

Стоимость и бюджеты (Cost)

Раздел **Cost** (/cost) показывает расход по агентам и позволяет задать лимиты.

Что учитывается

На шагах workflow и вызовах агентов:

- токены входа / выхода (LLM);
- задержка (latency, ms);
- внешние API-вызовы;
- оценка в USD (по настроенным тарифам окружения).

Работа в UI

1. Откройте **Cost**.
2. Просмотрите таблицу **Cost by agent** (spend, requests, latency).
3. Если данных нет — сначала выполните workflow (No cost data yet — run a workflow first).
4. Установите бюджет:
 - выберите агента;
 - укажите лимит в USD;
 - нажмите **Set budget**.

Баннер предупредит, если расход близко к лимиту или лимит **EXCEEDED**.

Алерты

- WebSocket: `cost.budget.alert`.
- При жёстком превышении следующие шаги сценария могут остановиться.

Через API

```
# Расход агента
curl "http://localhost:8001/api/v1/cost/agent/$AGENT_ID?period=last_7_days" \
  -H "X-API-Key: $TENANT_API_KEY"

# Бюджет
curl -X POST http://localhost:8001/api/v1/cost/budget \
  -H "Authorization: Bearer $JWT" \
  -H "X-API-Key: $TENANT_API_KEY" \
  -H "Content-Type: application/json" \
  -d '{
    "agent_id": "'$AGENT_ID'",
    "limit_usd": 100,
    "period": "monthly",
    "alert_threshold": 0.8
  }'
```

`alert_threshold: 0.8` — предупреждение при 80% бюджета.

Регламент оператора

- Ежедневно: агенты near limit / EXCEEDED.
- Еженедельно: пересмотр лимитов с владельцами агентов.
- При внезапном росте spend — корреляция с **Anomalies** и **Audit** (возможный цикл или атака).

Подробнее (EN): Cost.