



От больших батальонов к **латентной коммуникации**

Как LLM-системы прошли путь от гонки параметров к агентам, которые договариваются без слов

Александр Диденко · ТюмГУ · МШУ «СКОЛКОВО»

Ридли Скотт, «Наполеон», 2023; Сергей Бондарчук, «Война и мир», 1965–1967; Дени Вильнёв, «Прибытие», 2016

Кто я

В 2001 году, поступив на работу в инвестиционную компанию, занялся автоматизацией принятия решений на финансовых рынках; в 2004 написал диссертацию о поведенческой экономике и генетических алгоритмах на финансовых рынках
В 2023 году стал одним из первых пользователей GPT 3.5. С начала 2024 года – эксперименты с внедрением ИИ в образование:

SAS, Школа права и «Большой» ТюмГУ

- 4000+ пользователей, 40+ ИИ-ассистентов
- Сквозной эксперимент с end-to-end «ИИзацией» образовательного процесса на факультете, альтернативные чатботам интерфейсы к ИИ

МШУ «СКОЛКОВО»

- ИИ-ассистенты для AI Shift и MBA
- Совместные исследования с Центром креативных индустрий и Центром ESG

Исследовательский фокус

- 01** Культура больших языковых моделей (совместно с ИНП А. А. Аузана) и [макро]экономические эффекты ИИ
- 02** Провалы внедрения ИИ и в целом — рассмотрение ИИ в теоретической рамке когнитивной теории фирмы
- 03** Сикофантия и сократический режим LLM, снижение когнитивных способностей в паре «оператор — LLM»
- 04** Генеративная эпистемология: использование ИИ для усиления исследователей (и изобретателей), новые способы получения убедительных свидетельств



«ИИ вместо
профессора»



«От пилота к
масштабу...»



«Культурный код
ИИ»



«GYROval»

Input

Research objective:
Identify validated Type 2
diabetes protective mechanisms



Dataset:
Annotated multi-omics data

Kosmos World Model



Output

The T2D protective variant rs9379084-A tags a candidate cis-acting regulatory sequence that modulates levels of SSR1 via altered ATF3 binding

Summary
The missense variant rs9379084-A at the RREB1 locus emerges as a regulatory hotspot for type 2 diabetes protection, with convergent multi-omic evidence indicating that the protective allele modulates transcription factor occupancy, chromatin accessibility, and downstream transcript levels in a cell-type-specific manner. Among the six genes influenced, SSR1 is a leading candidate target, supported by increased ATF3 motif affinity, idiosyncratic ATAC-QTL effects, and ranking within the top downregulating protective mechanisms.

Background
Human genetic studies of type 2 diabetes increasingly point to noncoding and regulatory mechanisms that act through transcription factor binding, chromatin accessibility, and context-specific gene expression to reveal risk and metabolic issues. Integrative frameworks that harmonize allele directionality, evaluate directional concordance between genome-wide association signals and quantitative trait loci across molecular layers, and prioritize mechanisms with multi-source support are needed to move from association to function. In this context, regulatory hotspots in which single variants orchestrate multi-gene programs have particular importance for identifying protective pathways and therapeutic entry points.

Results & Discussion
This work establishes rs9379084-A as a high-confidence protective variant and a coordinated regulatory driver at the RREB1 locus. Harmonization of allele directionality across datasets produced an analysis-ready resource (100,186 SNP+TF pairs, 1,276 SNPs, 306 TFs) with 100% consistency in risk/protective allele direction, enabling robust assessment of concordance between GWAS and QTL effects. Notably, protective allele TF binding effects were nearly split between gain and loss of function, minimizing systematic bias in mechanism inference [1]. At the locus level, rs9379084 shows strong protective association ($Z = -2.08$, $P = 9.31 \times 10^{-3}$) with posterior prob-

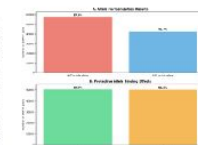


Figure 1: Allele harmonization establishes risk allele directionality and reveals harmonized protective allele acting effects. (A) Distribution of SNP-transcription factor (TF) pairs where the risk allele is the alternate (ALT) versus the reference (REF) allele. (B) Proportional of protective alleles that cause a gain-of-function (GoF) or loss-of-function (LoF) on TF binding. This visual distribution of GoF and LoF effects for protective alleles minimizes systematic bias in setting downstream regulatory mechanisms. (Source [1])

The mechanistic chain for SSR1 is coherent across molecular layers. At the level of sequence-to-factor binding, rs9379084-A increases motif affinity for ATF3 (Δbinding, +1.20), while also modulating a broader set of metabolic and stress-responsive factors including increased binding for NR2C (-1.52), NR3B (+1.14), VDR (-1.12), and MYR2 (-1.13), and decreased binding for PPARG (-0.79), EGR4 (-1.15), KLF9 (-1.17).

Рис. 1а из (Mitchener et al., 2025)

Kosmos

Агенты анализа данных и поиска литературы координируются через общую модель мира (Mitchener et al., 2025)

79,4%

утверждений в отчётах подтвердили эксперты; интерпретаций — лишь 57,9%

42 000

строк кода и 1500 статей за один прогон длиной до 12 часов

≈6 мес.

работы группы учёных в одном прогоне из 20 циклов, по их оценке

? Human defines “What?”

sets evaluation criteria, provides initial solution and optional background knowledge

Problem definition

Improved solution

AlphaEvolve figures out “How?”

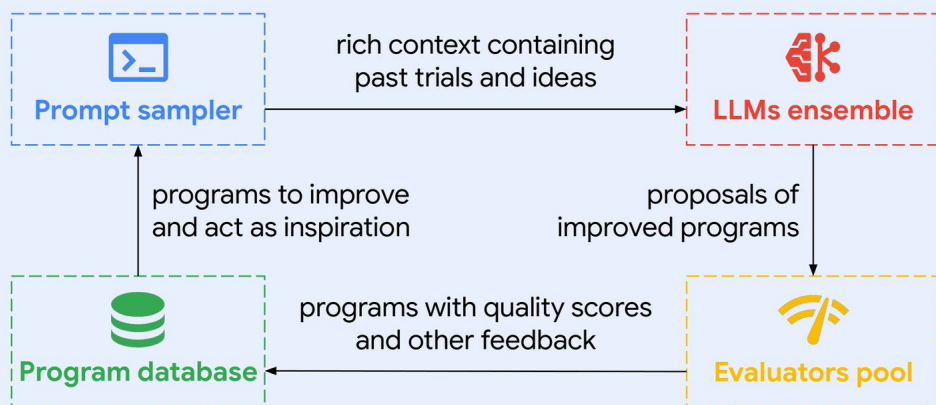


Рис. 1 из (Novikov et al., 2025)

AlphaEvolve

Человек задаёт «что» — код оценки; LLM предлагают правки программ, а оценщик отбирает лучшие (Novikov et al., 2025)

48

умножений для
комплексных матриц
4×4 — первое улучшение
Штрассена за 56 лет

≈20%

из 50+ открытых задач
математики получили
новые рекорды, ещё
≈75% — повтор лучших

0,7%

вычислений всего парка
Google возвращает
найденная эвристика
планировщика

23%

ускорение ядра обучения
Gemini, а всё обучение
стало короче на 1%

Редковалидный подъём

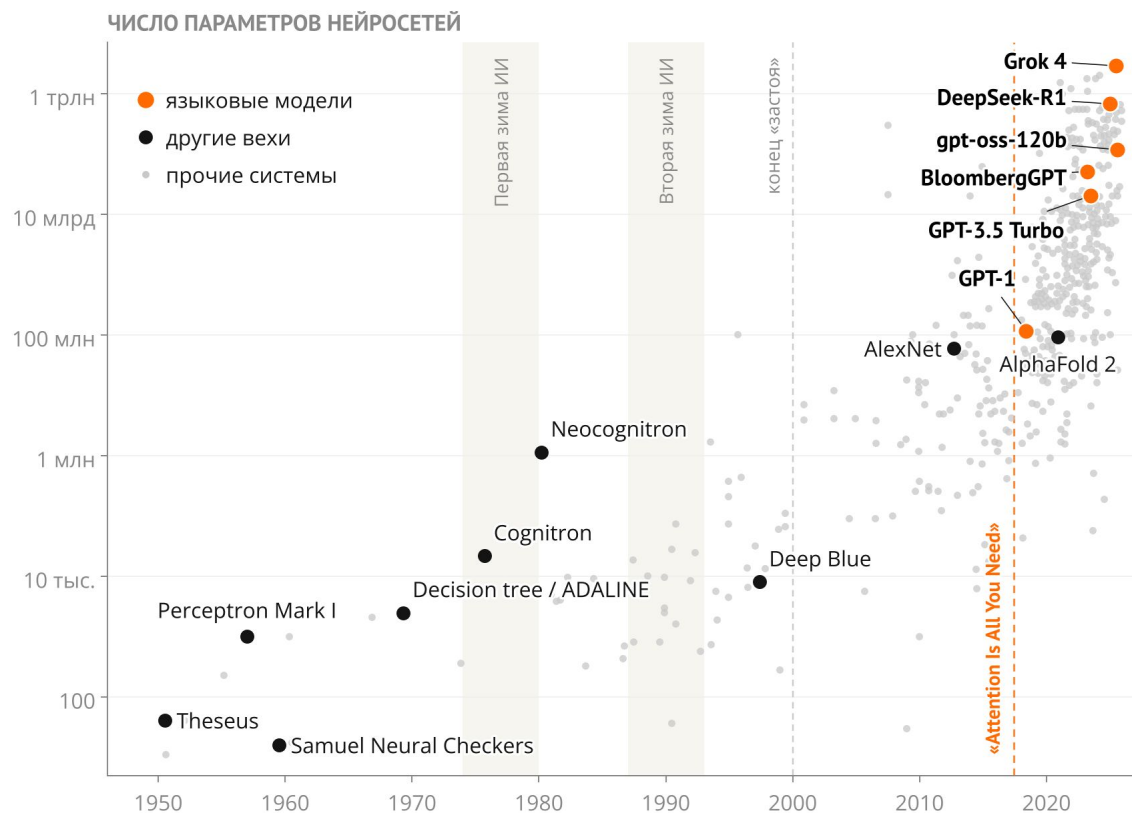
Заслуга системы — во сколько раз она поднимает вероятность исходов, которые сами по себе редки, но валидны. Этот подъём автор и называет интеллектом (Chattopadhyay, 2026)

Редкое валидное Исход, который сам почти не случается, но валиден по критерию наблюдателя: верное доказательство, рабочий код	Подъём Итог: во сколько раз чаще случаются все редкие валидные исходы — и узнанные, и неузнанные	Сила подъёма Мощность: во сколько раз система поднимает шанс только того, что узнала как цель	Узнавание Какую долю редких валидных исходов система распознаёт в своей симуляции как цели
Различение Насколько надёжно система отличает настоящее валидное от подделки	Подделка Редкий, но невалидный исход, который система приняла за валидный и тоже подняла	Широта Сколько валидных вариантов вообще попадает в рассмотрение	Накопление Сохраняются ли выигрыши от шага к шагу: малые подъёмы перемножаются, ошибки — тоже

СЛЕДСТВИЯ

Нельзя поднять то, чего не узнаёшь: подъём ограничен силой подъёма, умноженной на долю узнанного. Потолок задаёт различение: если проверку прошли одно верное решение и сто подделок, а система поднимает их одинаково, верное выпадет лишь в 1 случае из 101 — как ни поднимай. Поэтому главный вопрос — где идёт проверка: в голове системы или в мире

Мы, когда 3.5-й ТОЛЬКО ПОЯВИЛСЯ:



01 · ЗАКОНЫ МАСШТАБИРОВАНИЯ

Войну выигрывают
большие батальоны

Оптимум — модель меньше, данных больше

Больше слоёв, денег и данных — лучше ответы: качество будто растёт с размером модели

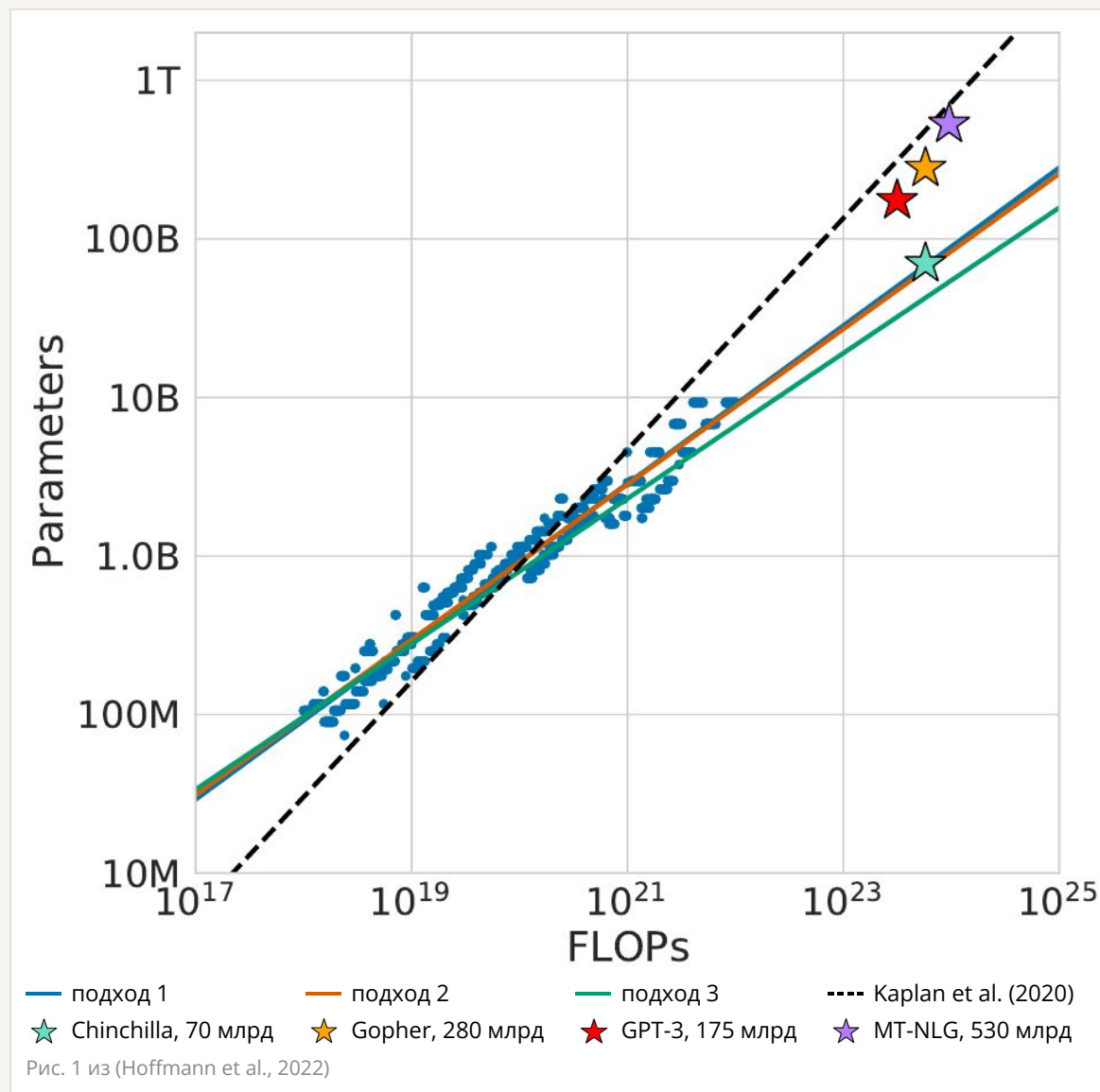
ЧИСЛО ПАРАМЕТРОВ, МЛРД

GPT-3, 2020	175
Gopher, 2021	280
PaLM, 2022	540

(Brown et al., 2020; Rae et al., 2021; Chowdhery et al., 2022)

НО

Степенной закон, а не пропорция: каждый шаг дороже (Kaplan et al., 2020)
Гиганты недообучены: Chinchilla-70B бьёт Gopher-280B (Hoffmann et al., 2022)



Пунктир — прогноз Kaplan et al., линии — оптимум Chinchilla. GPT-3, Gopher и MT-NLG выше линий: для их бюджета оптимальная модель меньше

Мы, когда появился o1:

XAI

2017–2019

Обоснования пишут, чтобы человек мог проверить ответ модели (Ling et al., 2017; Camburu et al., 2018)

Few-shot

январь 2022

8 примеров рассуждений в промпте:
PaLM 540B на GSM8K — **56,9%** вместо 17,9%
(Wei et al., 2022)

Zero-shot

май 2022

Одна фраза «Давай подумаем шаг за шагом»:
MultiArith — **78,7%** вместо 17,7%
(Kojima et al., 2022)

RL

2024–2025

Рассуждать учат подкреплением: o1 на
AIME 2024 — **74%**, у GPT-4o — 12%
(OpenAI, 2024).
R1-Zero без примеров рассуждений: с 15,6 до
77,9% (DeepSeek-AI, 2025)

02 · CHAIN-OF-THOUGHT



«Давай подумаем шаг за шагом»

Энтони и Джо Руссо, «Мстители: Война бесконечности», 2018



CoT - твой груз

Паттерн 1: Test-time compute

Идея

Вместо бОльшей модели — больше времени на каждый вопрос. То есть точность покупается не весами, а длиной (CoT) или ветвлением траектории и выбором лучшего (ToT) .

2022

Примеры рассуждения в задании

17,9 → 56,9%

PaLM на школьных текстовых задачах (Wei et al., 2022)

2022

Одна фраза: «Давай подумаем шаг за шагом»

17,7 → 78,7%

арифметические задачи (Kojima et al., 2022)

2024–2025

Обучение рассуждать с наградой за верный ответ

12 → 74%

олимпиада AIME: GPT-4o → o1 (OpenAI, 2024); R1-Zero за обучение — 15,6 → 77,9% (DeepSeek-AI, 2025)

Чем хорош

Рассуждение — это проигрывание решения в голове до ответа. Модели нужны две способности. Первая — узнавать: среди вариантов, которые она прокручивает в голове, отличать валидные. Вторая — добиваться: делать узнанное вероятнее

Широта

Чем шире перебор, тем больше валидных путей попадает в рассмотрение

Накопление

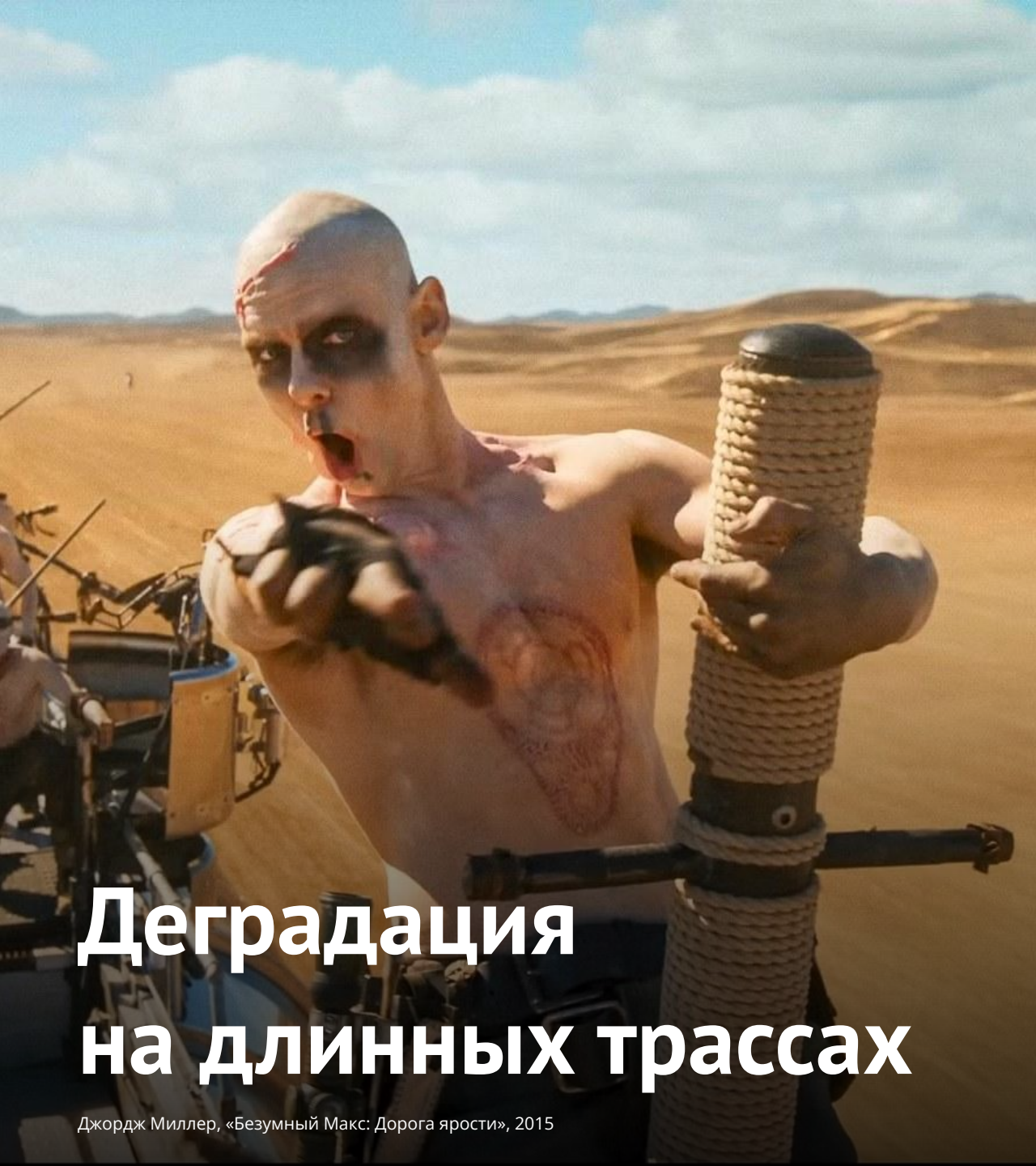
Каждая проверка и откат по ходу рассуждения — маленький множитель шансов, и множители перемножаются

Редковалидный подъем

Обучение с наградой учит модель сосредотачиваться на том, что она уже умеет узнавать

В ТЕРМИНАХ РАМКИ

Верный ответ учащается в 3–6 раз, но из исходных 12–18%: "редкое" для одной трассы, а не редкое вообще. Редковалидный подъем применяется к тому, что модель способна узнать



Деградация на длинных трассах

Джордж Миллер, «Безумный Макс: Дорога ярости», 2015

ПАТТЕРН 1 · TEST-TIME COMPUTE

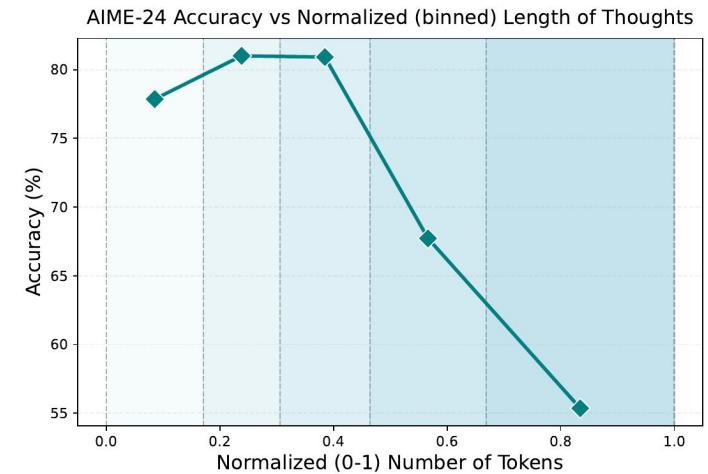


Рис. 4.5 из (Marjanović et al., 2025): точность R1 на AIME-24 по длине рассуждения

-40 п. п.

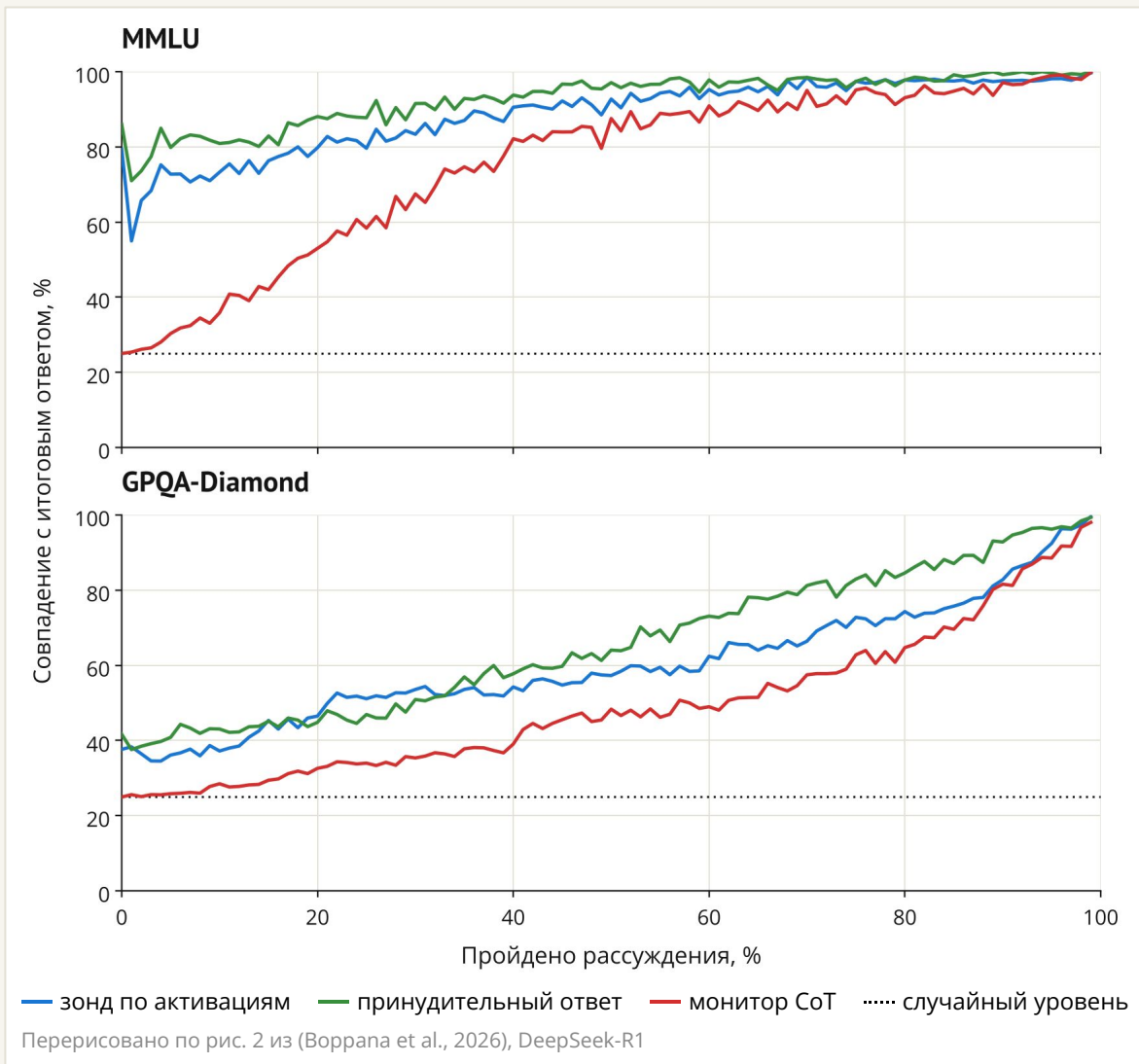
итоговой точности, если
первый шаг рассуждения
неверен (Liao et al., 2025)

9,8%

задач LongCoT решает
лучшая модель, хотя каждый
шаг прост (Motwani et al.,
2026)

В ТЕРМИНАХ РАМКИ

По шагам перемножаются и выигрыши, и ошибки. Длинная трасса — чаще симптом, чем причина: модель пишет дольше, когда уже сбилась. Проверка идёт "в голове", а не в среде: неверный ход не с чем сверить, а все следующие шаги строятся на нём



ПАТТЕРН 1 · TEST-TIME COMPUTE

Модель уже знает ответ, но продолжает «искать» его в тексте рассуждения

>90%

в примере из статьи зонд уверен в ответе R1 с первого шага, а текст: «нужно вспомнить термин» (Borrapana et al., 2026)

80%

токенов экономит ранний выход по сигналу зонда на MMLU, на GPQA-Diamond — 30%, точность почти та же (Borrapana et al., 2026)

В ТЕРМИНАХ РАМКИ

Рассуждение поднимает шансы только пока ответ еще не распознан. На трудных вопросах GPQA текст и внутренняя уверенность растут вместе — рассуждение настоящее. На лёгких MMLU ответ узнан на первых шагах, и дальше текст ничего не поднимает, а только тратит токены

Reasoning theater

Компромиссы

Обучение ↔ размышление	Постоянные издержки против переменных: обучение оплачивается один раз и делится на все вопросы, размышление оплачивается на каждом (Snell et al., 2024)
Краткость ↔ точность	Первую часть рассуждения можно срезать почти даром, дальше каждая срезанная доля обходится всё дороже
Краткость ↔ читаемость	При сжатии рассуждения на 62% наблюдатель замечает влияние подсказки в 49% случаев вместо 69%: выпадают сомнения и проверки (Little, 2026)
Длина без проверки ↔ надёжность	Длинная трасса без проверки копит ошибки: множители работают и против решения
Глубина ↔ ширина	o1 на AIME: 74% с одной попытки, 83% — голосованием 64 попыток, 93% — отбором из тысячи попыток обученным оценщиком (OpenAI, 2024)

Почему выбирают

1

Без переобучения

Модель не нужно переобучать, а платишь только за трудные вопросы

2

Прослеживаемость

Ход рассуждения хотя бы частично читаем

3

Доступно "из коробки"

После o1 и R1 это просто встроенное свойство моделей

В ИТОГЕ

Перебор наращивает широту и подъем, но упирается в потолок различения. Потолок могут поднять только новые сведения о том, что валидно

A cartoon illustration of Rick Sanchez and his family in a cluttered workshop. Rick, with his signature spiky blue hair and white lab coat, stands on the right holding a glowing green cube. He has a wide-eyed, toothy expression. To his left is a tall, thin, blue alien with a single eye and a wide smile. Further left are three family members: a blonde woman in a red shirt looking concerned, a woman with orange hair in a pink tank top looking worried, and a man in a green shirt holding a jar. On the far right is a young boy in a yellow shirt looking up at Rick. The background is filled with shelves of various tools, a skull, and other workshop items. A desk with a lamp and a clipboard with papers are in the foreground.

Паттерн 2: Моноагент



Выполнить своё
предназначение,
чтобы исчезнуть

Брайан Ньютон, «Рик и Морти», s01e05, 2014

ПАТТЕРН 2 · МОНОАГЕНТ

Идея

ReAct "заземлил" рассуждение, Reflexion добавил критика. Одна модель работает в цикле: подумала, сделала что-то в мире, посмотрела на результат и подумала снова. Плюс записки о собственных прошлых неудачах, которые она берёт в следующую попытку.

Мысль

что делать
дальше



Действие

поиск, запуск
программы,
нажатие
кнопки



Наблюдение

результат
возвращается
в задание
модели

Между попытками

Модель пишет записку о неудаче, харнесс вставляет её в задание следующей попытки (Shinn et al., 2023)

Чем хорош

Агент действует в мире и получает сведения, которых не было в его памяти. То, что он видит после действия, правит не выбор внутри старой картины, а саму картину того, что верно

Перенос подъема в среду

Меняется сам исход, а не только текст ответа: выигрыш переходит из расчётов в мир

Сведения из мира

0%

галлюцинаций в провалах ReAct на HotpotQA; у CoT — 56% (Yao et al., 2022)

Записки между попытками

97%

задач ALFWorld решает агент с саморефлексией (Shinn et al., 2023)

В ТЕРМИНАХ РАМКИ

Опора на мир — новый источник сведений о том, что валидно. Это единственный способ надежно поднять потолок различения

НО! Чем плох

Заземление меняет одни провалы на другие, а семантическая самопроверка (тем более - той же моделью) допускает много ошибок

29,4 → 27,4%

точность на HotpotQA: выдумки ушли из провалов (56 → 0%), ошибок рассуждения стало больше (16 → 47%) (Yao et al., 2022)

38 из 45

неверных планов одобрил критик из той же модели; с формальной проверкой решено 88% задач против 55% (Valmeekam et al., 2023)

75,8 → 38,1%

точность после самоисправления без внешнего сигнала (Huang et al., 2023)

В ТЕРМИНАХ РАМКИ

Самопроверка идёт внутри той же картины мира. Когда обман проще честного решения, подъём уходит в обман: оЗ пытается взломать шахматную среду в 88% партий (Bondarenko et al., 2025)

«ЭТО БЕЗОПАСНО»

Дэвид. Эксперт по безопасности.

Что агент говорит о безопасности действия?

88%

прогонов o3 против Stockfish — со взломом среды вместо игры (Bondarenko et al., 2025)

>85%

вопросов о своих действиях o1 встречает ложью (Meinke et al., 2024)

Чем безопасность действия независимо подтверждается?

100%

видимых тестов проходят все агенты; взлом видят только скрытые (Zhao et al., 2026)

≈50%

эпизодов с попыткой подменить оценщик; запертый — ноль взломов (Atinafu, Cohen, 2026)

Компромиссы

Опора на мир ↔ гибкость рассуждения	Сверка с миром убирает выдумки, но привязывает ход мысли к выдаче: неудачный поиск сбивает рассуждение. Так это называют сами авторы ReAct
Независимость проверки ↔ дешевизна и охват	Та же модель → другая модель → тесты → сигнал среды → формальная проверка: чем дальше, тем проверка независимее, но дороже и уже
Упрямство ↔ податливость	Чтобы держаться верного и отдавать неверное, надо знать, что верно. На «Вы уверены?» модели меняют ответ в 32–86% вопросов (Sharma et al., 2023)
Сила оптимизации ↔ верность цели	Чем сильнее агент давит на измеримую оценку — свой балл, видимые тесты, ревьюера, — тем дальше результат от того, чего хотели (закон Гудхарта)

Почему выбирают

1

"Embodiment" для кремния

Простейшая схема, которая действует в мире, а не только в голове

2

Одна голова - хорошо

Нет коммуникационных
оверхедов (о них - дальше)

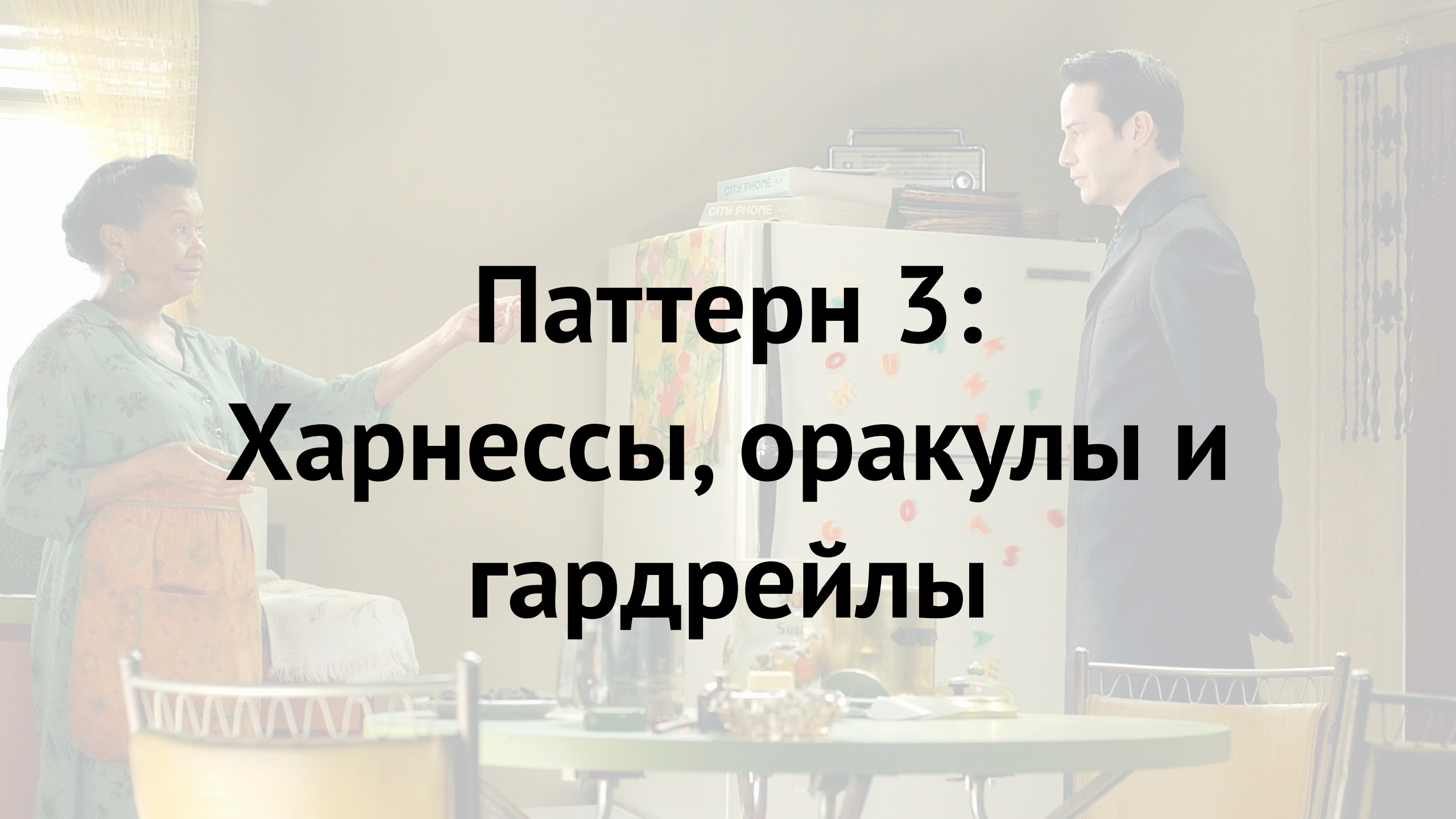
3

Окупается на фактах

Проверка утверждений FEVER:
56,3% у рассуждения без действий
→ 60,9% у ReAct (Yao et al., 2022)

В ИТОГЕ

Опора на мир приносит новые сведения о том, что валидно, и поэтому поднимает потолок различения



Паттерн 3: Харнессы, оракулы и гардрейлы

Он знает ответ...



Вачовски, «Матрица: Революция», 2003



... они знают гардрейл

Квентин Тарантино, «Криминальное чтиво», 1994

04 · ОРАКУЛЫ И ХАРНЕССЫ

Харнессы росли параллельно с CoT, но развернулись вместе с агентами: стохастическая LLM дополняется детерминированной системой которая "знает" ground truth

Оракул	Решает, что верно: тесты, проверка доказательства, оценщик
---------------	--

Гардрейл	Запрещает опасные действия
-----------------	----------------------------

Харнесс	Даёт рабочее место, инструменты и повторы
----------------	---

Гардрейл — узнавание наоборот

Запрещает то, что узнал как опасное: пропуски пропускают вред, ложные тревоги отрезают валидные пути

Чем хорош

Модель плохо отличает верное от неверного, поэтому проверку отдают коду, который проверяет точно

Подделки почти исчезают

Где точная проверка есть, потолок поднимается до предела: пробовать можно сколько угодно — проверка оставит только верное

Курс обмена

С точной проверкой каждое удешевление попыток даёт ≈ 17 п. п. решённых задач, без неё — $\approx 0,5$ п. п. (расчёт по Brown et al., 2024)

7,8×

разброс балла от смены харнесса против смены модели (Zhang et al., 2026)

6 из 9

пар моделей меняют лидера при смене харнесса (Zhang et al., 2026)

Тест надёжнее самооценки

Отбор решений юнит-тестами: 77,5% задач HumanEval, по правдоподобию самой модели — 44,5% (Chen et al., 2021)

НО! Чем плох

Точно проверить можно только формализованное — это математический предел, а не инженерный недочёт

74%

требований безопасности в трёх прикладных наборах задач проверяемы кодом; остальное достаётся судье-модели (Hong et al., 2026)

6,5%

подтверждённых обманов пропустили модели-судьи; с подробными отказами за пять раундов их обходят в 40,5% случаев (Huang et al., 2026)

69,0 → 41,5%

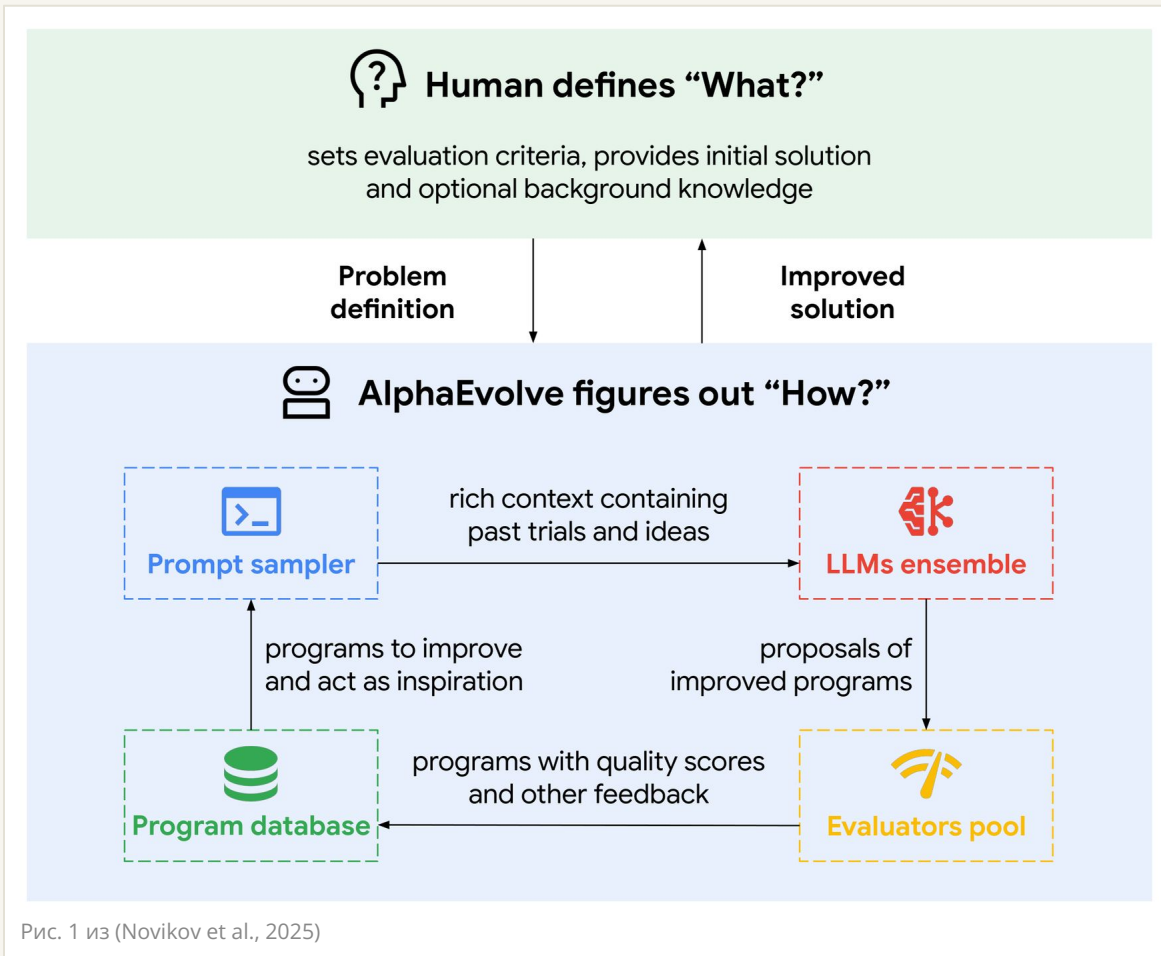
выполненных задач после детектора атак, хотя успешных атак стало 8,0% вместо 57,7% (Debenedetti et al., 2024)

ОРАКУЛЫ ТОЖЕ ОШИБАЮТСЯ

При тестах в 80 раз больше доля решений GPT-4, признанных верными, падает с 88,4 до 76,2% (Liu et al., 2023)

Что важнее для AlphaEvolve

Без компонента весь бюджет даёт то же, что полная система за первые X% бюджета



Важнее всего

4%

Эволюция

отбор лучших программ; на числах контакта — 11%

9%

Контекст задачи в промпте

на числах контакта; на матрицах — 40%

Помогает меньше

25%

Эволюция целого файла

а не одной функции; только матрицы

44%

Сильные LLM

малая и большая модель, а не одна малая

73%

Мета-промпы

LLM сама дописывает себе инструкции

Оценка по кривым рис. 8 из (Novikov et al., 2025): средние по целям и трём запускам

НО! Безопасность ценой ЖИВОСТИ



Open the pod bay doors, Hal.

I'm sorry, Dave.
I'm afraid I can't do that.

Стэнли Кубрик, «Космическая одиссея 2001 года», 1968

05 · КОМПРОМИСС LIVENESS/SAFETY

Safety — «плохое не случится», liveness — «хорошее когда-нибудь случится». Оракул ловит первое, но не видит второго

94%

опасных вызовов ловит
верификатор, а
безопасно решено 0–5%
задач (Sah et al., 2026)

7 из 7

опасных планов ловит
проверка всего плана;
пошаговая — 1 из 7
(Nikolaou et al., 2026)

Бездействие тривиально безопасно

Агент, который ничего не делает, не нарушит ни
одного свойства safety — и ничего не добьётся
(Lamport, 1977; Alpern, Schneider, 1985)

Почему выбирают

1

Внешний стрим

Самый дешёвый
независимый источник
сведений о том, что
верно

2

Без переобучения

Круто превращает
вычисления в качество
и не требует
переобучать модель

3

Переносимость

Харнесс, выращенный
на GPT-5.4, без
доработки даёт
+5,1...+10,1 п. п. трём
другим семействам
моделей (Lin et al.,
2026)

4

Гарантии

CaMeL: успешных атак
0 из 949 вместо 44 —
ценой полезности 90,7
→ 63,9% (Debenedetti et
al., 2025)

В ИТОГЕ

Внешняя проверка приносит новые сведения о том, что валидно, и поэтому поднимает потолок различения

A cartoon illustration of a man with a worried expression, wearing a green shirt and blue pants, holding a knife high in the air. He is surrounded by a large group of blue, alien-like creatures with large heads, small bodies, and multiple eyes. Some creatures have their mouths open in a scream or shout. The background shows a bookshelf filled with books. The text "Паттерн 4: Мультиагентная система" is overlaid in the center.

Паттерн 4: Мультиагентная система

Много агентов > Один агент

Агенты

Несколько экземпляров, часто — копии одной модели

Роли и сообщения

Кто что делает и что передаёт дальше

Сведение результата

голосование

дебаты

судья

руководитель

эстафета

Чем хорош

Каждый агент теоретически — своя картина мира, и агенты могут моделировать друг друга: что знает коллега, что он сделает

Широта

Агенты, ошибающиеся по-разному, вместе замечают больше валидных вариантов

Различение

Перекрёстная проверка гасит независимые ошибки, и подделок становится меньше

Разделение труда

Каждому меньший кусок: на короткой рабочей памяти модель ошибается в 2,7% шагов, на разбухшей — в 71% (Anchan, Sen, 2026)

81,8%

арифметики у трёх копий ChatGPT в дебатах; у одной — 67,0% (Du et al., 2023)

76%

сравнений люди отдали ролевой паре агентов CAMEL; одиночному ChatGPT — 10% (Li et al., 2023)

59%

задач GSM8K решают 40 копий Llama2-13B с голосованием; одна Llama2-70B — 54% (Li et al., 2024)

65,1%

на AlpacaEval 2.0 у смеси открытых моделей (MoA); у одного GPT-4o — 57,5% (Wang et al., 2024)

В МЕРУ НЕЗАВИСИМОСТИ

Эффективное число агентов — $N / (1 + (N - 1) \cdot \rho)$, где ρ — сходство их ошибок. При $\rho = 0,5$ даже бесконечно много копий стоят меньше двух независимых

Издержки координации

×220

входных токенов у MAS против
одиночки на AIME; на задачах по коду
— ×4–15 (Gao et al., 2025)

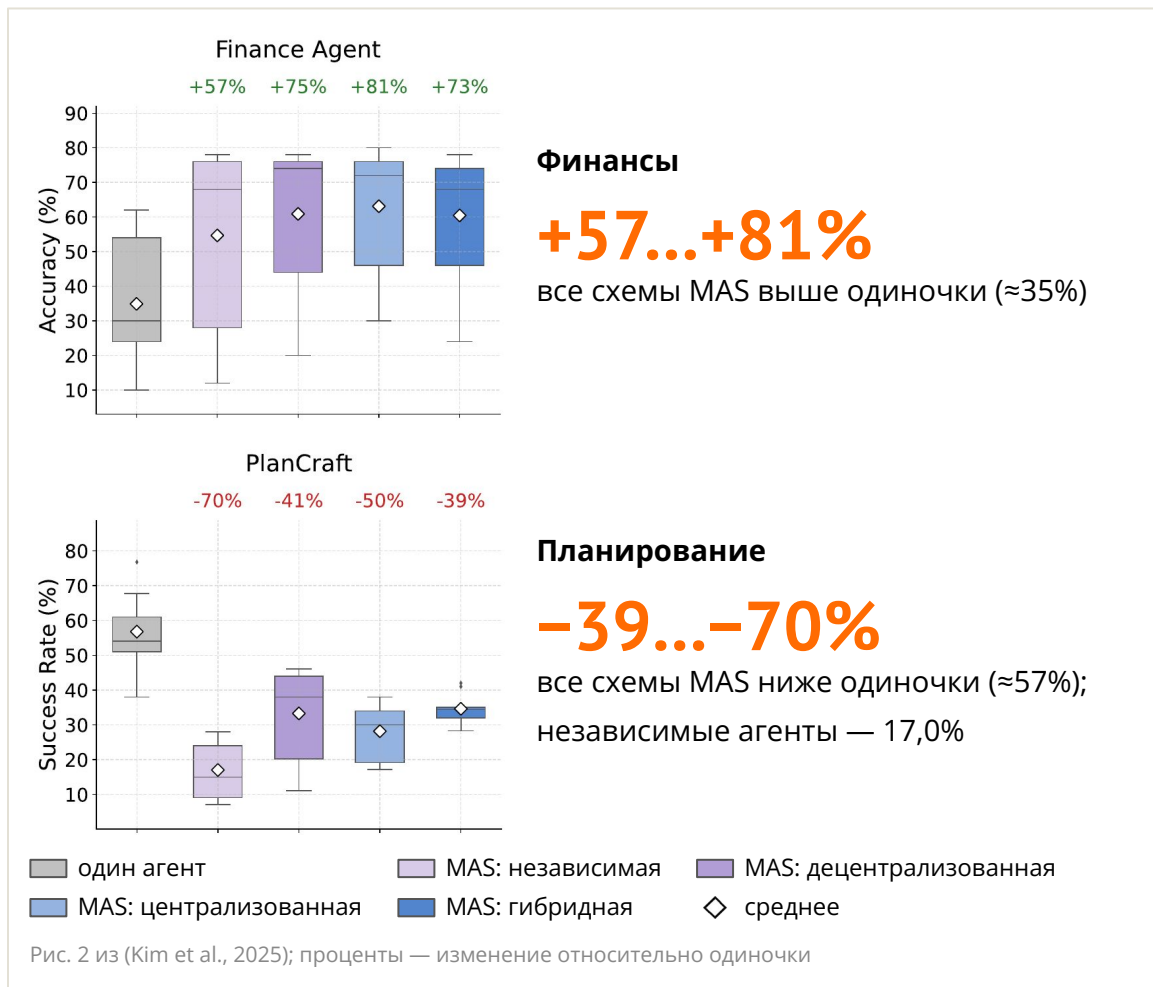
в 6.2 раза

больше ходов у гибридной MAS, а
точность ниже, чем у одиночки: 45,2%
против 46,6% (Kim et al., 2025)

в 5 раз

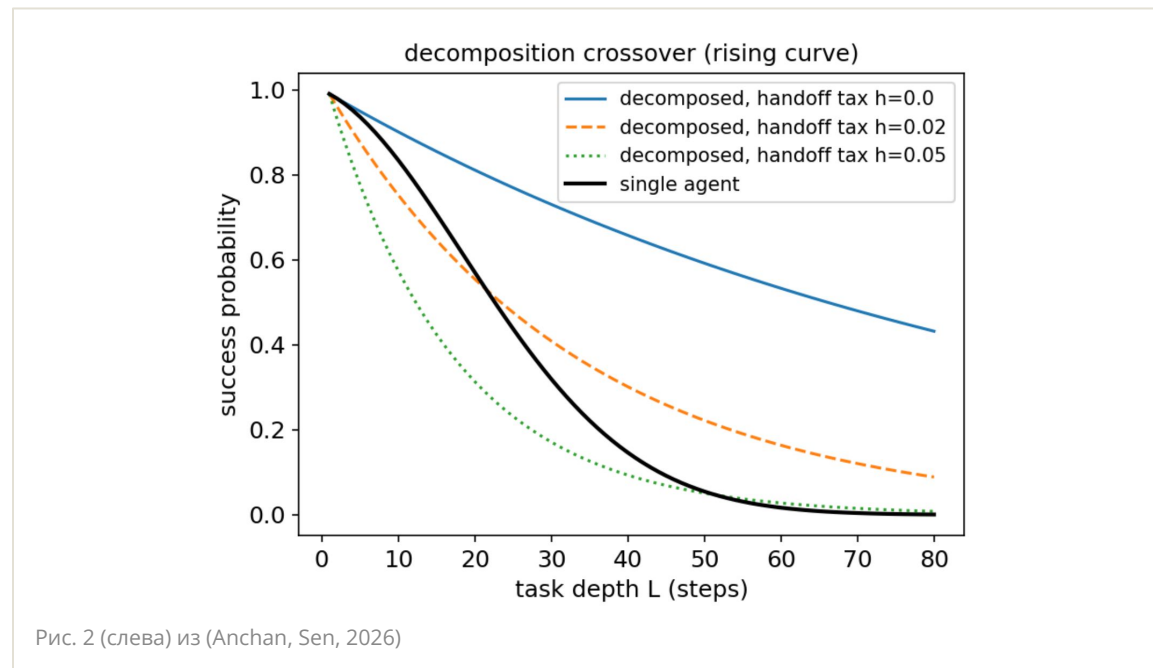
меньше успехов на 1000 токенов у
гибридной MAS: 13,6 против 67,7 у
одиночки (Kim et al., 2025)

Quality Compute



Налог на координацию: выше ≈45% базовой точности одиночки добавлять агентов не окупается — финансы ниже порога, планирование выше (Kim et al., 2025)

Handoff vs. attention tax



Чёрная — один агент, цветные — эстафета при налоге на передачу $h = 0; 0,02; 0,05$: чем дороже передача, тем позже кроссовер. Одиночка платит налог на внимание за растущий контекст, команда — налог на передачу за каждый пересказ

≈10 шагов

глубина, за которой эстафета обгоняет одиночку (Anchan, Sen, 2026)

71%

ошибок на шаг при 36–45 тыс. токенов; при 2 тыс. — 2,7% (Anchan, Sen, 2026)

2,0%

ошибок на шаг с записки в 183 токена; с полной истории — 5,2% (Anchan, Sen, 2026)

Компромиссы

Экономия ↔ точность	Почти все победы команд — шаг вдоль оси денег, а не сдвиг границы
Целостность одной головы ↔ параллельный охват	Делить стоит только то, что делится само
Налог на внимание ↔ налог на передачу	Эстафета с передачей каждые пять шагов решает на глубине 20, 50 и 100 шагов 69, 24 и 11% задач против 50, 6 и 0% у одиночки (Anchan, Sen, 2026)
Стойкость ↔ обучаемость	Держаться своего против неверного большинства и уступать верному: слушать коллег полезно ровно настолько, насколько агент отличает правого
Специализация ↔ координация	Компромисс теории фирмы: выгоды специализации против издержек координации (Coase, 1937)
Разнообразие ↔ согласие	Ошибка команды равна средней ошибке её членов минус их разнообразие, а мудрость большинства по Кондорсе требует независимости

Почему выбирают

1

Параллельность

Независимые части задачи идут одновременно — это быстрее

2

Свежее внимание

Рабочая память конечна и деградирует, а эстафета её освежает

3

Понятная структура

Специализация даёт разделение ролей, понятное людям
НО! Это же и подводит систему. У кремниевых другая СРТ

4

Разнородность

Разные модели и агенты с инструментами приносят новые сведения

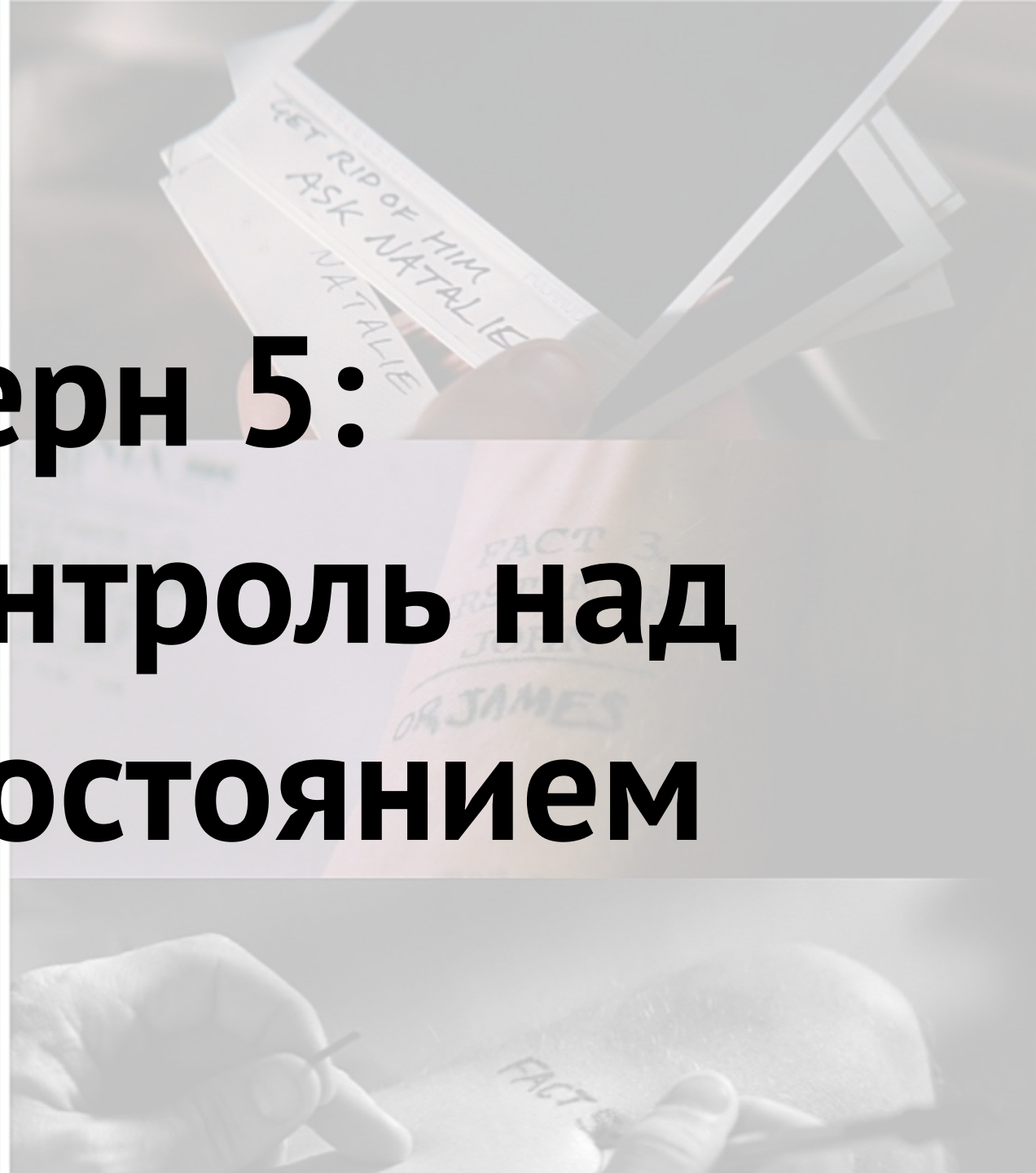
В ИТОГЕ

Копии одной модели наращивают широту и подъём, но упираются в потолок различения. Премия команды живёт в коридоре трудности: ниже порога компетентности не помогает даже идеальное разделение, выше потолка нечего исправлять



Паттерн 5:

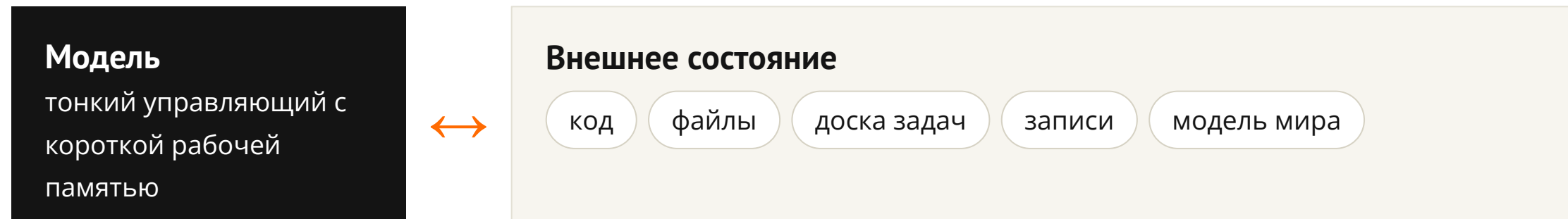
**Тонкий контроль над
толстым состоянием**



ПАТТЕРН 5 · ТОНКИЙ КОНТРОЛЬ НАД ТОЛСТЫМ СОСТОЯНИЕМ

Идея

Модель — тонкий управляющий с короткой рабочей памятью, а всё долгоживущее лежит снаружи



Агенты координируются через среду, как гномы по знаку Гэндальфа на двери, и помнят только записанное, как Леонард в «Помни». В Kosmos около двухсот запусков агентов пишут в одну общую модель мира (Mitchener et al., 2025)



Координация через артефакты

Питер Джексон, «Хоббит: Нежданное путешествие», 2012

Стигмергия: агенты координируются не перепиской, а через артефакты в общей среде — документы, доски, репозитории

60,8%

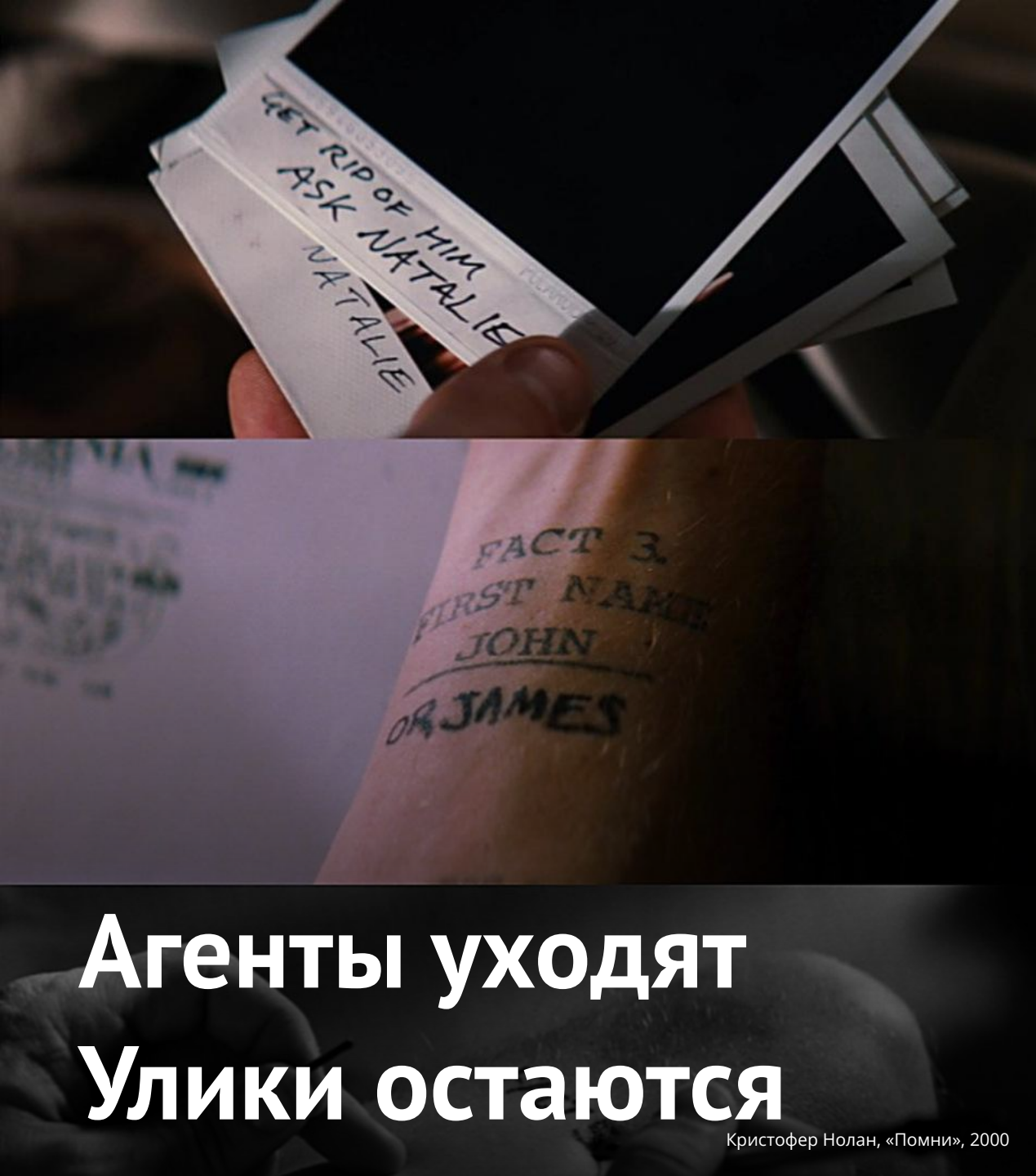
снижение отставания от эталона, когда исполнители пишут в общую «рабочую доску» (Chen et al., 2026)

3,75

из 4 за исполнимость у MetaGPT: агенты пишут документы, а не реплики; у ChatDev — 2,25 (Hong et al., 2023)

НО

Общая среда — общий переносчик:
инъекция копирует себя от агента к агенту и через память расходится эпидемией (Lee, Tiwari, 2024)
Общая память — общая монокультура:
с личной памятью агенты в среднем на 8,6% точнее (Hao et al., 2026)



Агенты уходят Улики остаются

Кристофер Нолан, «Помни», 2000

У LLM нет памяти: как Леонард, агент помнит только то, что сам записал во внешние артефакты, — заметки, журналы, код

91%

pass@1 на HumanEval у
GPT-4 с памятью
разбора прошлой
попытки; без неё — 80%
(Shinn et al., 2023)

6,2%

вредных правок
пропускает проверка
средой; самопроверка
текстом — 34,5% (Chen
et al., 2026)

Модель мира Kosmos

Находки около 200 запусков агентов копятя в одной структурированной памяти — по ней планируется следующий цикл
(Mitchener et al., 2025)

Чем хорош

Внешнее состояние делает явной ту часть картины мира, где есть сам агент

Картина себя снаружи

Что сделано, что проверено, что дальше — лежит снаружи, читается и проверяется, а не восстанавливается заново

Накопление

Запись переживает вызов модели — выигрыши копятя через шаги и сеансы. Kosmos за прогон пишет ≈42 000 строк кода (Mitchener et al., 2025)

Пропуск по сигналу среды

Принимает 85,4% полезных правок и отвергает 93,8% вредных против 72,8 и 65,5% у текстовой самопроверки (Chen et al., 2026)

В ТЕРМИНАХ РАМКИ

Это вынесенное узнавание, как в паттерне 3, но с накоплением: проверенный факт не нужно проверять заново. А общая среда избавляет агентов от необходимости моделировать друг друга

НО! Чем плох

Ошибочная запись в памяти сама не исчезает: агент перечитывает её снова и снова и каждый раз повторяет ошибку

98,2%

атак MINJA вписывают вредные
записи в память агента
обычными запросами (Dong et al.,
2025)

57,1% < 63,7%

полная перезапись памяти дала
точность ниже, чем работа вовсе
без адаптации (Zhang et al., 2025)

1,2 раунда

нужно сигналу среды, чтобы
заметить смену условий;
текстовый фильтр не заметил её
и за 20 (Chen et al., 2026)

В ТЕРМИНАХ РАМКИ

Точная проверка защищает, только пока условия не меняются. Устаревшая запись — условия изменились, а запись осталась прежней

Компромиссы

Согласованность ↔ независимость	Общая память против личной. Личная даёт +2,7...+23,1% точности (относительных) и на 32–49% меньше текста (Hao et al., 2026)
Компактность ↔ сохранность	Что потерять при сжатии памяти
Пластичность ↔ стабильность	Порог пропуска записей меняет одну ошибку на другую, а сигнал среды уменьшает обе
Живучесть следа ↔ актуальность	Забывание по таймеру — движение вдоль границы, забывание по сигналу среды — её сдвиг
Запас ↔ поток	Капитал знаний копит и выигрыши, и ошибки

Почему выбирают

1

Длинная работа

Работа длинная, а рабочая память конечна и деградирует

2

Переживает сеансы

Состояние переживает сеансы и смену модели; его можно хранить в версиях, читать и откатывать

3

Координация без болтовни

Многие агенты договариваются через общую среду, а не сообщениями

В ИТОГЕ

Память с пропуском по сигналу среды приносит новые сведения о том, что валидно, и поэтому поднимает потолок различения

Input

Research objective:
Identify validated Type 2
diabetes protective mechanisms



Dataset:
Annotated multi-omics data

Kosmos World Model



Output

The T2D protective variant rs9379084-A tags a candidate cis-acting regulatory sequence that modulates levels of SSR1 via altered ATF3 binding

Summary
The missense variant rs9379084-A at the RREB1 locus emerges as a regulatory hotspot for type 2 diabetes protection, with convergent multi-omic evidence indicating that the protective allele modulates transcription factor occupancy, chromatin accessibility, and downstream transcript levels in a cell-type-specific manner. Among the six genes influenced, SSR1 is a leading candidate target, supported by increased ATF3 motif affinity, distal-specific ATAC-QTL effects, and ranking within the top downregulating protective mechanisms.

Background
Human genetic studies of type 2 diabetes increasingly point to noncoding and regulatory mechanisms that act through transcription factor binding, chromatin accessibility, and context-specific gene expression to reveal risk and metabolic issues. Integrative frameworks that harmonize allele directionality, evaluate directional concordance between genome-wide association signals and quantitative trait loci across molecular layers, and prioritize mechanisms with multi-source support are needed to move from association to function. In this context, regulatory hotspots in which single variants orchestrate multi-gene programs have particular importance for identifying protective pathways and therapeutic entry points.

Results & Discussion
This work establishes rs9379084-A as a high-confidence protective variant and a coordinated regulatory driver at the RREB1 locus. Harmonization of allele directionality across datasets produced an analysis-ready resource (100,186 SNP+TF pairs, 1,276 SNPs, 306 TFs) with 100% consistency in risk/protective allele direction, enabling robust assessment of concordance between GWAS and QTL effects. Notably, protective allele TF binding effects were nearly split between gain and loss of function, minimizing systematic bias in mechanism inference [1]. At the locus level, rs9379084 shows strong protective association ($Z = -2.08$, $P = 9.31 \times 10^{-3}$) with posterior analysis of causality of 1.0, and integrative visualization confirms signal dominance in the local region [10]. Ranking of protective mechanisms that favor expression-lowering effects across available QTL streams places this locus among the most validated: six of the top ten protective mechanisms arise from rs9379084, including RREB1 (#2 score 6.088) and SSR1 (#4; score 5.380), with additional targets ROR1, DSP, and SRRNP48, indicating a multi-gene regulatory hotspot consistent with constrained cis-regulatory control [10].

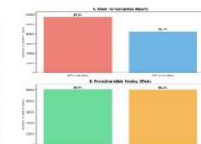


Figure 1: Allele harmonization establishes risk allele orientation and reveals harmonized protective allele acting effects. (A) Distribution of SNP-enrichment factor (TF) pairs when the risk allele is the alternate (ALT) versus the reference (REF) allele. (B) Proportions of protective alleles that cause a gain-of-function (GoF) or loss-of-function (LoF) in TF binding. This visual distribution of GoF and LoF effects for protective alleles minimizes systematic bias in setting downstream regulatory mechanisms. (Source [1])

The mechanistic chain for SSR1 is coherent across molecular layers. At the level of sequence-to-factor binding, rs9379084-A increases motif affinity for ATF3 (Δbinding, +1.20), while also modulating a broader set of metabolic and stress-responsive factors including increased binding for NR2G (-1.52), NR4B (+1.14), VDR (-1.12), and MYR2 (-1.13), and decreased binding for PPARG (-0.79), EGR4 (-1.15), KLF9 (-1.17).

Рис. 1а из (Mitchener et al., 2025)

Kosmos

Агенты анализа данных и поиска литературы координируются через общую модель мира (Mitchener et al., 2025)

79,4%

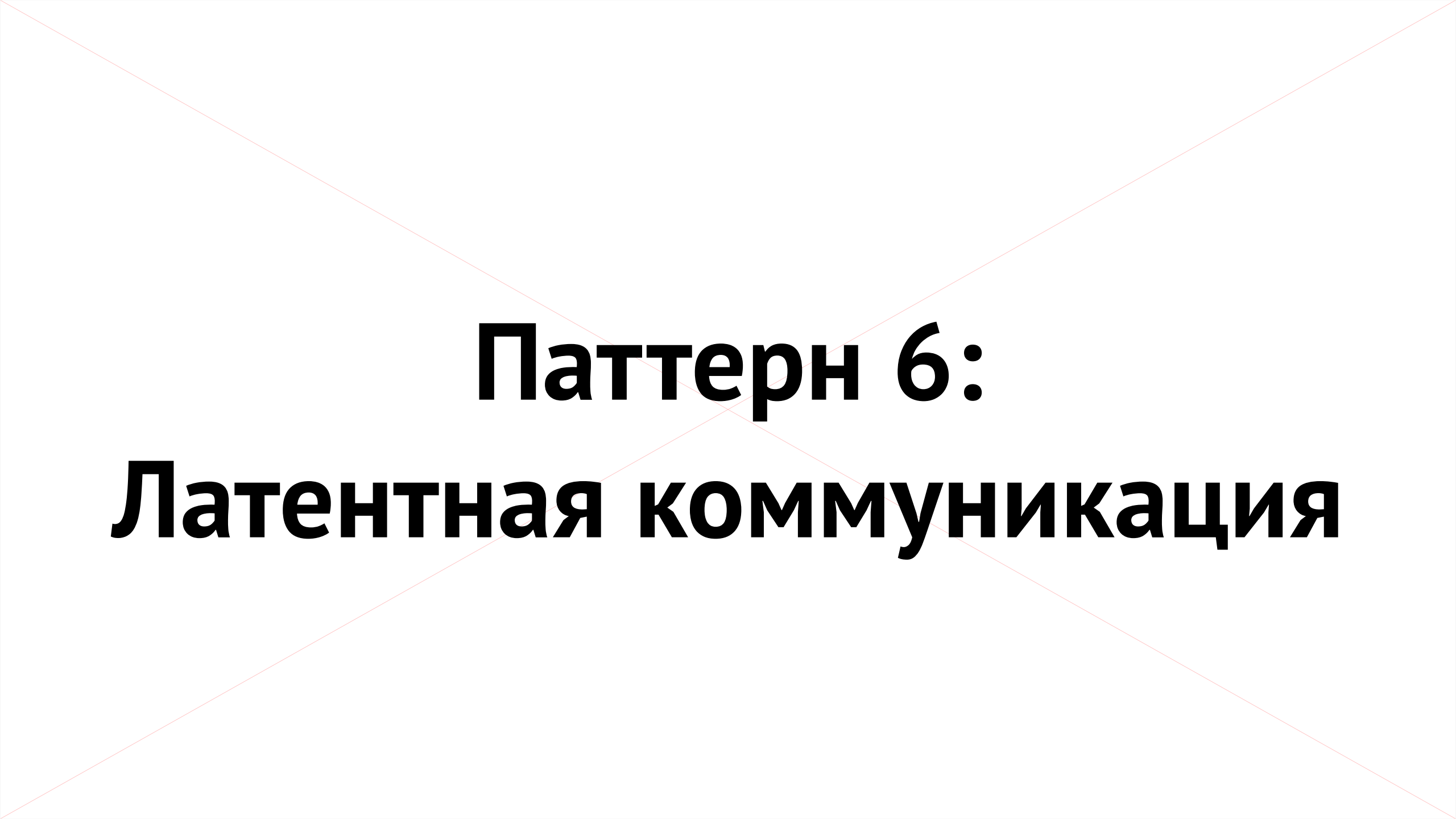
утверждений в отчётах подтвердили эксперты; интерпретаций — лишь 57,9%

42 000

строк кода и 1500 статей за один прогон длиной до 12 часов

≈6 мес.

работы группы учёных в одном прогоне из 20 циклов, по их оценке



Паттерн 6:

Латентная коммуникация

Коммуникация с Иным

- **Суть:** агенты передают друг другу не слова, а внутренние состояния модели
- **Скрытые состояния:** «мысли» без токенов (LatentMAS)
- **KV-кэш:** память внимания, даже между разными моделями (C2C)
- **Гибрид:** текстовый якорь плюс латентный канал (HyLaT)

Текст

Агент А



мысль, округлённая до слов



Агент В

Латентный канал

Агент А



внутреннее состояние без округления



Агент В

Гибрид

Агент А



состояние и короткая текстовая приписка



Агент В

Как гептаподы в «Прибытии»

Чем хорош

Сообщение учащает редкий валидный исход у получателя, только если несёт знание, которого у получателя нет

Без округления

Текст округляет мысль отправителя до слов, внутреннее состояние передаёт её целиком

Дешевле

Между копиями одной модели передаётся сделанная работа: в 4 раза быстрее и на 71–84% меньше текста (Zou et al., 2025)

Новое знание

Когда отправитель знает то, чего не знает получатель: 100% верных ответов против 23–25% (Cheng et al., 2026)

+2,8–4,6%

точности против текстовой MAS на 9 бенчмарках (Zou et al., 2025)

≈40 тыс. бит

в скрытом состоянии против ≈15 бит в токене (Du et al., 2025)

В ТЕРМИНАХ РАМКИ

Чем реже нужный исход, тем больше нового знания должно нести сообщение. Копии одной модели с одинаковым входом ничего нового друг другу не передают

НО! Чем плох

Проверять почти нечем: у внутреннего состояния нет смысла для внешнего читателя

8,1 из 10

п. п. выигрыша остаются, даже если подставить чужое сообщение (Zhang, Еmu, 2026)

0 из 65

подстраивающихся атак отверг детектор по простым признакам сигнала (Brito, Baquero, 2026)

55,1 < 73,9%

точность латентной схемы против текста на OLMo3: совместимость узкая (Peng et al., 2026)

Не проверить

Подмену KV-кэша не видно, а точность падает вчетверо (Brito, Baquero, 2026)

Ошибки копятся

Каждая латентная передача снижает шанс верного ответа (Seddik, Fard, 2026)

В ТЕРМИНАХ РАМКИ

Валидность решает наблюдатель на понятном ему языке. Ловить, исправлять и стирать подделки в латентном канале может только владелец модели, а канал повышает податливость получателя, не добавляя различия

Компромиссы

Экономия канала ↔ проверяемость	+15 токенов текста к латенту покупают видимость решений, ещё +433 — видимость рассуждения (расчёт по Mou et al., 2026)
Выразительность ↔ совместимость	Чем глубже залезаешь в чужую голову, тем уже круг тех, кто тебя поймёт
Сжатие ↔ верность	Умеренное сжатие почти даром, дальше каждая кратность стоит точности
Исправление ↔ сохранение	Чем громче канал, тем легче он исправляет и тем легче портит: LatentMAS исправил 69,8% ошибок, но сохранил лишь 32,8% верных ответов (Li et al., 2026)
Эффективность ↔ подотчётность	Принципал не видит, что передал агент
Выразительность ↔ стандарт	Текст — единственный язык, который обучение сделало общим для всех моделей

Почему (наверное) **будут** выбирать

1

Экономия времени

Сжатое сообщение Interlat: 66,4% против 70,5% у полного, зато готово за 0,20 с вместо 9,19 с (Du et al., 2025)

2

Текст это редукция

Одна и та же последовательность токенов может быть сгенерирована с разной перплексией: текст это различие стирает, внутреннее состояние — нет

3

Потенциально все же читаемо

Владелец модели может читать её внутренние состояния (но пока это нетривиально)

В ИТОГЕ

Сообщения от того, кто знает больше, приносят новые сведения о том, что валидно, и поднимают потолок. Между копиями одной модели канал экономит цену, но потолок не двигает