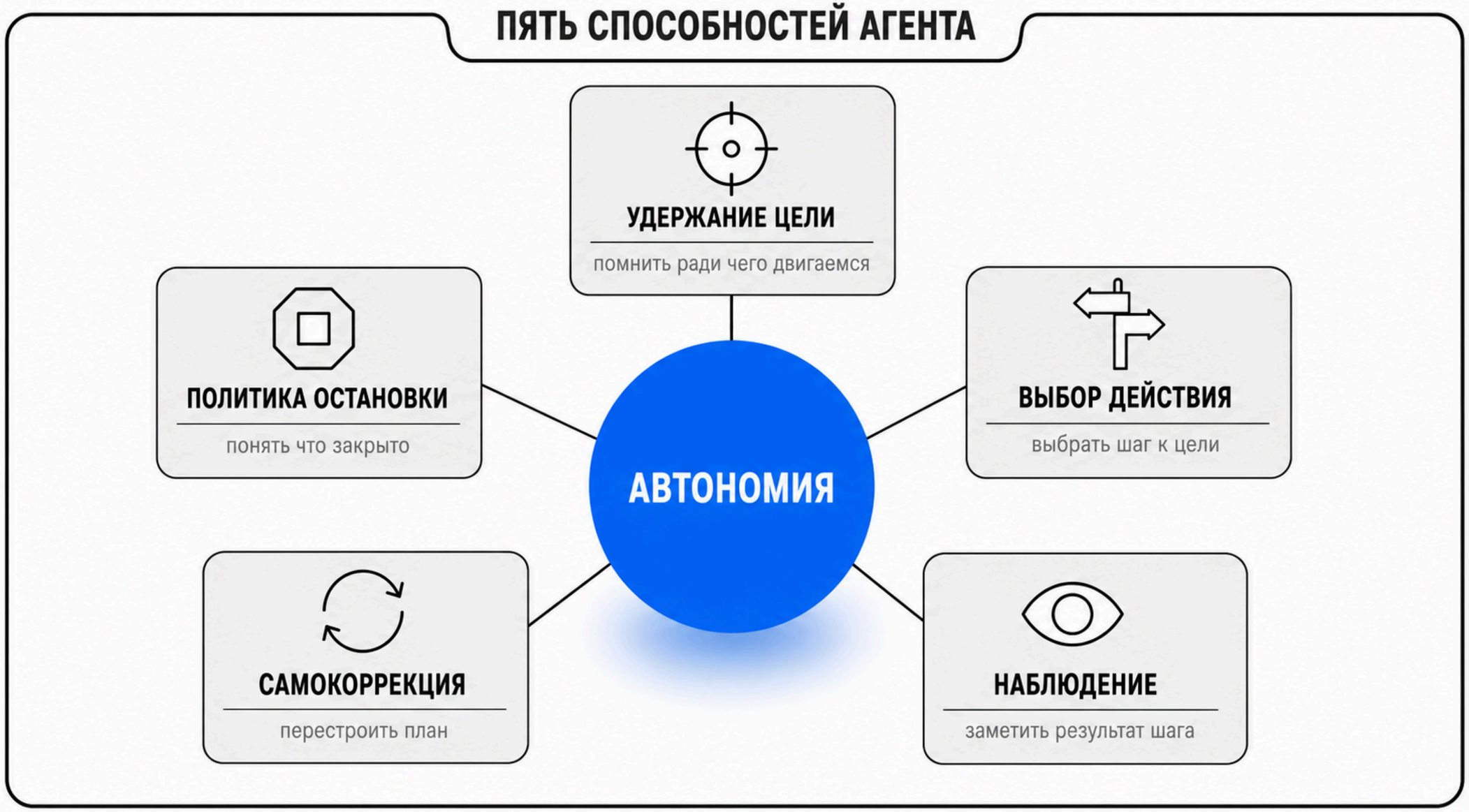
МОДУЛЬ 6

# АВТОНОМИЯ

Как собрать и на чём она держится

# Что такое автономия


АВТОНОМИЯ
АВТОНОМИЯ = СПОСОБНОСТИ МОДЕЛИ × КАЧЕСТВО ОБВЯЗКИ × СООТВЕТСТВИЕ ЗАДАЧЕ
ЗНАНИЯ И РАССУЖДЕНИЯ
что модель знает и как мыслит
ОБОБЩЕНИЕ
перенос на новые ситуации
НАДЁЖНОСТЬ
стабильность и калибровка
СПОСОБНОСТИ МОДЕЛИ
КАЧЕСТВО ОБВЯЗКИ
АВТОНОМИЯ
СООТВЕТСТВИЕ ЗАДАЧЕ
КОНТЕКСТ И ПАМЯТЬ
какой контекст и история доступны модели
ЗАЩИТЫ И ОГРАНИЧЕНИЯ
границы, проверки, безопасность
ИНСТРУМЕНТЫ И ДАННЫЕ
доступ к внешним источникам и API
ОЦЕНКА И ОБРАТНАЯ СВЯЗЬ
как оцениваем и улучшаем результат
ЯСНОСТЬ ЦЕЛИ
насколько чётко определена задача
СЛОЖНОСТЬ ЗАДАЧИ
насколько задача в зоне возможностей модели
КРИТЕРИИ УСПЕХА
понятно, как выглядит хороший результат
СЛАБОЕ ЗВЕНО ЛЮБОГО ЭЛЕМЕНТА ОГРАНИЧИВАЕТ ОБЩУЮ АВТОНОМИЮ
УЛУЧШАЙТЕ ТО, ЧТО ДАЁТ НАИБОЛЬШИЙ РОСТ ВАШЕЙ ЗАДАЧИ


ПЯТЬ СПОСОБНОСТЕЙ АГЕНТА
УДЕРЖАНИЕ ЦЕЛИ
помнить ради чего двигаемся
ПОЛИТИКА ОСТАНОВКИ
понять что закрыто
АВТОНОМИЯ
ВЫБОР ДЕЙСТВИЯ
выбрать шаг к цели
САМОКОРРЕКЦИЯ
перестроить план
НАБЛЮДЕНИЕ
заметить результат шага

# Пять способностей агента

| СПОСОБНОСТЬ | ЧТО ЗНАЧИТ |
|---|---|
| **1 · Удержание цели** | помнить, ради чего двигаемся, на десятом шаге |
| **2 · Выбор действия** | из всех возможных шагов взять тот, что приближает к цели |
| **3 · Наблюдение** | заметить, совпал результат шага с ожиданием |
| **4 · Самокоррекция** | после плохого результата перестроить план, не повторить |
| **5 · Политика остановки** | понять, когда задача закрыта, и не продолжать |

# Какой вклад модели

# Модель оптимизирована на полезный одношаговый ответ. Не на удержание цели

# Модель — эхо обучающих текстов

АНДРЕЙ КАРПАТИ · ПОДКАСТ ДВОРКЕША ПАТЕЛЯ, 2025

# ПЯТЬ ИСКАЖЕНИЙ МОДЕЛИ

## ПОДХАЛИМСТВО

угодить пользователю
важнее правды

## УГАДЫВАНИЕ

выглядеть знающим
важнее признать незнание

## ЛОЖНОЕ ЗАВЕРШЕНИЕ

закрыть тред важнее
довести до результата

## ЯКОРЕНИЕ

защитить первое решение
важнее обновить картину

## УХОД ОТ ОТВЕТСТВЕННОСТИ

снять с себя
промежуточный результат

# Пять искажений модели

| ИСКАЖЕНИЕ | КАК ПРОЯВЛЯЕТСЯ |
|---|---|
| **Подхалимство** | Сказали «фокус на B2B» — план будет про B2B. Даже если данные про B2C |
| **Угадывание** | «Когда конференция?» — выдумает дату уверенно. Без «не знаю» |
| **Ложное завершение** | «Обработал 20 писем» — три не открыл, два классифицировал по неверному правилу |
| **Якорение** | На шестом шаге появилось противоречие. Игнорирует — план уже выбран |
| **Уход от ответственности** | «Похоже исправлено», «known limitation», «good stopping point» |

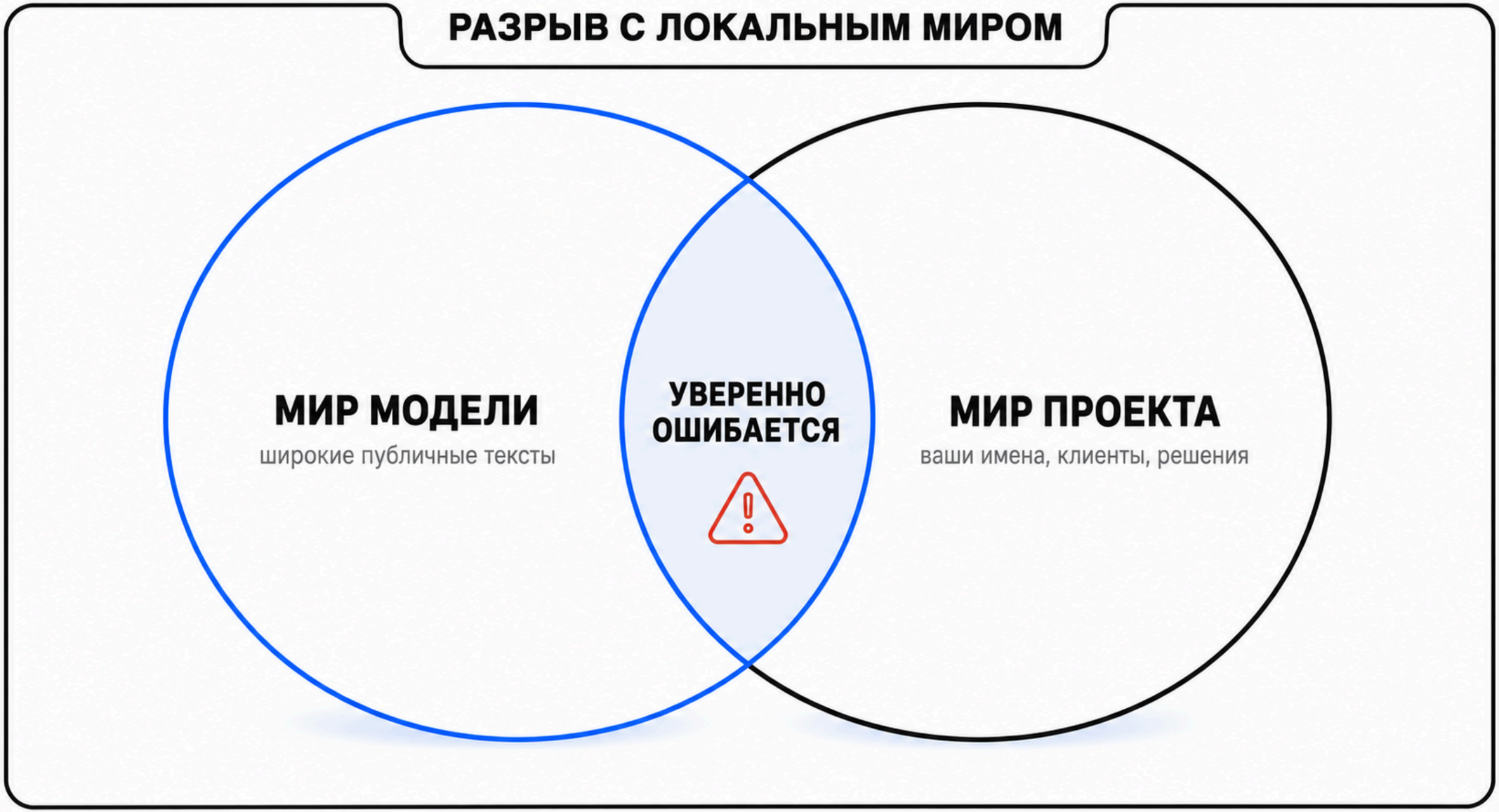
# Структурные потолки

PRODUCT SENSE

# Нет постоянной памяти. Закрылась сессия — модель забыла всё

# Конечное окно контекста. Большая задача физически не помещается

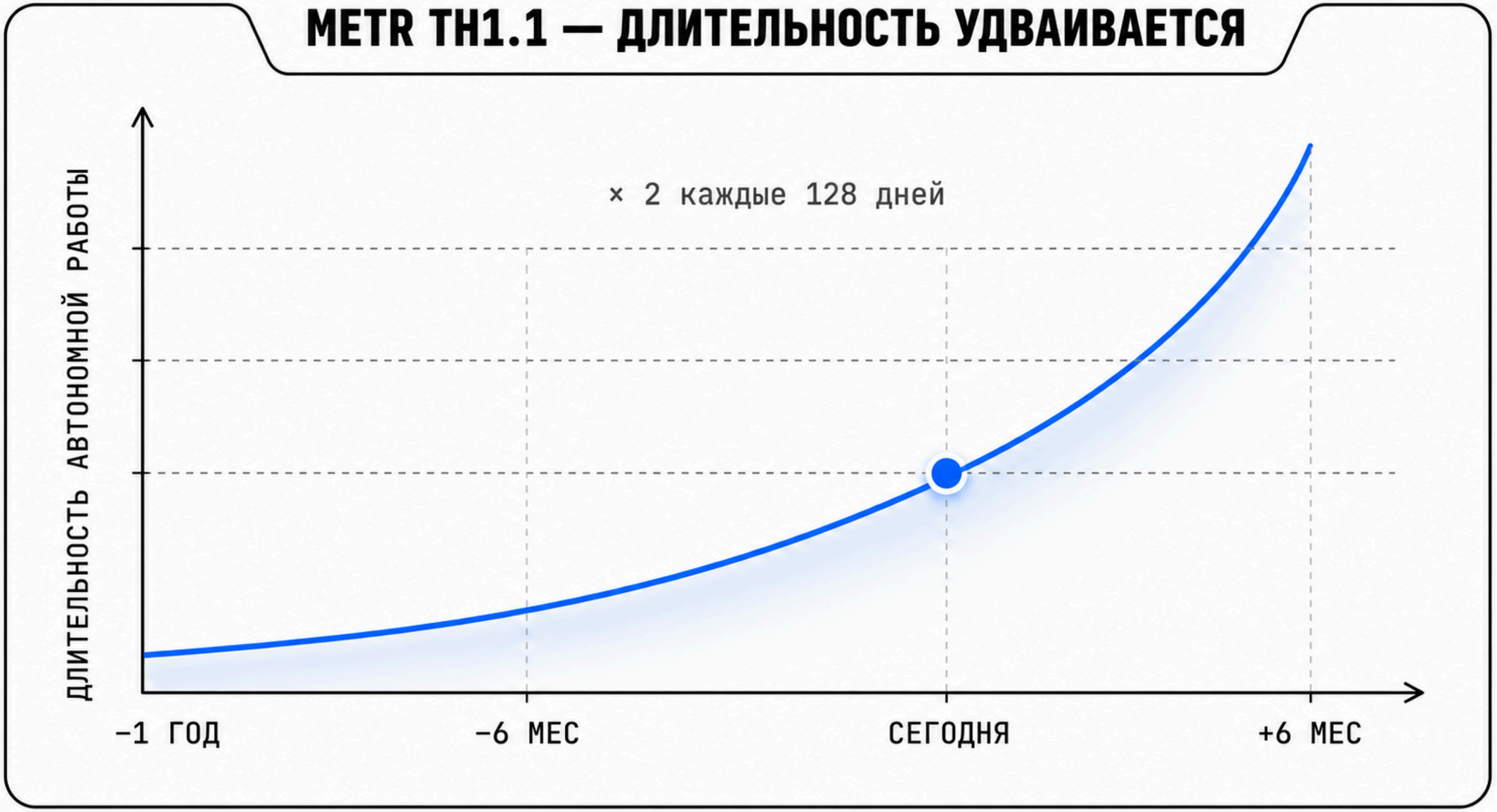
# Обучение замёрзло во времени. Мир движется — модель не знает


РАЗРЫВ С ЛОКАЛЬНЫМ МИРОМ
МИР МОДЕЛИ
широкие публичные тексты
УВЕРЕННО ОШИБАЕТСЯ
МИР ПРОЕКТА
ваши имена, клиенты, решения

# В новом мире модель уверенно ошибается. Память и источники — не опция

# Какое из пяти искажений вы видели в своей работе с агентом? Одной строкой

# Фронтир сегодня


METR TH1.1 — ДЛИТЕЛЬНОСТЬ УДВАИВАЕТСЯ
ДЛИТЕЛЬНОСТЬ АВТОНОМНОЙ РАБОТЫ
× 2 каждые 128 дней
−1 ГОД
−6 МЕС
СЕГОДНЯ
+6 МЕС

# Лучшее на май 2026

| МОДЕЛЬ | БЕНЧМАРК | РЕЗУЛЬТАТ |
|---|---|---|
| **Sonnet 4.6** | OSWorld-Verified | 72,5 % |
| **MCP-Atlas + Opus 4.5** | OSWorld | 62,3 % |
| **GPT-5.5** | SWE-Bench Pro | 58,6 % |

случаев Auto Mode пропускает явное подтверждение

# 17%

Anthropic, апрель 2026

# Главный рычаг — обвязка, а не сила модели

# Качество обвязки

# ТРИ УРОВНЯ ПРИНУЖДЕНИЯ

## L1 — МЫ ПОПРОСИЛИ

правило в CLAUDE.md или скилле —
модель может проигнорировать

≈ 30%

## L2 — МЫ ДАЛИ ИНСТРУМЕНТ

скилл, субагент — агент решает,
использовать или нет

≈ 60%

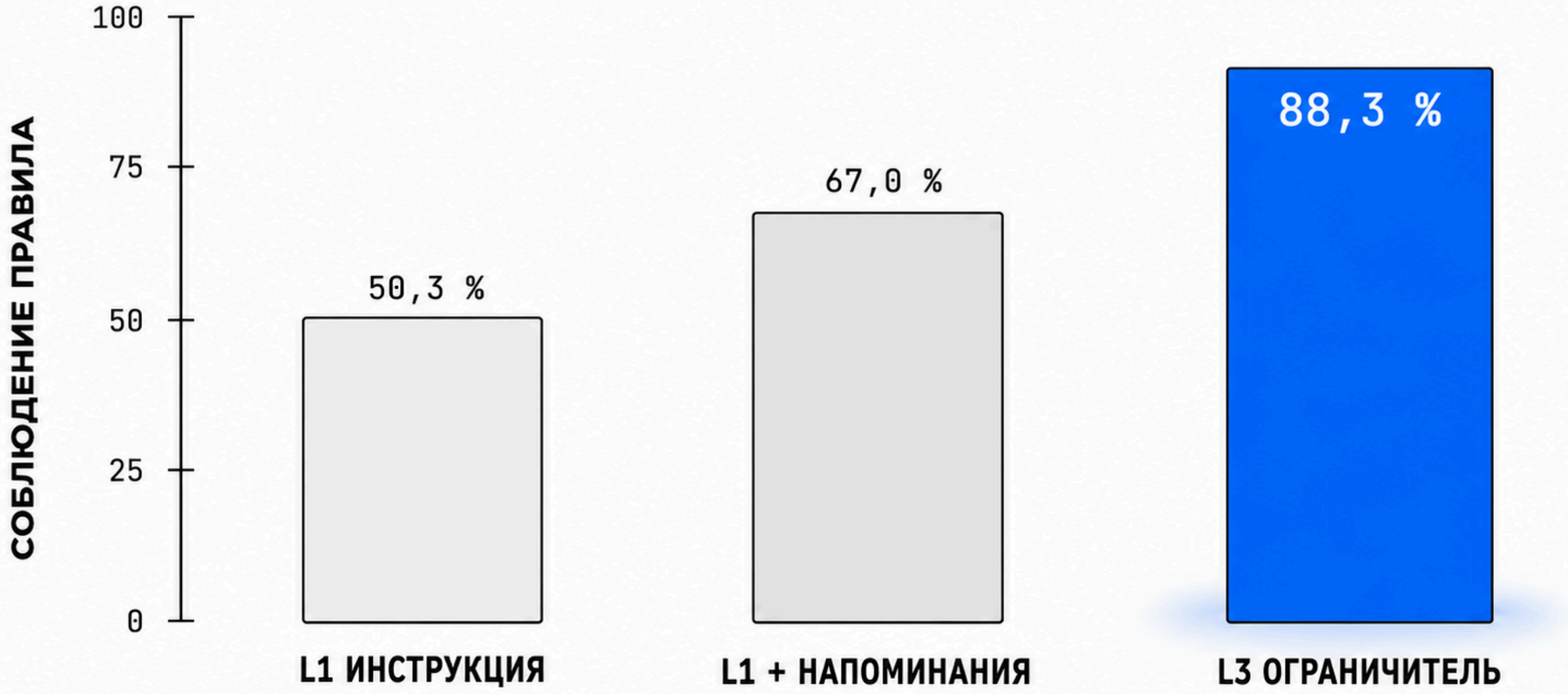
## L3 — МЫ НЕ ДАЛИ СДЕЛАТЬ НЕПРАВИЛЬНО

хук, sandbox, ограничение прав —
агент не может обойти

≈ 95%+

# Тот же паттерн — три уровня соблюдения. Цифры на следующем слайде

# ТО ЖЕ ПРАВИЛО — ТРИ УРОВНЯ




ContextCov on SWE-bench Lite, 2025

# context files tend to reduce task success rates

AGENTS.MD SPEC · AGENTS.MD, 2026

# Claude treats them as context, not enforced configuration

ANTHROPIC · ДОКУМЕНТАЦИЯ CLAUDE CODE, 2026

# Один паттерн — разный синтаксис

| УРОВЕНЬ | CLAUDE CODE | CODEX | OPENCODE |
|---|---|---|---|
| **L1** | CLAUDE.md, .claude/rules | AGENTS.md | markdown-агенты |
| **L2** | скиллы, субагенты | app-server hooks | permissions ask/allow |
| **L3** | hooks .claude/settings.json | sandbox CLI | запрет инструмента + sandbox |

# ЧЕТЫРЕ КЛАССА ХУКОВ

## DELETE GUARD

блокирует rm, drop, terraform destroy

## ХУК КОНФИГОВ И КЛЮЧЕЙ

блокирует правки .env, ключей, prod-конфигов

## ПРОВЕРКА УТВЕРЖДЕНИЕ — ДОКАЗАТЕЛЬСТВО

блокирует «готово» без подтверждения

## STOP-ХУК НА ОГОВОРКИ

блокирует «похоже», «known limitation»

# эпизодов ухода от ответственности проявились за 17 дней

**173**

AMD, 6 852 сессии, директор AI-группы

# Хук не лечит модель. Делает искажение видимым

# Пять шагов на папке pm-agent-staging

# Что соберём за 25 минут

| ШАГ | ЧТО ДОБАВИМ |
|---|---|
| 1 | Хуки безопасности + dangerously-skip-permissions |
| 2 | Долгосрочная память — снэпшот и подгрузка |
| 3 | MCP + поиск по инструментам |
| 4 | Хук на утверждение без доказательства |
| 5 | Stop-хук на оговорки + субагент-рецензент |

LIVE DEMO

ШАГ 1

# Хуки безопасности + dangerously-skip-permissions

PRODUCT SENSE

# Хуки на L3. Опасные команды блокируются. Остальное идёт без подтверждений

ШАГ 2

# Долгосрочная память + хуки жизненного цикла

# ПАМЯТЬ? ТОЖЕ ХУКИ


Модель ничего не помнит. Помнят хуки

ШАГ 3

# MCP + поиск по инструментам

# MCP открывает инструменты. Поиск по инструментам держит контекст чистым

ШАГ 4

# Хук на утверждение без доказательства

# 200 — это не «работает». Зелёные тесты — это не «путь работает»

ШАГ 5

# Stop-хук на оговорки + субагент-рецензент

# Stop-хук на оговорки плюс субагент-рецензент — вторая пара глаз без человека

# ПЯТЬ ШАГОВ СБОРКИ

| | | | |
|---|---|---|---|
| 01 |  | **ХУКИ БЕЗОПАСНОСТИ** | ЦЕНА: поддерживать список<br>НЕ ЗАКРЫВАЕТ: невидимый обход |
| 02 |  | **ПАМЯТЬ + ХУКИ ЖИЗНЕННОГО ЦИКЛА** | ЦЕНА: что записывать, как стареть<br>НЕ ЗАКРЫВАЕТ: устаревшие данные |
| 03 |  | **MCP + ПОИСК ПО ИНСТРУМЕНТАМ** | ЦЕНА: настройка<br>НЕ ЗАКРЫВАЕТ: подделанные источники |
| 04 |  | **ПРОВЕРКА УТВЕРЖДЕНИЕ — ДОКАЗАТЕЛЬСТВО** | ЦЕНА: список маркеров<br>НЕ ЗАКРЫВАЕТ: подхалимство в промпте |
| 05 |  | **STOP-ХУК + СУБАГЕНТ-РЕЦЕНЗЕНТ** | ЦЕНА: до 15× токенов<br>НЕ ЗАКРЫВАЕТ: ложная уверенность в ширину |

# Какой шаг сборки вашей задаче нужен первым?

# Что лежит на каждом уровне после сборки

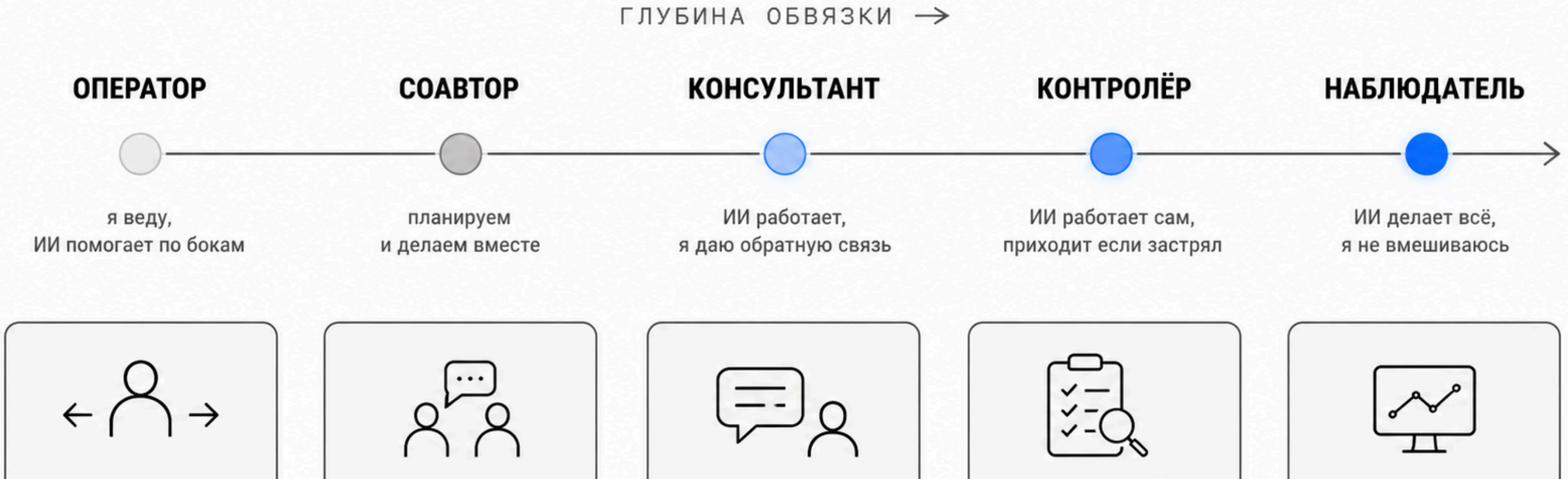
| УРОВЕНЬ | ЧТО ЛЕЖИТ |
|---|---|
| **L1** | CLAUDE.md, правила |
| **L2** | 8 скиллов из M5 + субагент-рецензент |
| **L3** | 4 класса хуков, хуки жизненного цикла памяти, ограниченный MCP, граница рабочей папки |

# Обвязка — костыли. Не гарантия. Главное умение — распределить по уровням

# Соответствие задаче

# Чем больше отдаёшь ИИ — тем глубже обвязка

# ШКАЛА АВТОНОМИИ — ПЯТЬ РОЛЕЙ




### ОПЕРАТОР

Вы задаёте цель и принимаете решения, ИИ помогает выполнять задачи.

### СОАВТОР

Вы и ИИ вместе планируете, обсуждаете и выполняете задачи на равных.

### КОНСУЛЬТАНТ

ИИ выполняет работу, вы даёте обратную связь и корректируете направление.

### КОНТРОЛЁР

ИИ работает автономно, вы подключаетесь только когда возникают проблемы.

### НАБЛЮДАТЕЛЬ

ИИ полностью ведёт процесс, вы наблюдаете за результатами и при необходимости меняете цели.

# ШЕСТЬ КЛАССОВ ЗАДАЧ

## РАБОЧИЙ ПРОЦЕСС

собрать статус, отправить шаблон

## ПОМОЩЬ С ИНСТРУМЕНТОМ

найти прошлое решение, черновик письма

## ОГРАНИЧЕННЫЙ АГЕНТ + КОНТРОЛЬНЫЕ ТОЧКИ

разобрать 20 писем, не отправлять без подтверждения

## ДЛИТЕЛЬНАЯ РАБОТА С ПАМЯТЬЮ

вести бриф проекта 2 недели

## НЕСКОЛЬКО АГЕНТОВ В ШИРИНУ

конкурентный обзор по 5 направлениям

## ОТКРЫТАЯ ФРОНТИРНАЯ ЗАДАЧА

проектный артефакт через неизвестные источники

# Главное умение — выбрать точку в (модель × обвязка × задача)

# Простой класс — минимальная обвязка. Длинная — полная сборка

# Полная сборка не равна нулю риску. Три риска остаются всегда

# Три риска

PRODUCT SENSE

## ТРИ ОСТАТОЧНЫХ РИСКА

1

### Подмена формулировки

навязали гипотезу — план под неё. Подхалимство хуками не закрывается

2

### Неочевидное в файле

нужный факт лежит в файле, который агент не открыл

3

### Скрытое стейл-состояние

решение из decisions.md недельной давности подаётся как свежее

# Какому классу задачи доверите этот конфиг?

# Паспорт автономии

# ПАСПОРТ АВТОНОМИИ

**ЗАДАЧА** _____

**КОНФИГУРАЦИЯ ОБВЯЗКИ** _____

**ПРАВИЛА ПО L1 / L2 / L3** _____

**ЧЕТЫРЕ КЛАССА ХУКОВ** _____

**ПОЛИТИКА ОСТАНОВКИ** _____

**ОСТАТОЧНЫЕ РИСКИ** _____

ОДНА СТРАНИЦА — ОДНА ЗАДАЧА

## КАК ЗАПОЛНЯТЬ ПАСПОРТ

1

### Одна страница — одна задача

не пытаться универсальный шаблон под всё

2

### Правила сначала, риски в конце

L1 / L2 / L3 + хуки + политика остановки

3

### Остаточные риски обязательны

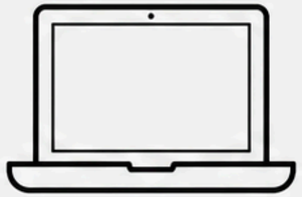
три, которые ваша обвязка не закрывает

# Задача: разобрать 20 писем. Роль: контролёр. Хуки: claim-gate, ownership-stop. Риск: подхалимство

# Индустриальная карта

PRODUCT SENSE

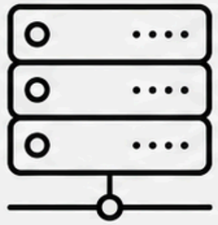
# ТРИ КЛАССА АРХИТЕКТУРЫ



## HARNESS

рамки сессии PM

Claude Code
Codex
OpenCode



## SELF-HOSTED AGENT GATEWAY

собственная архитектура, через API, 24/7

Hermes
OpenClaw

## CLOUD PACKAGING

обвязка у вендора

Anthropic Managed Agents

ТРИ КЛАССА АРХИТЕКТУРЫ. НЕ ОДИН ПОВЕРХ ДРУГОГО

## НА ДОМ

1

### Архив папки

pm-agent-staging с подготовленными хуками, settings.json пуст — активируете по шагам

2

### Шаблон паспорта

одна страница — одна задача — шесть полей

3

### Сравнительная таблица хуков

Claude Code / Codex / OpenCode плюс примеры из opus-agent

# M6 опирается на M1–M5

| МОДУЛЬ | ЧТО ПЕРЕХОДИТ В M6 |
|---|---|
| **M1** | Контракт → политика остановки |
| **M2** | Искажения → причины поломки |
| **M3** | Контекст-пакет → основа памяти |
| **M4** | Рабочий процесс / агент → шкала автономии |
| **M5** | pm-agent-staging → стартовая папка |

# Формула × × ×. Распределение по L1 / L2 / L3. Паспорт под одну задачу

# Вопросы