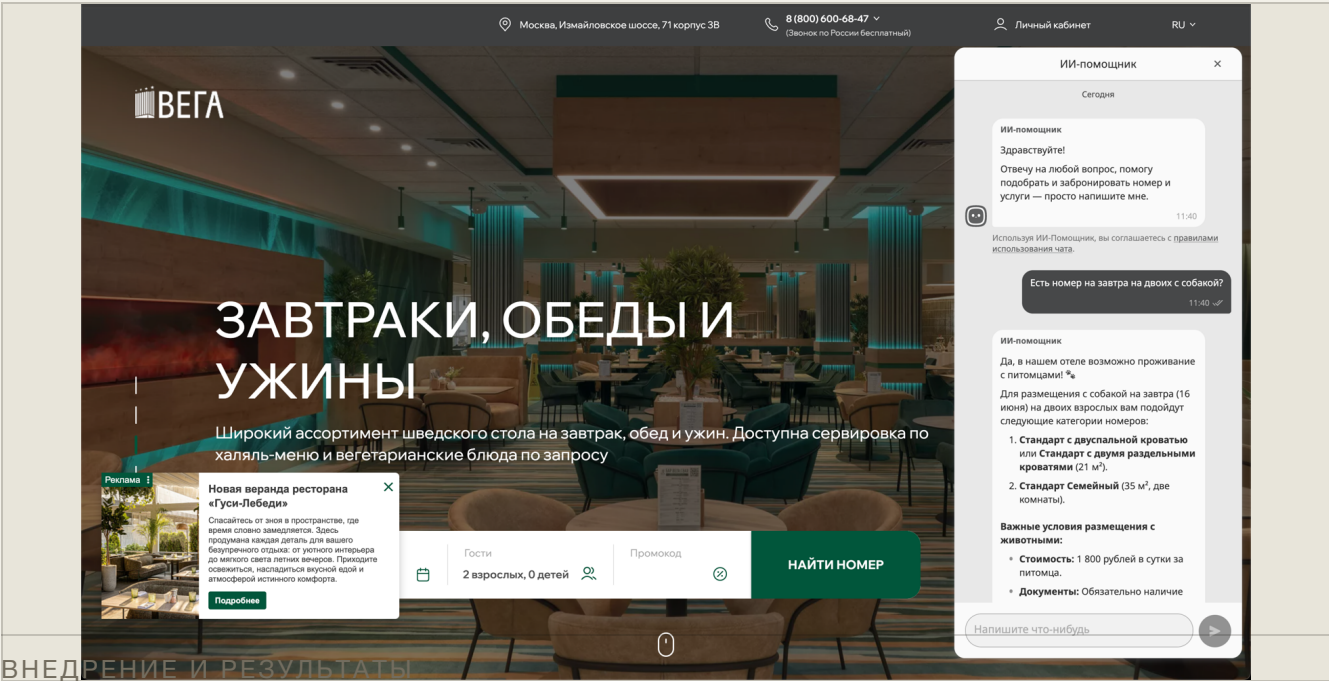
отелей — гость и отельер

50 000+

диалогов гостей

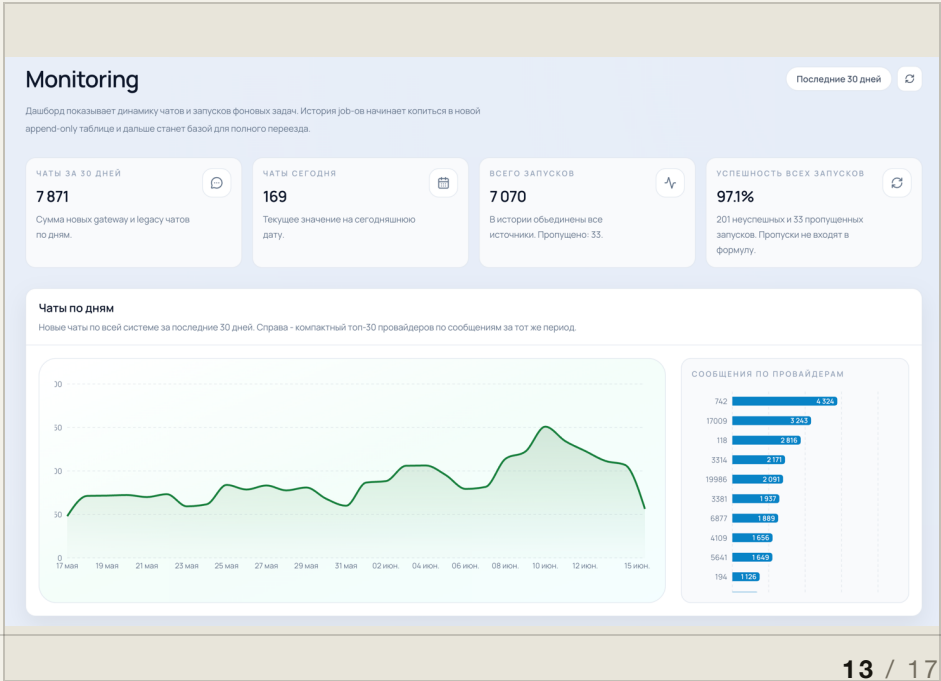
1 000+

диалогов отельера



ВНЕДРЕНИЕ И РЕЗУЛЬТАТЫ

ВЕБ-ВИДЖЕТ АССИСТЕНТА ГОСТЯ НА САЙТЕ ОТЕЛЯ



МОНИТОРИНГ ИСПОЛЬЗОВАНИЯ · ДИНАМИКА ДИАЛОГОВ

Результаты работы

ИТОГИ ПО ЗАДАЧАМ

- **Мультиагентная система сопровождения бронирования**
ассистент гостя (Google ADK) и ассистент отеля
- **Переиспользуемые компоненты**
библиотека среды выполнения агентов (agent runtime) со специализированным протоколом и шлюз сообщений на Go
- **Обеспечение персонализации и надёжности транзакций**
промт-цепочки отелей + RAG (Qdrant); двухшаговое подтверждение изменяющих операций человеком (HITL)
- **Автоматизированное развёртывание**
Docker · Kubernetes · CI/CD; потоковая доставка ответов по SSE
- **Трёхуровневый контур контроля качества**
детерминированные проверки · LLM-судья · мониторинг → приоритизированный бэклог доработок
- **Опытно-промышленная эксплуатация**
система успешно прошла опытную эксплуатацию
- **Научный результат**
модель диалогового ассистента как интерфейса консультирования и управления продуктом

В ЦИФРАХ

600+
отелей — гость и отельер

50 000+
диалогов гостей

1 000+
диалогов отеля

128
сценариев · полное прохождение в эталонном окружении

176
событий мониторинга · 107 диалогов · 15 классов

Новизна · инженерное ядро · значимость

ВЫПОЛНЕННЫЕ ЗАДАЧИ

01 Анализ и требования

Ф1–Ф5 / Н1–Н5

02 Архитектура

Go-шлюз, мосты, ассистент гостя на ADK, переиспользуемый runtime

03 Персонализация

промт-цепочки, RAG, контекст

04 Развёртывание и приёмочные испытания

опытно-промышленного контура

05 Качество

128 сценариев + 176 событий мониторинга

НОВИЗНА

Ассистент как интерфейс консультирования
и управления — HITL, замкнутый контур

ИНЖЕНЕРНОЕ ЯДРО

Переиспользуемый агентный runtime + Go-
шлюз

ВНЕДРЕНИЕ

600+ · 50 000+ · 1 000+

Направления развития

01

Системы фильтрации спама и противодействия флуду

Интеграция средств автоматической изоляции аномальных сессий (honeypot).

02

Абляционный анализ разработанной архитектуры

Сравнительное тестирование многоагентных и одноагентных конфигураций на фиксированном наборе данных.

03

Полный набор инструментов TravelLine

Интеграция в агентный каркас.

04

Долговременная память

Предпочтения гостя между сессиями.

05

Авто-оптимизация промптов

По данным мониторинга диалогов.

06

Экономический эффект

Оценка по метрикам конверсии.

Спасибо за внимание

Мультиагентная LLM-система
персонализированного сопровождения
бронирования гостиничных услуг.

Готов ответить на ваши вопросы

ВЫПОЛНИЛ

Мухин Дмитрий Игоревич

НАУЧНЫЙ РУКОВОДИТЕЛЬ

Бородин А. В. · к. э. н., доцент