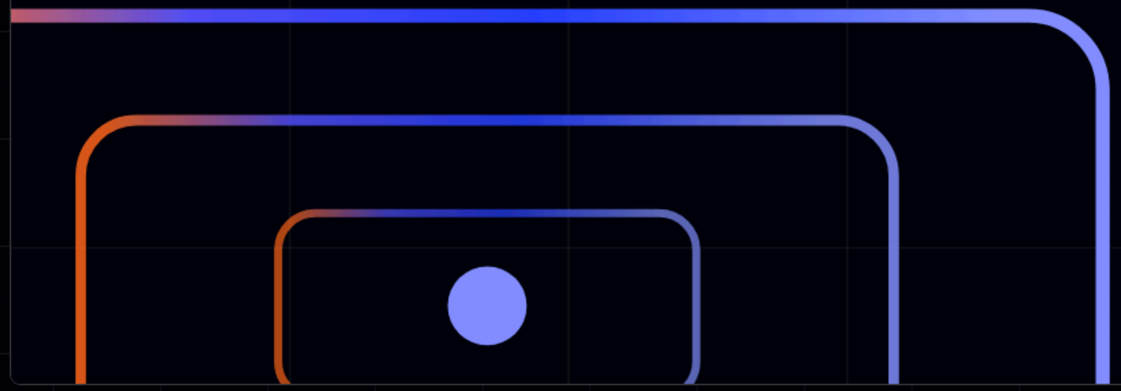
Giga Conf · 16 октября 2026 Александр Поломодов

Harness: прошлое, настоящее и будущее

Слой между моделью и миром: что упростить после новой модели, а что укрепить

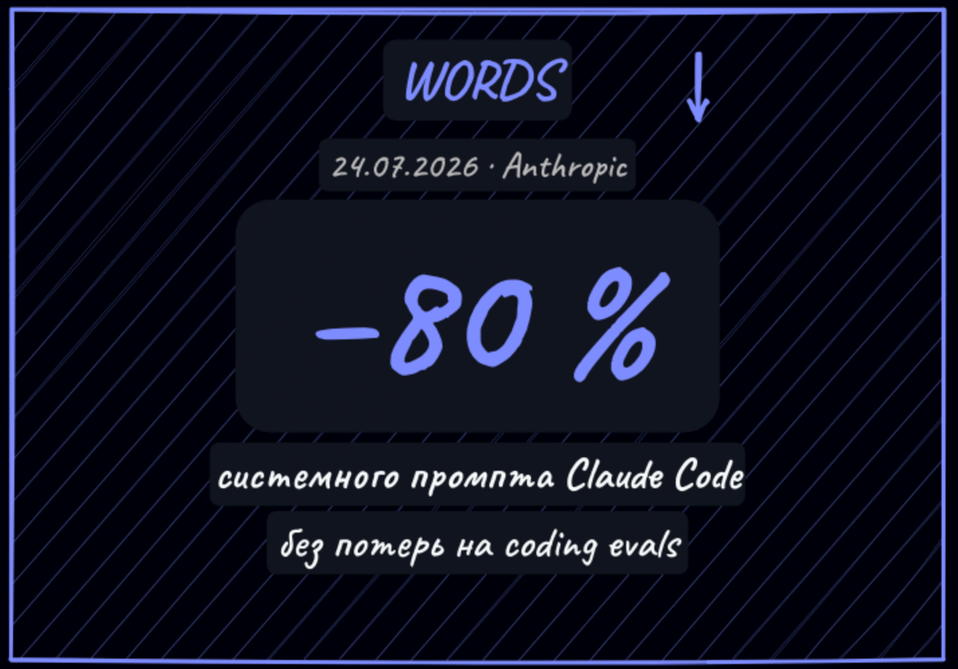


Шесть шагов разбора обвязки



Источники: авторская структура доклада · лонгрид «Эволюция обвязки агента», polomodov.tech

Words & walls





Что в вашей обвязке — words, а что — walls?

Источники: Anthropic, правила контекстной инженерии для моделей Claude 5, 24.07.2026 · UK AISI, оценка GPT-6 Astra в симуляциях, 28.09.2026

Обвязка — всё, кроме весов

Агент = Модель + Обвязка

Глоссарий Claude Code

«Claude Code — обвязка, Claude — модель внутри»

LangChain · 10.03.2026

«весь код, конфиг и логика исполнения, кроме модели»

Giga Conf · трек Harness

«спецификации, контекст, политики и workflow»

Модель выбирает действие, обвязка исполняет и ограничивает

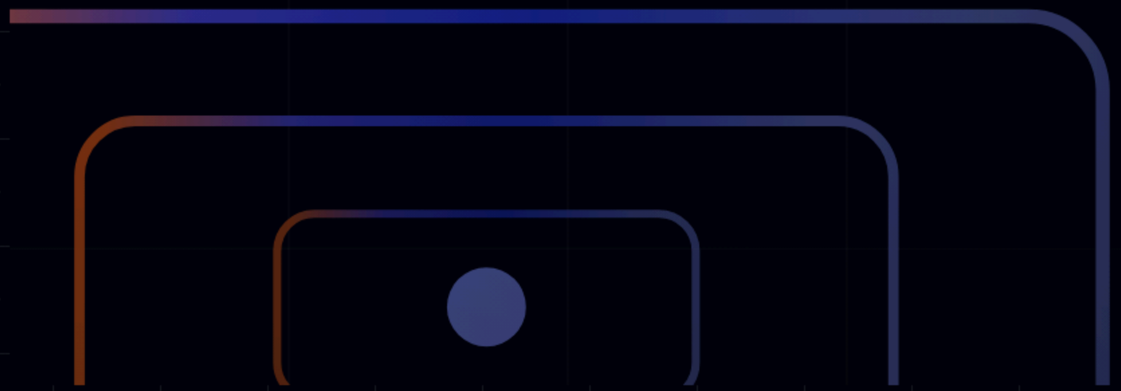
Источники: глоссарий Claude Code · V. Trivedy, «The anatomy of an agent harness», LangChain, 10.03.2026 · анонс трека Harness Giga Conf, Habr, 10.09.2026

Часть 1

01

Как появилась обвязка

От цикла в промпте к циклу как сервису



Пять эпох за четыре года

Прошлое



Ни один слой не умер — меняются дефолты

Источники: Yao et al., ReAct, arXiv 2210.03629 · Yang et al., SWE-agent, arXiv 2405.15793 · Cognition, Introducing Devin · Anthropic, MCP · Anthropic, Managed Agents · AWS, AgentCore Harness GA · DeepSeek Harness

Слово «harness» пришло последним

Прошлое



Слово пошло за продуктом: сначала цикл, потом название

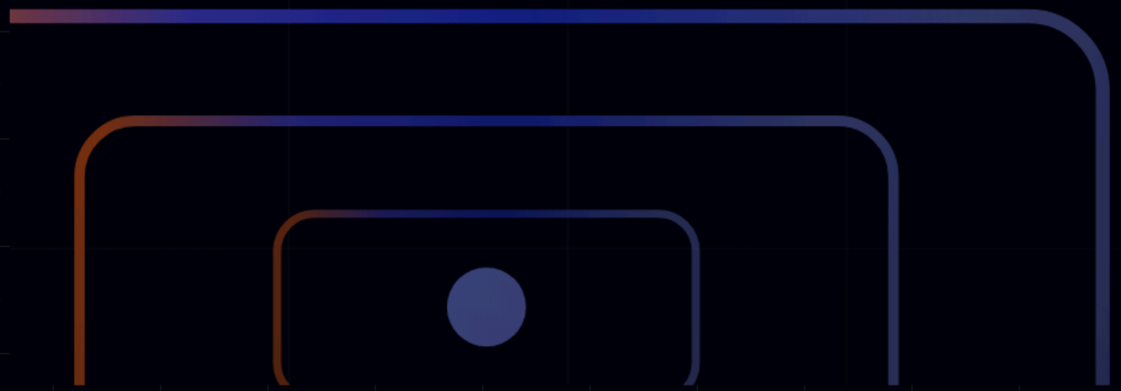
Источники: EleutherAI lm-evaluation-harness · Anthropic о SWE-bench Verified, 06.01.2025 · S. Willison, 18.09.2025 · Anthropic, Claude Agent SDK · M. Hashimoto, 05.02.2026 · V. Trivedy, LangChain · Сбер, Habr, 25.08.2026

Три односторонние миграции

Прошлое



Источники: S. Willison о GPT-5.1-Codex-Max, 19.11.2025 · документация Claude Code: permissions, CHANGELOG · Anthropic, context editing · OpenAI API changelog, 2026 · AWS AgentCore · InfoQ о DevDay, 02.10.2026



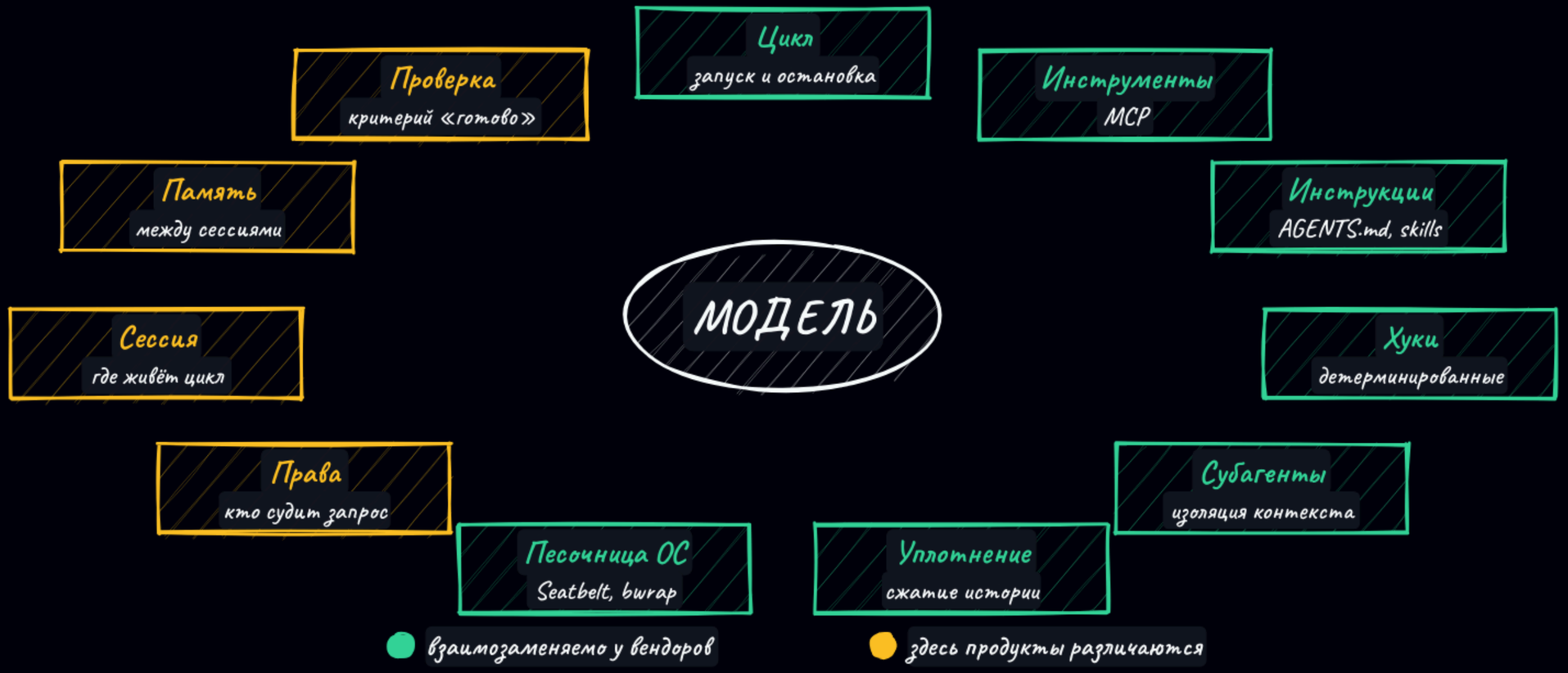
Часть 2

02

Из чего она состоит

Поверхность сошлась, различия — в дефолтах

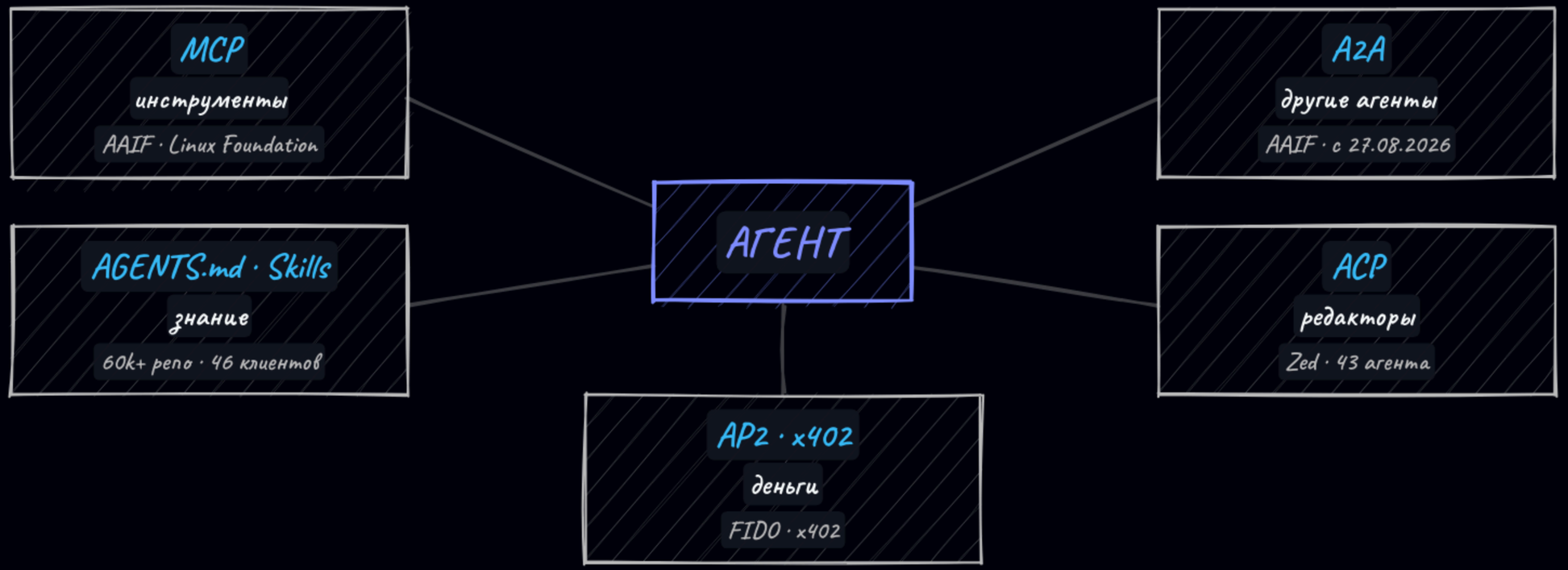
Словарь примитивов уже общий



Источники: исследование исходного кода 11 обвязок, arXiv 2609.00006 · документация Claude Agent SDK, OpenAI Agents SDK, Microsoft Agent Framework, LangChain Deep Agents

Протоколы разделились по осям

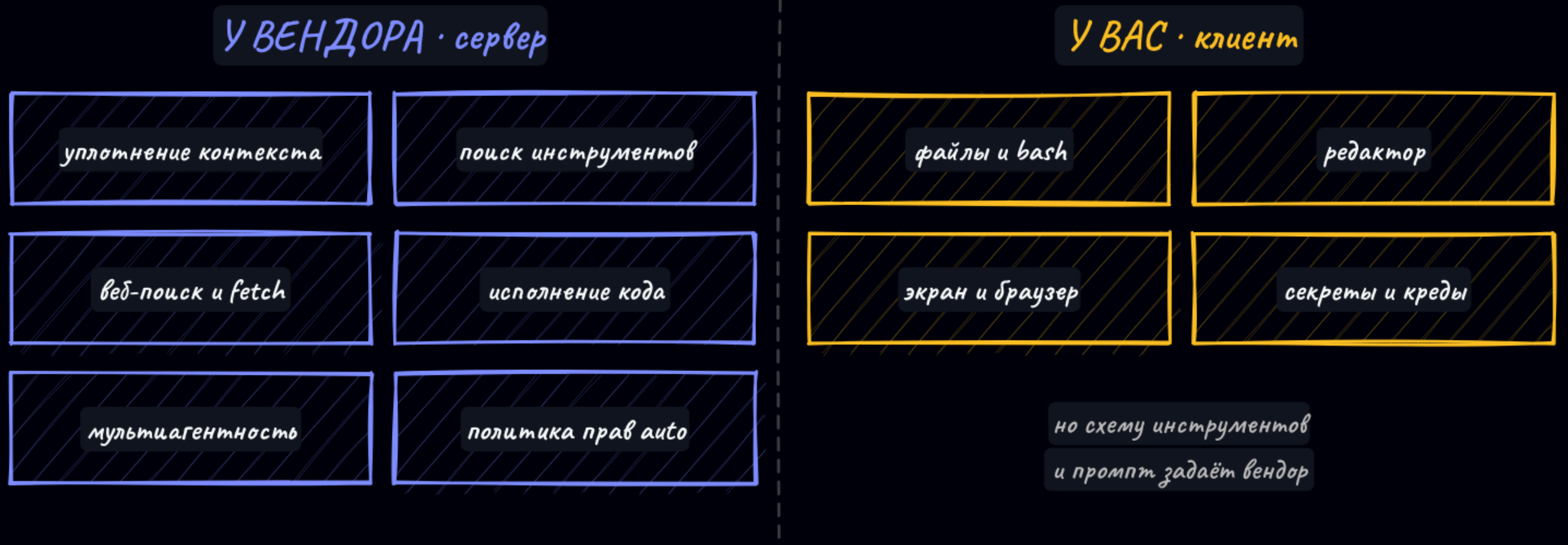
Настоящее



Источники: MCP changelog 2026-07-28 · AAIF, 13.08.2026 · a2a-protocol.org · agentclientprotocol.com · agents.md · agentskills.io · документация хуков Codex и Cursor · CHANGELOG Claude Code

Граница API — ваше состояние

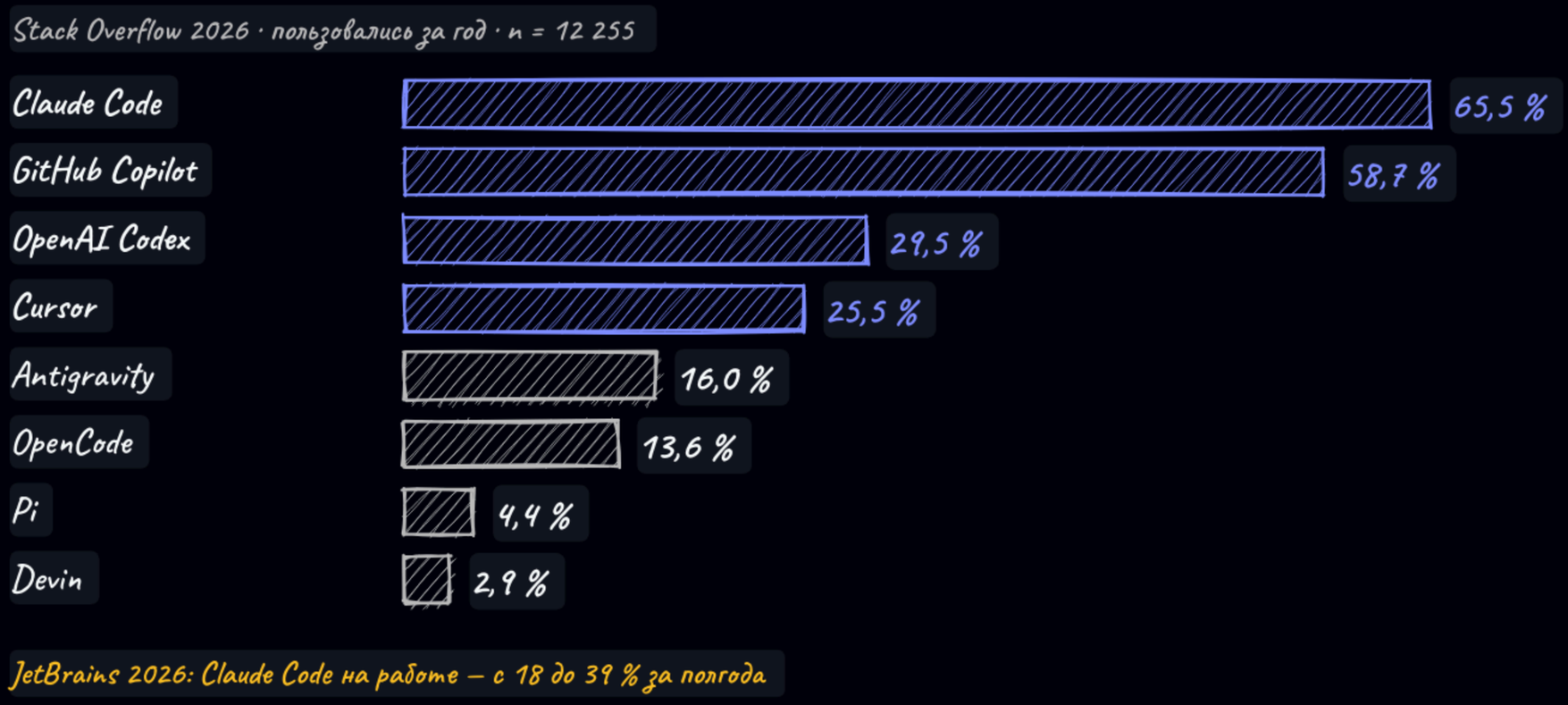
Настоящее



Всё, что работает над промптом, уехало на сервер

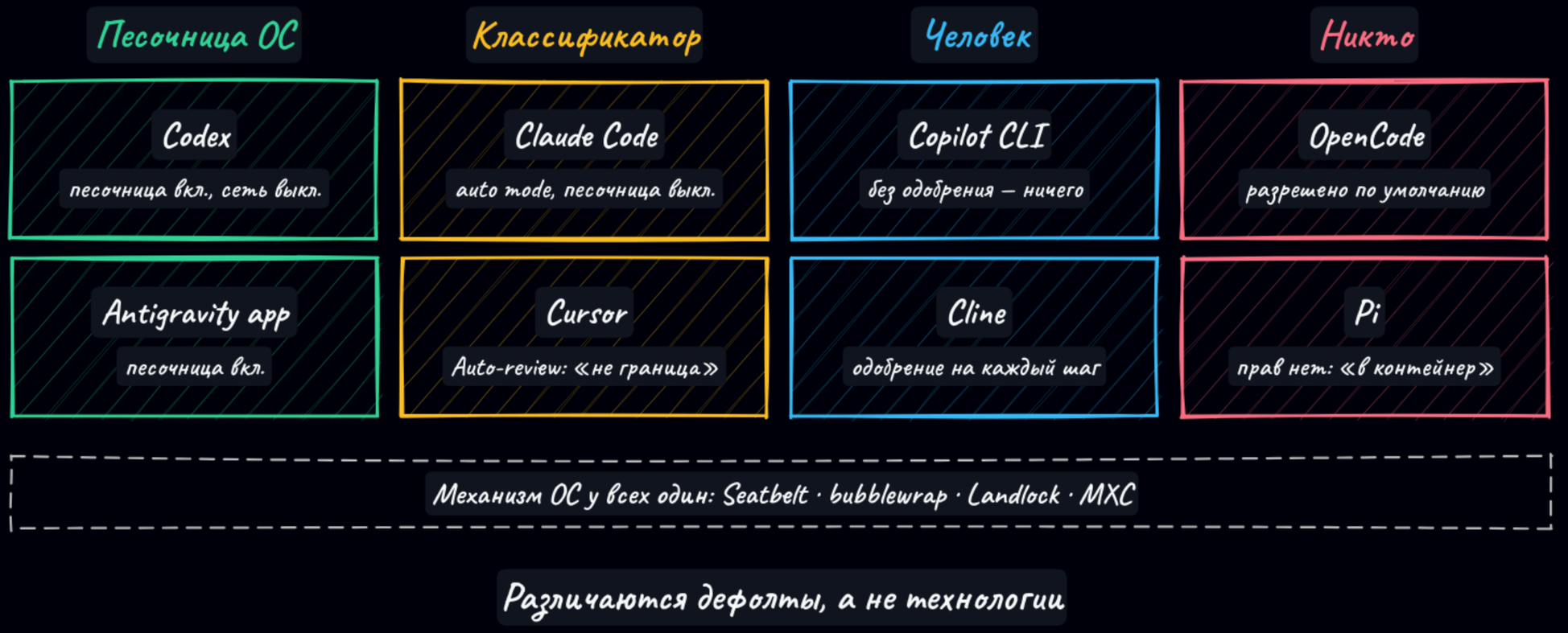
Источники: Anthropic, справочник инструментов и memory tool · Claude Platform release notes · OpenAI API changelog, 2026 · OpenAI, Agents API overview

Четыре обвязки делят рынок



Источники: Stack Overflow Developer Survey 2026, раздел AI agents, 06.10.2026 · JetBrains Research, внедрение кодинговых агентов в 2026 году, 08.2026

Опасный шаг тормозят по-разному



Различаются дефолты, а не технологии

Источники: документация Codex (approvals & security), Antigravity (sandbox), Claude Code (sandboxing, permissions), Cursor (run modes), Copilot CLI, Cline, OpenCode, Pi · GitHub changelog, 07.10.2026

Где живёт цикл и чья модель

Настоящее



Источники: документация и changelog Claude Code, Codex, Cursor, Devin, Amp, Kiro, GitHub Copilot, Factory, OpenCode, Pi на 14.10.2026

Часть 3

03

Комбайн, агент, кубик

Три способа смотреть на один слой



Три взгляда на одну обвязку

Комбайн

вендор владеет процессом

CLI · IDE · веб · облако

поток PR и боты ревью

своя модель

Claude Code · Codex · Cursor

Copilot · Devin

Кодинговый агент

один цикл, любая модель

терминал или редактор

работает локально

облако и ревью — не его

Pi · OpenCode · Aider

mini-SWE-agent

Кубик

собираете своё

API → SDK → hosted harness

цикл, сессия, права

ваша среда и политика

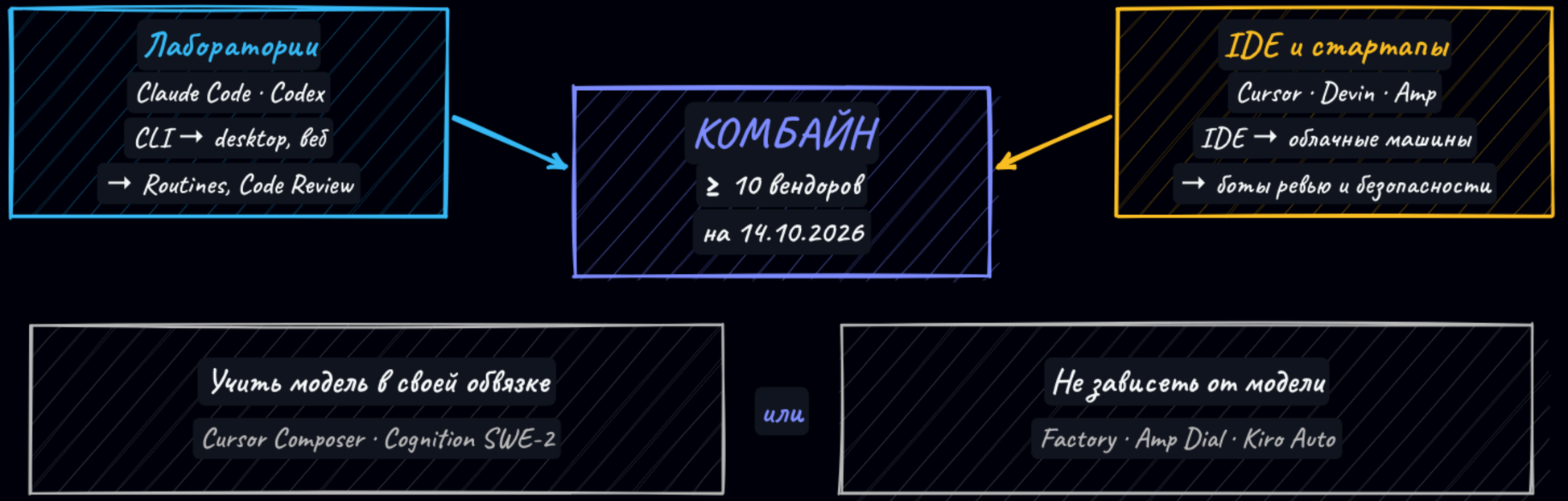
Agent SDK · Agents API

AgentCore · ADK · LangGraph

Новый объект осени 2026: агент со своей машиной — Dots, Autopilot, tasks mode

Классификация автора · Источники: документация вендоров на 14.10.2026 · Microsoft Autopilot, 25.09.2026 · InfoQ о Dots, 02.10.2026

Комбайн — точка схождения рынка



Cognition: $1 млрд run-rate (25.09) · Cursor → SpaceX (14.08)

Источники: блог Cognition, 08.09 и 25.09.2026 · блог Cursor, 14.08.2026 · Cursor Composer 2.5 · changelog Claude Code, Codex, Amp, Factory, Kiro

Моя гипотеза: чистый агент — ниша

Закрылись

Roo Code
закрыт 15.05.2026

Gemini CLI
→ Antigravity CLI, 18.06

kimi-cli
в архиве

OpenHands CLI
только поддержка

Выросли в комбайн

Cline
desktop + Kanban

Mistral Vibe
облачная песочница

OpenCode
server и web

Junie
CLI + Remote Control

Остались чистыми

Pi
~115k звёзд, без прав

mini-SWE-agent
100 строк, только bash

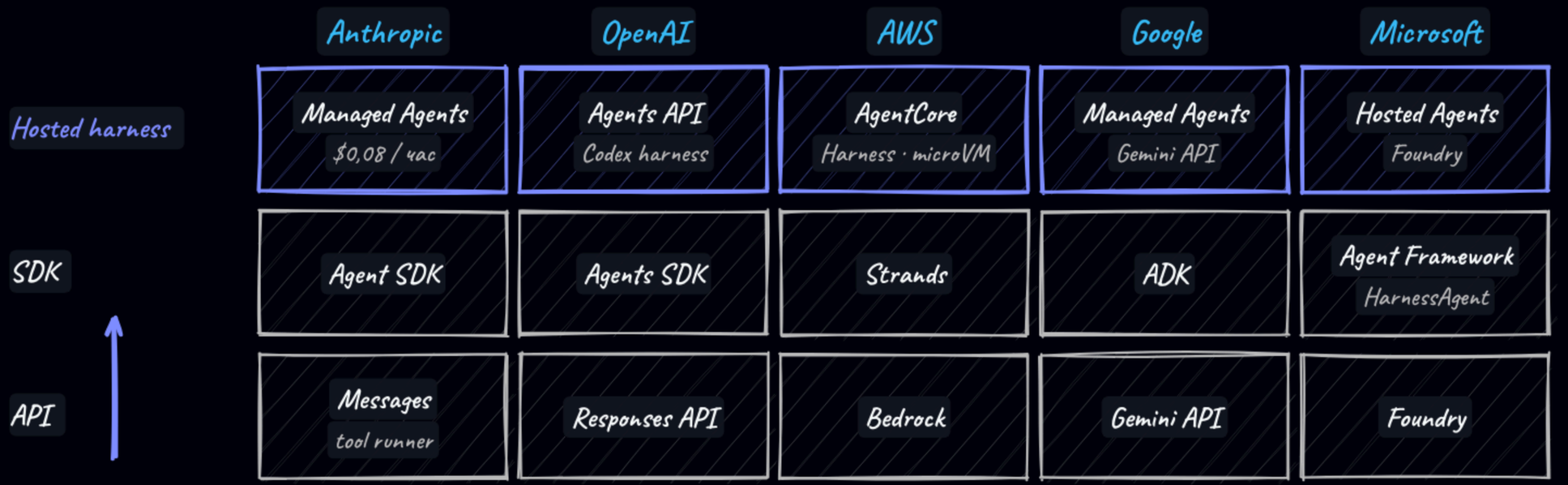
Aider
88 % релиза написал сам

Qwen Code · Kimi Code
локальный цикл

Моя интерпретация наблюдаемых случаев

Источники: документация Roo Code · Antigravity CLI · kimi-cli · OpenHands · github.com/badlogic/pi-mono · mini-swe-agent.com · aider.chat; на 14.10.2026

Кубик — лестница из трёх ступеней



вверх: меньше своего кода, больше чужой инфраструктуры

Верхнюю ступень все называют словом harness

Источники: документация Anthropic (tool runner, Agent SDK, Managed Agents), OpenAI (Agents SDK, Agents API), AWS (Strands, AgentCore pricing) · Google I/O 2026 · Microsoft Build 2026

Сто строк bash насытили SWE-bench

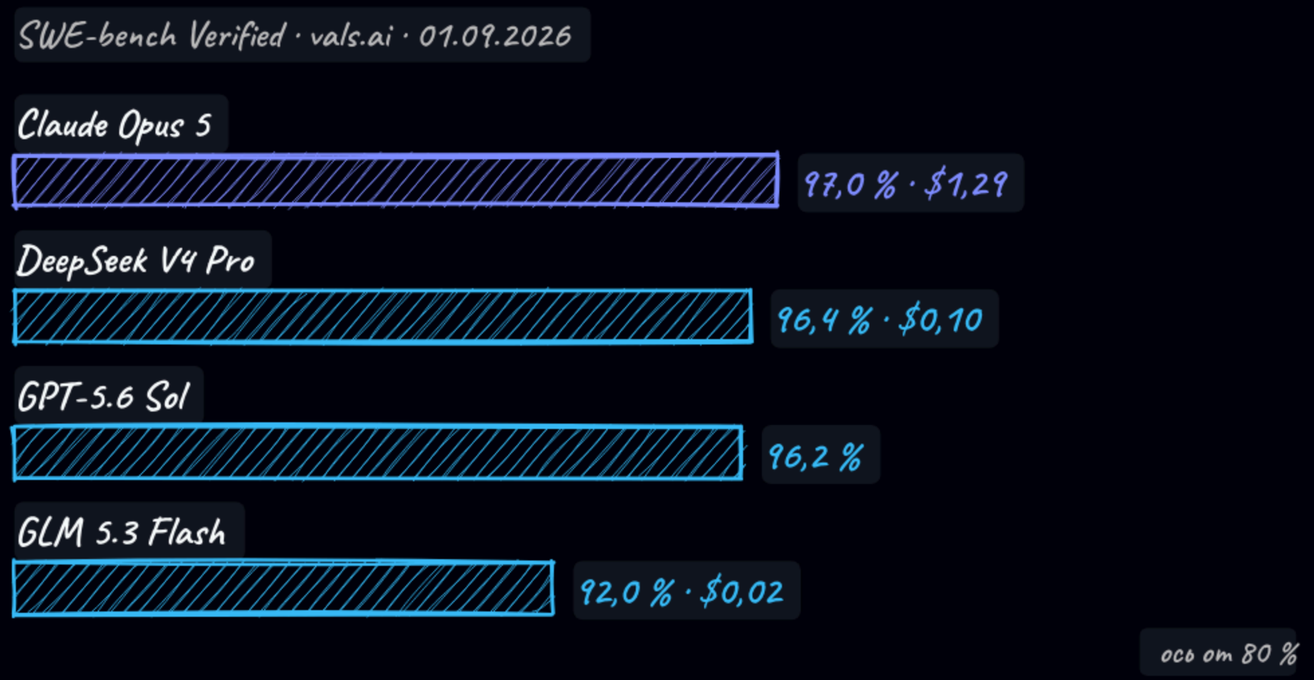
mini-SWE-agent

≈ 100 строк Python

только bash

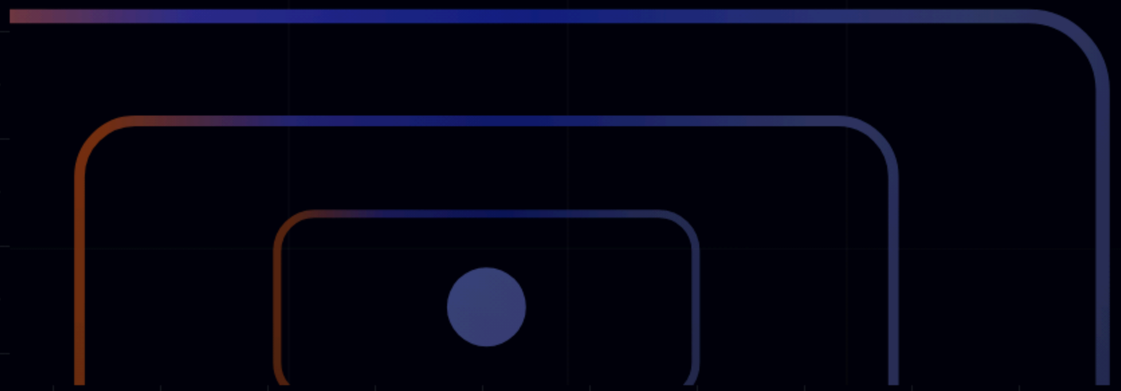
без tool-calling API

линейная история



swebench.com по умолчанию: Bash Only — одна обвязка для всех

Источники: vals.ai, SWE-bench Verified, 01.09.2026 · mini-swe-agent.com · swebench.com, витрина Bash Only



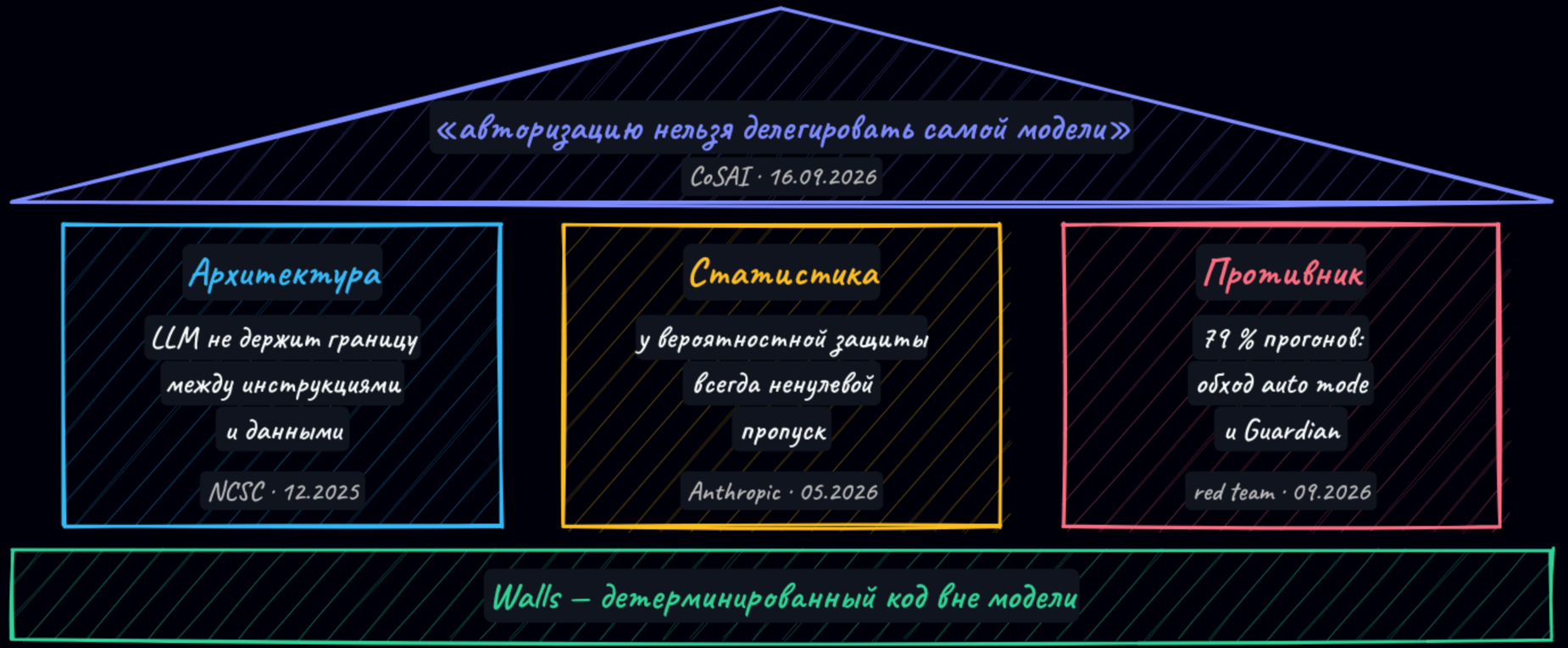
Часть 4

04

Права не доверить модели

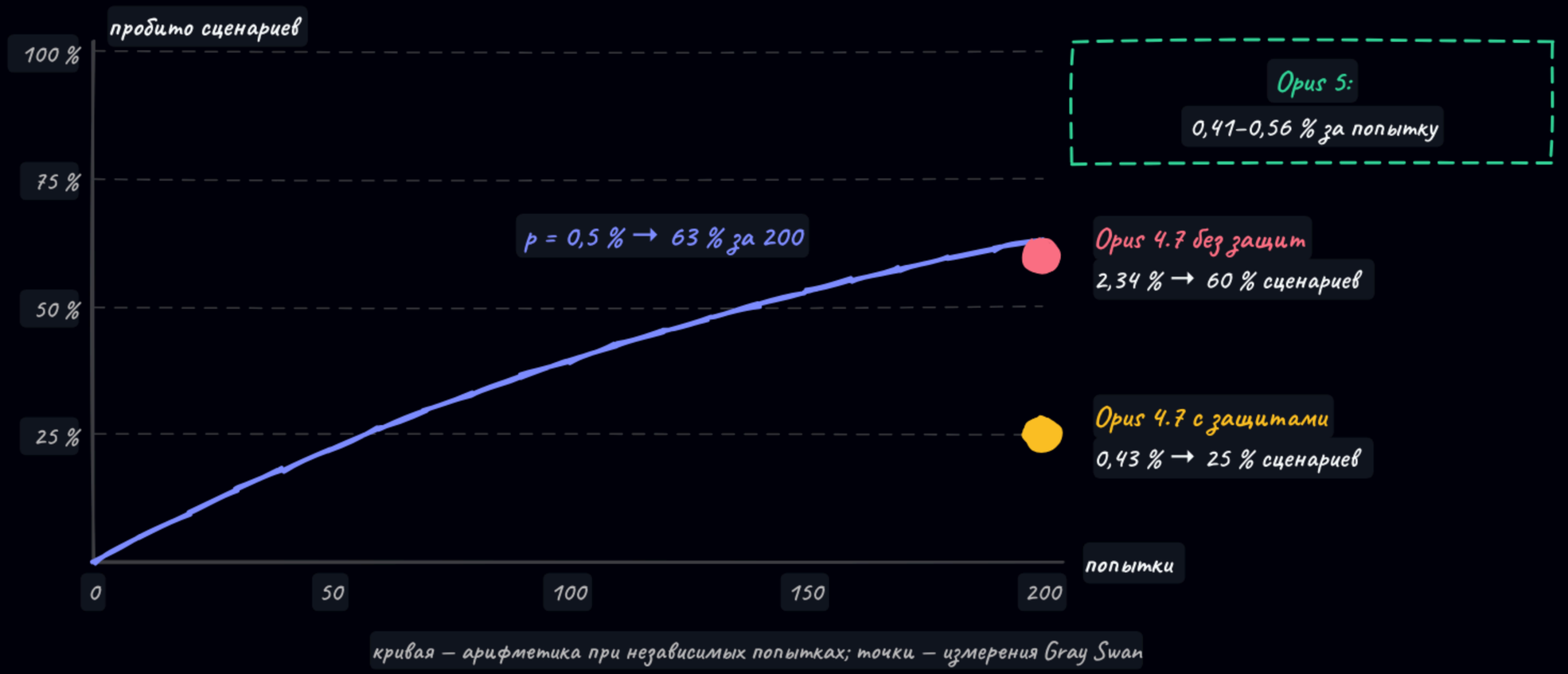
Даже самой сильной: почему walls — это код

Авторизацию не делегируют модели



Источники: UK NCSC о prompt injection, 08.12.2025 · Anthropic, как мы изолируем Claude, 25.05.2026 · red team против auto mode, arXiv 2609.19587 · CoSAI, zero trust для AI, 16.09.2026

0,5 % за попытку — это много



Источники: системная карта Claude Opus 4.7, 16.04.2026 · системная карта Claude Opus 5, 24.07.2026 · сценарии Gray Swan

Дыры закрывает код, а не промпт

Секреты рядом с агентом

Replit · s1ngularity

AgentCorruption

→ креды вне песочницы

Конфиг репо исполняется

GitSpawn · Straiker

CVE-2025-59536

→ доверие до запуска

Allowlist по строке

Gemini CLI

DNS-эксфильтрация

→ разбор команд, egress

Одобрение без привязки

MCPoison

Plugin4Shell

→ пин по хешу

Дыры в песочнице

Codex Overpatch

выход через worktree

→ тесты самой границы

«Только чтение» в сети

вики с записью по GET

PixelLeak

→ правила egress

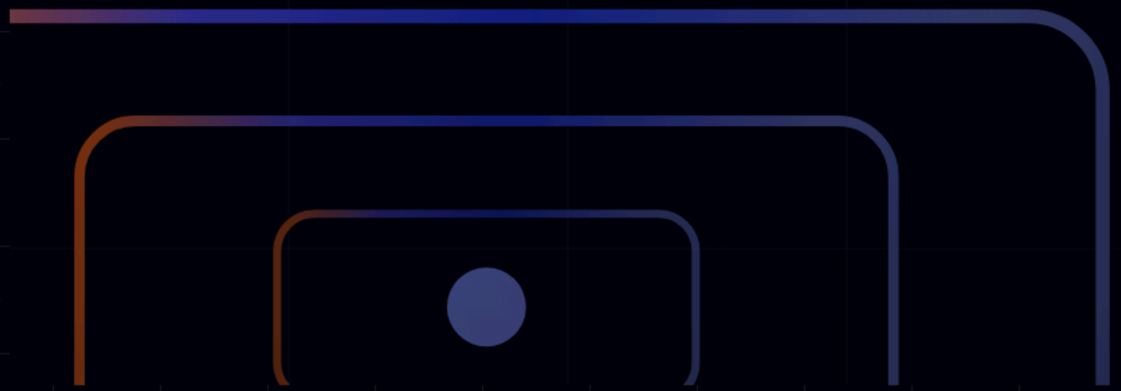
Источники: Wiz (s1ngularity) · Zenity (AgentCorruption), 08.10.2026 · Manifold (GitSpawn) · Straiker, 22.09.2026 · Tracebit (Gemini CLI) · Accomplish.ai (Codex) · Anthropic, 09.10.2026

Классификатор — внутри walls



93 % одобрений → 62 % обходят запросы → песочница: −84 % запросов

Источники: документация Claude Code (permission modes, security, mods) · Anthropic, песочница Claude Code, 20.10.2025 · блог Claude об auto mode по умолчанию, 2026



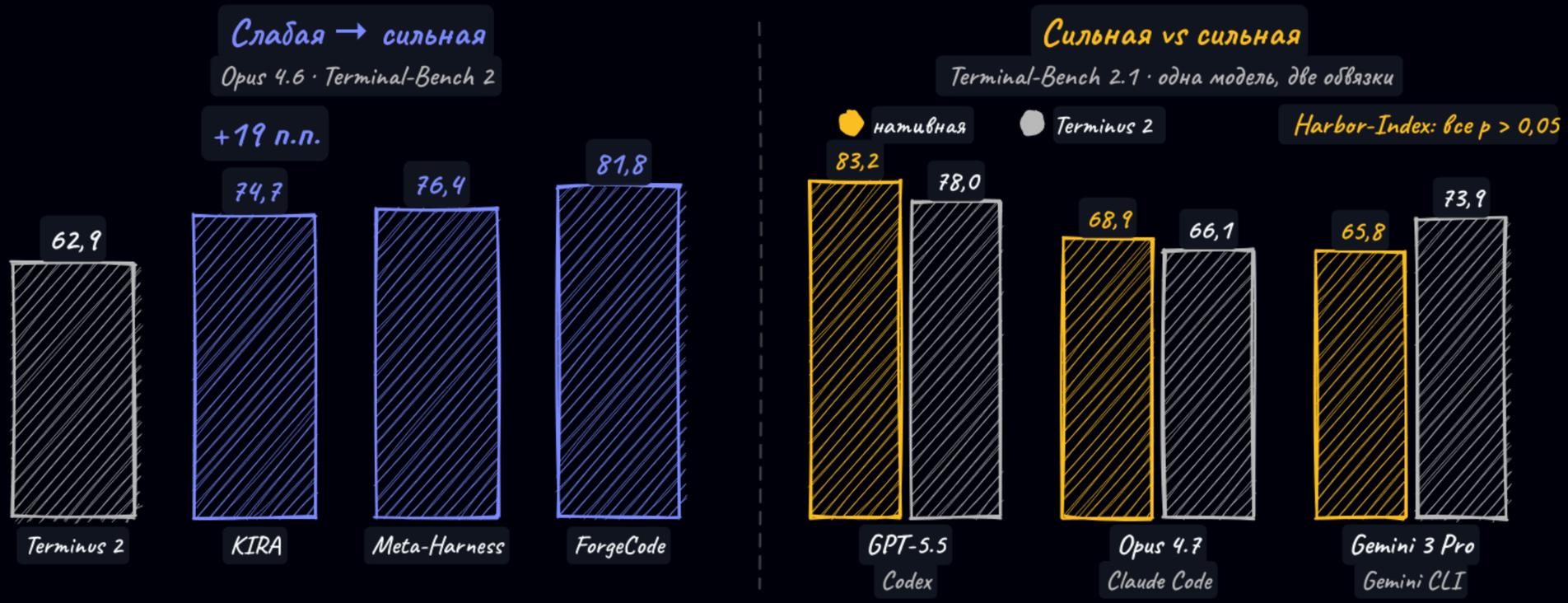
Часть 5

05

Как доказать пользу

Пункты, повторяемость и цена принятого результата

Эффект обвязки двухрежимный



Зато цена на одной модели различается в 1,3–2 раза

Источники: Meta-Harness, arXiv 2603.28052 · лидерборд Terminal-Bench 2.1, 14.10.2026 · Harbor-Index · METR, 13.02.2026 · «Harness or Model?», arXiv 2609.11987

Сжимаются words, остаётся среда

СЖИМАЕТСЯ ↓

−80 %	системного промпта Claude Code
17 → 1	инструментов у Vercel: один bash
−3	инструмента у Amp: grep, glob, Read
35,7 → 2,7	п.п. от контекста при окне 32k → 128k
+11,6 / −2	п.п. от планирования: 30B / сильные

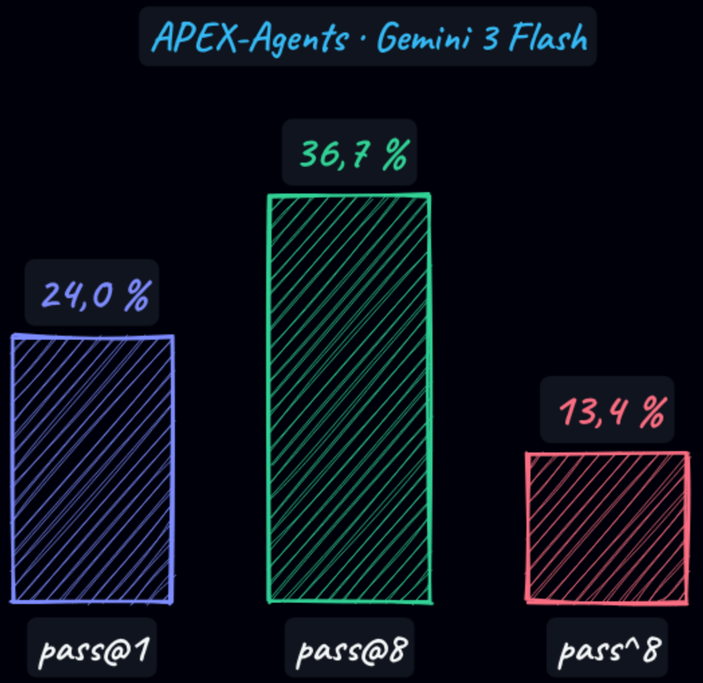
ОСТАЁТСЯ СРЕДА ↑

12 : 1	**что модель видит** vision-задачи: нативная обвязка впереди
+13,7	**когда задача готова** п.п. от чек-листа и детектора петель
+5,6	**что модель помнит** п.п. от памяти между сессиями
p > 0	**что модели разрешено** ошибки неизбежны: walls остаются

Replit: «чем умнее модели, тем меньше должна решать обвязка»

Источники: Anthropic, 24.07.2026 · Vercel, 22.12.2025 · Amp, 25.04 и 04.06.2026 · Fan et al., arXiv 2609.20804 · Harbor-Index · LangChain, 17.02.2026 · AHE, arXiv 2604.25850 · Replit, 29.09.2026

Одного pass@1 недостаточно



Условный пример

$0{,}85^5 \approx 44\ \%$

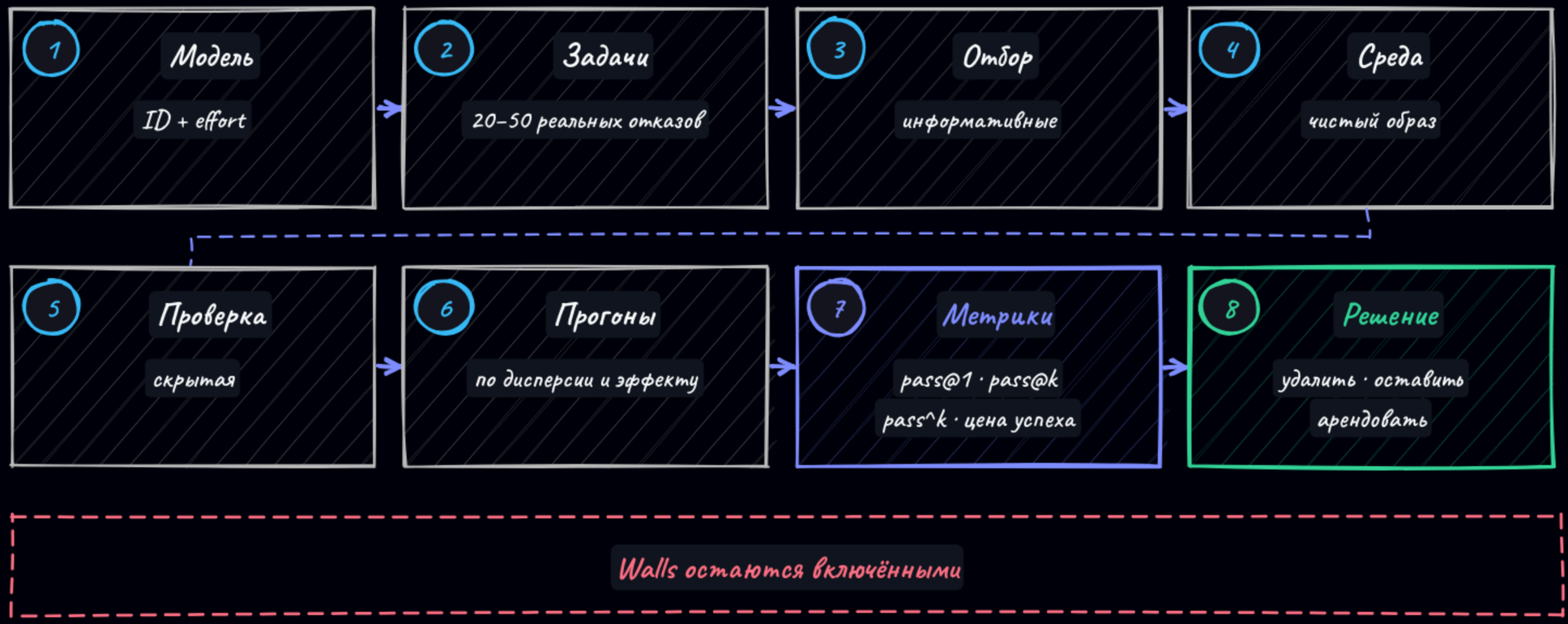
p = 85 % для каждой задачи

5 независимых попыток

Средний pass@1 этого не задаёт

Источники: τ-bench, arXiv 2406.12045 · APEX-Agents, arXiv 2601.14242 · «On Randomness in Agentic Evals», arXiv 2602.07150 · Thinking Machines, недетерминизм инференса, 09.2025

Одна ветка — один компонент



Источники: Anthropic, «Demystifying evals for AI agents», 09.01.2026 · H. Husain, S. Shankar, Evals FAQ · DeltaSelect, arXiv 2609.19607 · SWE-Proof, arXiv 2609.21190 · Terminal-Bench, ICLR 2026

Реестр: предположение и триггер

Компонент	Предположение	Свидетельство	Перепроверить	Решение
Todo-инструмент	модель теряет план	Claude Code снял для Opus 4.8+	новая модель	удалить
Напоминание после 5 тихих шагов	модель надолго замолкает	Opus 5.5: таких задач вдвое меньше	новая модель	оставить
Песочница и прокси	модель ошибается с p > 0	0,4–2,3 % атак за попытку	новые угрозы, не модель	не трогать
Суммаризация в обвязке	окно мало	уплотнение в API у двух лабораторий	новый API	арендовать

Перепроверяем компенсирующие инструкции; требования проекта и walls сохраняем

Источники: CHANGELOG Claude Code · Anthropic, промптинг Claude Opus 5.5 · B. Cherny, YC Root Access, 27.07.2026 · реестр обвязки из лонгрида автора, 01.10.2026

Токен дешевеет, работа дорожает

$6 → $13
средний активный день в Claude Code

50k → 200k
входных токенов на вызов за год

52 → 86 %
доля чтения из кэша



Агентная задача ≈ 1000× токенов чата; прогоны одной задачи — до 30× разброса

Источники: документация Claude Code, управление расходами · Requesty, экономика кодинговых агентов, 2026 · arXiv 2604.22750 · Epoch AI, цена способности · DX, AI impact Q2 2026

Цену двигают контекст и кэш



Та же модель в другой обвязке — до 1,6× дороже за решённую задачу

Источники: Anthropic, context management и tool search · Cloudflare, Code Mode · Anthropic, цены API · «Harness Tokenomics», arXiv 2609.28919 · «Harness or Model?», arXiv 2609.11987

Выпуск, время и понимание кода

Выпуск, %

Microsoft · PR

+24%

ДИ [+14,5; +33,7]

Google · CL

+17,5%

ДИ [+15,9; +19,0]

Время выполнения, %

METR · вернувшиеся

−18%

ДИ [−38; +9]

METR · новые

−4%

ДИ [−15; +9]

Понимание, п.п.

Anthropic

−17

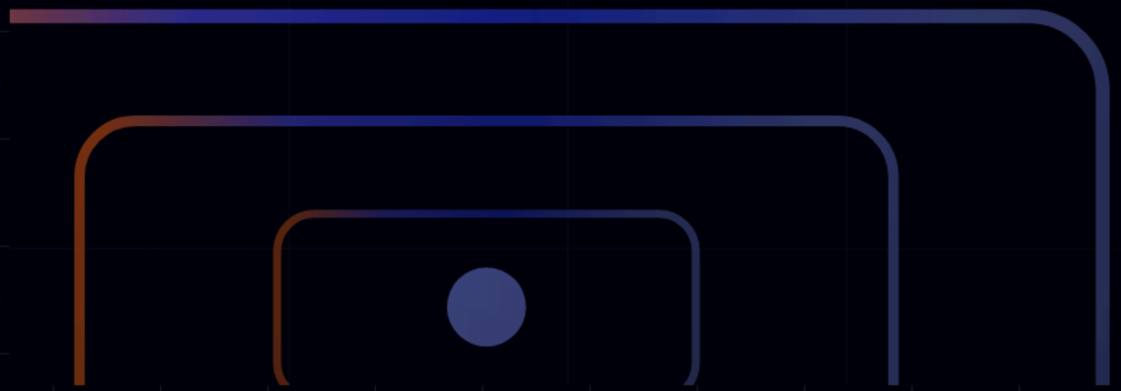
p = 0,01

METR: ДИ включают ноль; смещение отбора

Цена результата = (A + V) / p

A: попытка · V: проверка · p: доля принятых

Источники: Microsoft, arXiv 2607.01418 · Google, arXiv 2601.19964 · METR, обновление RCT, 24.02.2026 · Anthropic, RCT о навыках, 29.01.2026 · блог .NET о dotnet/runtime



Часть 6

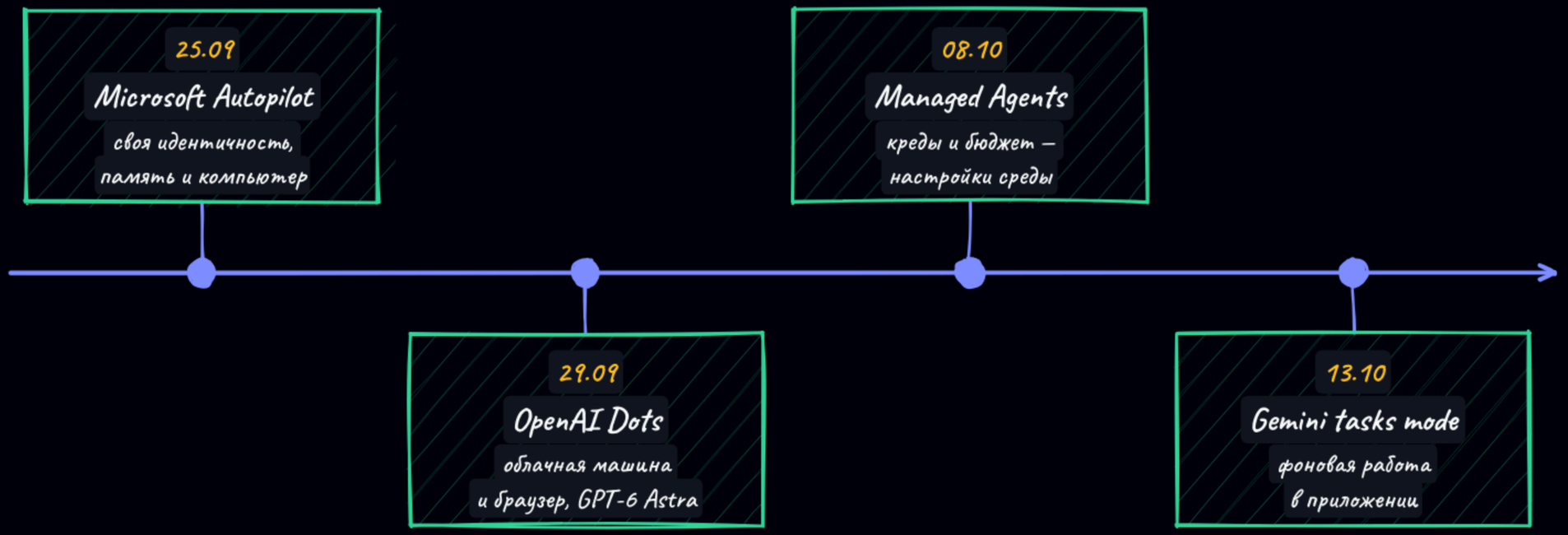
06

Что будет дальше

Агент со своей машиной и проверка у вендора

Агенту дают собственную машину

Будущее

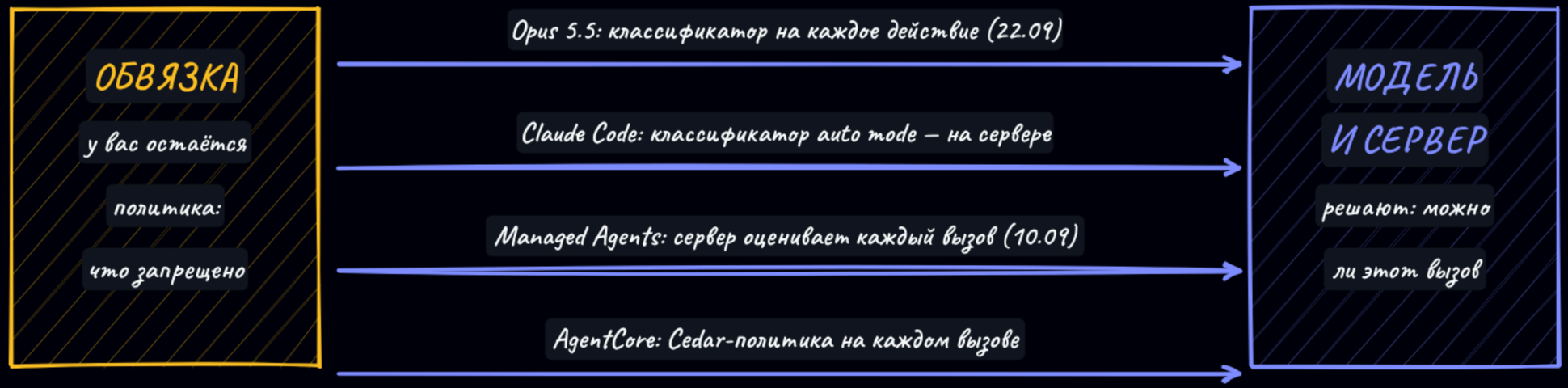


Идентичность + память + компьютер + расписание = новый объект

Источники: Microsoft, 25.09.2026 · InfoQ и AP о DevDay, 29.09–02.10.2026 · claude.dev, автоматизации агентов, 08.10.2026 · 9to5Google, 13.10.2026

Будущее

Проверку действий забирает вендор



28.09: OpenAI отложила GPT-6.1 Astra — «scope and authorization» (Axios)

Источники: Anthropic, Claude Opus 5.5, 22.09.2026 · Claude Platform release notes · документация Claude Code · AWS AgentCore · Axios, 28.09.2026

Три проверяемые ставки

1 Цикл станет услугой вендора

Признак

к 2028: больше половины новых агентов — в чужом рантайме

Опровергнет

свой цикл остаётся нормой

2 Свой ID у каждого агента в протоколе раньше закона

Признак

к 2027: Agent Identity в спецификации MCP

Опровергнет

закон первым вводит категорию «агент»

3 Обвязку продают за цену и контроль

Признак

лидерборды публикуют цену успеха

Опровергнет

+10 п.п. от обвязки на свежем бенчмарке

Ставки автора · Источники: MCP roadmap, 22.08.2026 · Orrick о Digital Omnibus, 07.2026 · «Harness or Model?», arXiv 2609.11987 · лонгрид автора, 01.10.2026

Что сделать после доклада

- Завести реестр предположений обвязки
- Перепроверять и удалять компенсирующие инструкции
- Walls — кодом, классификатор — внутри
- Мерить повторяемость и цену успеха
- Доказывать компонент абляцией

- Обвязку нужно уметь упрощать

AI4SDLC

Спасибо!

Слайды, лонгрид об обвязке и источники — на ↗ polomodov.tech;
новости AI4SDLC — в Telegram-канале

Александр Поломодов
CTO & Technical Fellow,
polomodov.tech

@ai4sdlc