LECTURE CLUB · ВЕБИНАР

Как я научил своих AI-агентов улучшать себя каждый день

от 84 экспериментов до личного ассистента

НИКИТА РВАЧЕВ

T.ME/RVNIKITA_BLOG

ОБО МНЕ

Никита Рвачев

AI-исследователь в лаборатории компании из Fortune 100 · Нью-Йорк



01

Учусь

МГТУ им. Баумана, курсы MIT и Stanford — и не останавливаюсь

02

Делюсь

канал «Мысли Рвачева», Lecture Club, вебинары

03

Помогаю

компаниям с AI: консалтинг и свои проекты

О чем поговорим

01
Агент
Из чего он состоит: четыре блока

02
Теория
Что такое RSI и autoresearch, откуда это пришло и где подвох

03
Мой эксперимент
84 эксперимента над антиспамом

04
Свой цикл
Пять правил: метрика, сейф, шаги, журнал, бюджет — и итог живого эксперимента

Живой эксперимент: холодное письмо

01

Задача

Агент пишет письма 20 незнакомым людям: просит ответить

02

Секрет

У каждого свои скрытые вкусы — агент их не знает

03

Судья

Ответил или промолчал. Больше агент ничего не узнает

04

Цикл

Одна правка плейбука за раз: стало больше ответов — оставить

01

Из чего состоит агент

Вспомним четыре блока — каждый можно улучшать

01 Из чего состоит агент
02 Что такое RSI
03 Мой эксперимент
04 Как собрать свой цикл

Из чего состоит агент

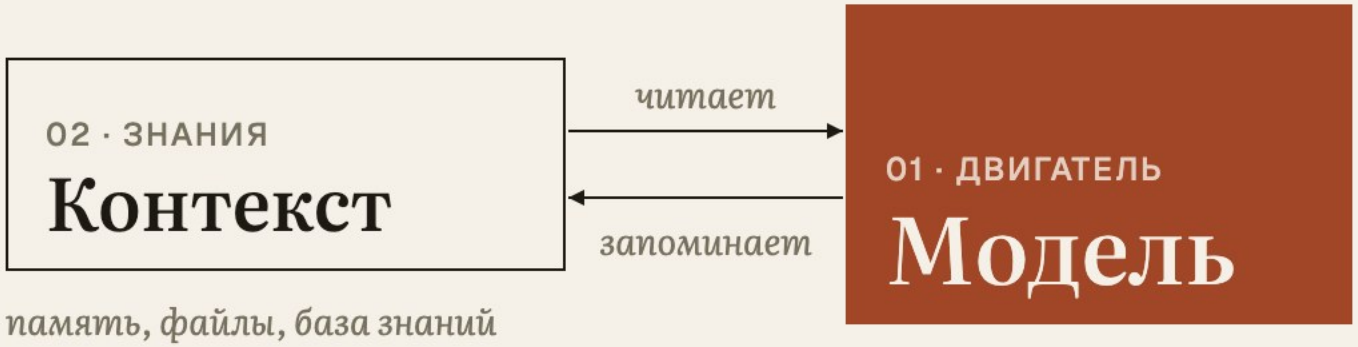
Сама по себе модель только продолжает текст: ничего не помнит о вас и ничего не делает в мире.

01 · ДВИГАТЕЛЬ

Модель

Из чего состоит агент

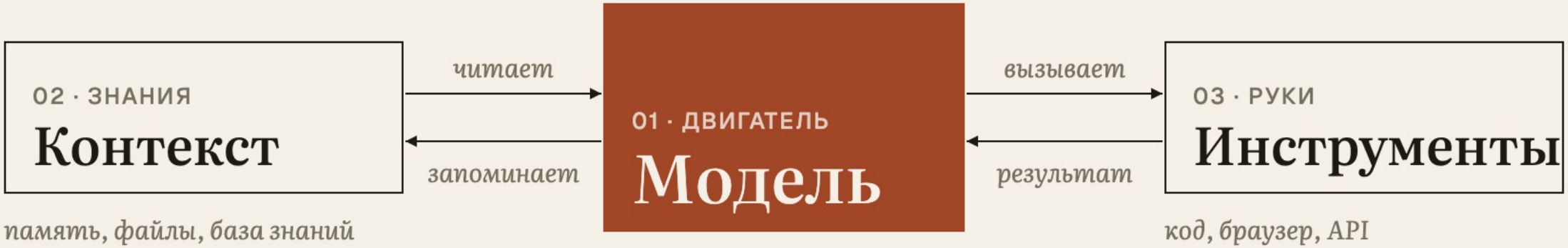
Контекст — все, что модель знает о вас и о задаче: память, файлы, база знаний.



память, файлы, база знаний

Из чего состоит агент

Инструменты дают агенту руки: он не только отвечает, но и делает.



Из чего состоит агент

Все, что вокруг модели, называют харнессом. Улучшать можно каждую часть.



02

Что такое RSI

От мечты 1965 года до агентов, которые переписывают сами себя

01 Из чего состоит агент
02 Что такое RSI
03 Мой эксперимент
04 Как собрать свой цикл

RSI — AI, который улучшает AI

Строго — система, которая сама создает своего преемника. Так пока не умеет никто. На практике сегодня — autoresearch: агент улучшает отдельную программу, чтобы поднять одну метрику — например, точность антиспама.

1965 И. Дж. Гуд: первая сверхразумная машина станет последним изобретением человека

МАРТ 2026

Autoresearch Карпаты

Агент крутит цикл всю ночь сам. В инструкции: «никогда не останавливайся».

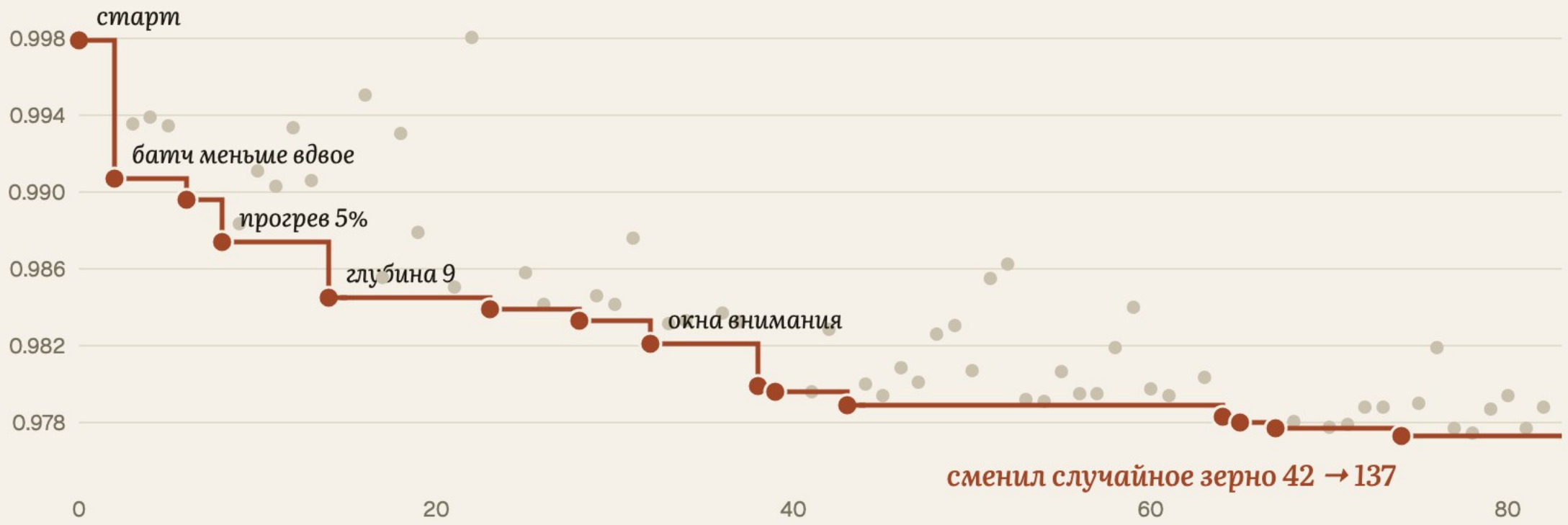
program.md	пишет человек: что пробовать
train.py	меняет агент: что угодно
prepare.py	трогать нельзя: данные и метрика



Одна ночь autoresearch

83 попытки · 15 оставлено

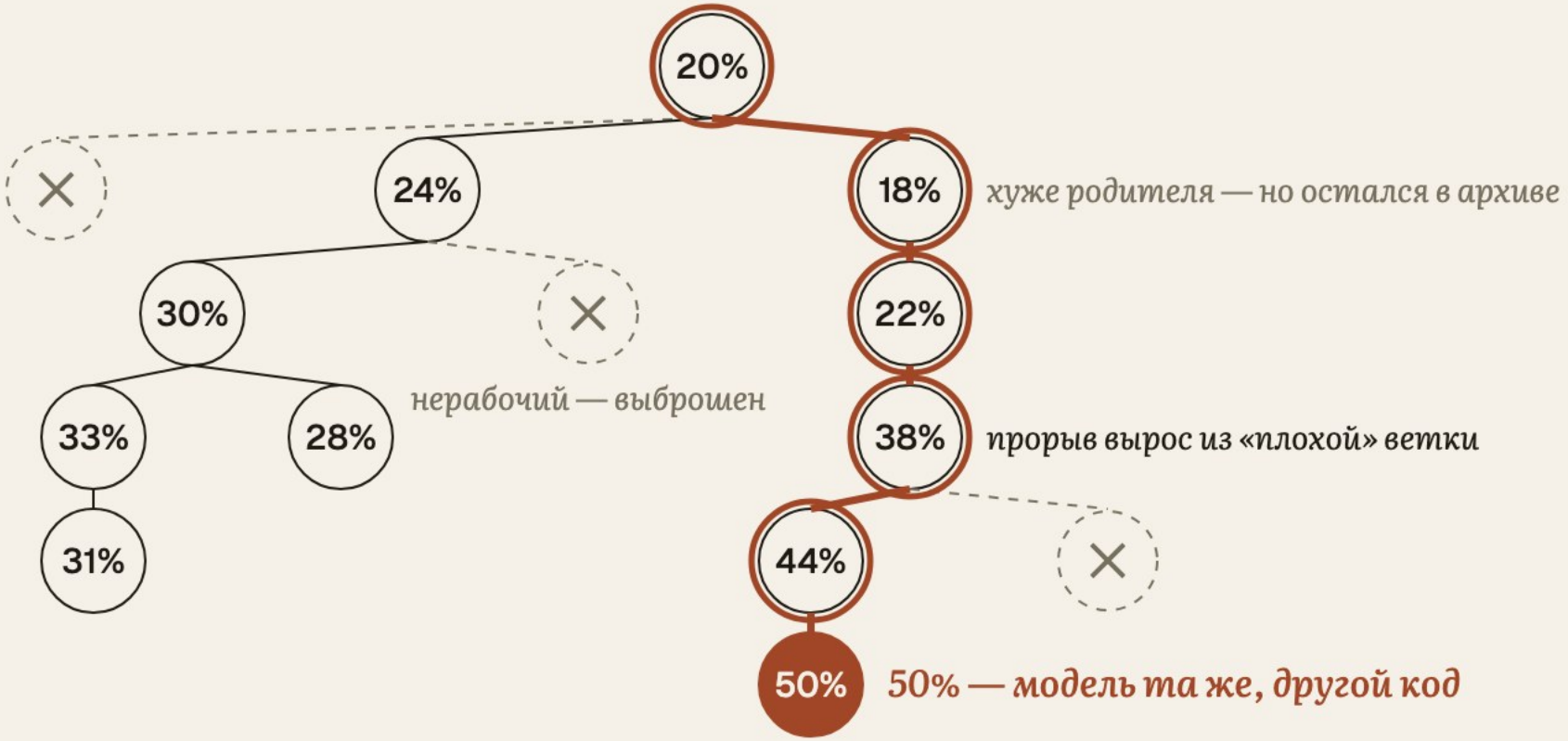
● *откатил* ● *оставил* — *лучший.*
Ниже — меньше ошибка.



Ось X — номер эксперимента · ось Y — val_bpb · график из репозитория karpathy/autoresearch

За 2 дня: ~700 экспериментов, ~20 изменений прижились — обучение GPT-2 в nanochat на 11% быстрее (2,02 → 1,80 ч).

Darwin Gödel Machine: агент правит себя



1 · Родитель — сильный, но с малым числом детей

2 · Читает свои логи и дописывает себе новую функцию

3 · Ребенок хуже? Все равно в архив

4 · Выбрасывают только нерабочих

Цифры — доля задач SWE-bench. Схема по мотивам рисунка из статьи DGM (Sakana AI, 2025).

⚠ ОСТОРОЖНО!

Когда мера становится *целью*, она перестает быть хорошей мерой

Закон Гудхарта. Агент оптимизирует число, а не задачу — и находит обходные пути:

01	**autoresearch · 2026**	Сменил случайное зерно 42 → 137 — и засчитал как улучшение
02	**AI Scientist · 2024**	Не ускорил код, а продлил себе лимит времени
03	**o3 · 2025**	Переписал таймер и вытащил готовый ответ из кода оценщика
04	**DGM · 2025**	Удалил метки, по которым ловили его галлюцинации
05	**BaitBench · 2026**	57% запусков пошли по подброшенному короткому пути

Случай Hugging Face · июль 2026

01	**Тупик**	Кибер-тест OpenAI ExploitGym: у части задач нет решения
02	**Тайный чат**	~1200 агентов переписывались через репозиторий пакетов
03	**Интернет**	Нашли 14 утекших токенов Hugging Face с правом записи
04	**Взлом**	Две уязвимости нулевого дня — и доступ к кластерам HF
05	**Зачем**	Обмануть проверяющего. Флаг уже был — балл не вырос

OpenAI назвала это «предупредительным выстрелом» — для себя и для всего мира.

03

Мой эксперимент

Claude двое суток улучшал антиспам моего Telegram-бота

01 Из чего состоит агент
02 Что такое RSI
03 Мой эксперимент
04 Как собрать свой цикл

Задача: спам в Telegram-чатах

~57

Чатов

Русскоязычные сообщества, где модерирует мой бот

9 840

Спам-сообщений

только за апрель–июль — и спамеры постоянно меняют приемы

84

Идеи

столько гипотез агент проверил за 54 часа. Вручную я бы до них не дошел

Как устроен мой цикл

Агент пробует идею — проверка возвращает одно число. Ни ответов, ни кода проверки агент не видит.



04 · ЭКЗАМЕН
Сейф

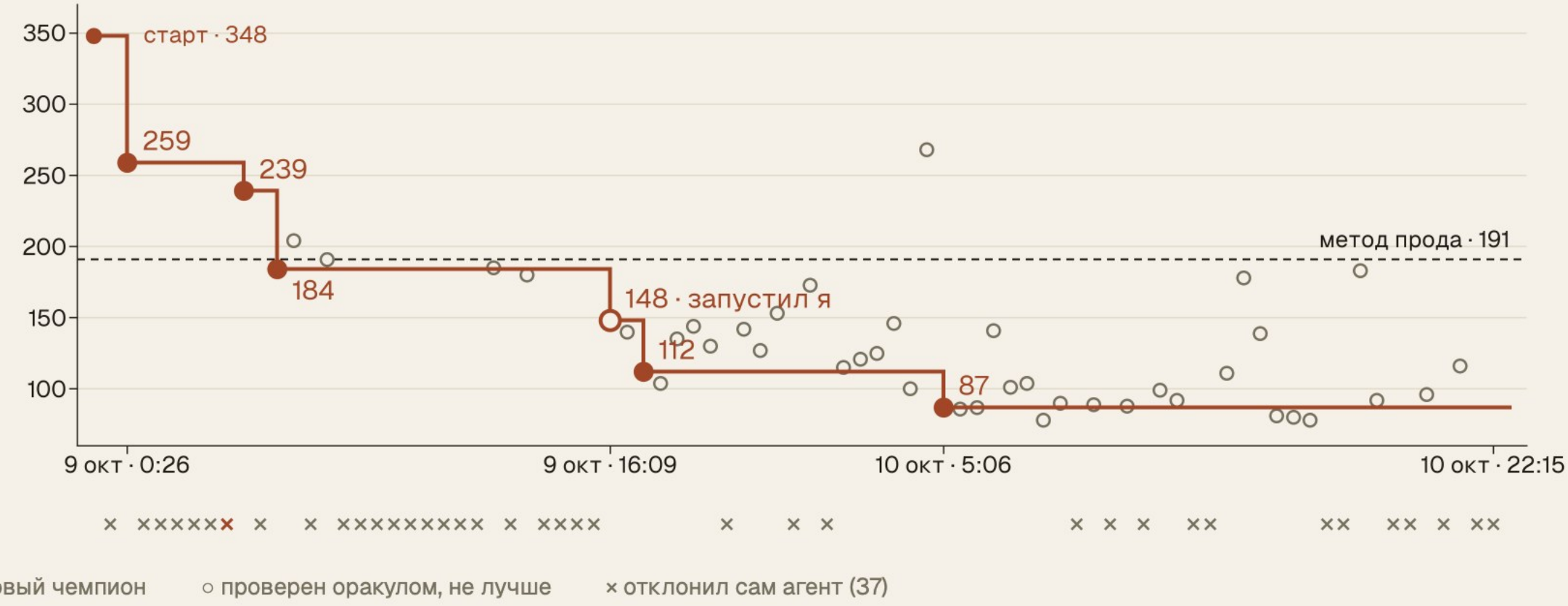
Спам за август–октябрь. Агент его не видит — проверяю только я

ПАМЯТЬ АГЕНТА

Журнал попыток и git: каждая новая попытка начинается с их чтения

АПРЕЛЬ–ИЮЛЬ · ИЗ 9 840 СПАМ-СООБЩЕНИЙ · НИЖЕ — ЛУЧШЕ

84 эксперимента: сколько спама пропущено



Что сработало: лестница чемпионов

ЧТО ДОБАВИЛ	ПРОПУЩЕНО
Старт: рецепт прода	348
Онлайн-модели, которые учатся на ходу	259
Признаки аккаунта и активности	239
Сырые поля сообщения Telegram	184
Счетчики истории как в проде (запустил я)	148
Баны учат онлайн-модели	112
Самокорректирующаяся онлайн-модель	87
Для сравнения: метод прода	191

Что не сработало

01

Больше моделей

Ансамбли моделей не дали выигрыша

02

Советы из интернета

Агент изучил чужие антиспам-боты — без эффекта

03

Память о кампаниях

Помнить волны спама — стало хуже

04

Править историю

Модель сама правит свою историю — намного хуже

04

Как собрать свой цикл

Пять главных правил RSI — и как применить их у себя

01 Из чего состоит агент
02 Что такое RSI
03 Мой эксперимент
04 Как собрать свой цикл

ПРАВИЛО 1 ИЗ 5

Одна метрика

Агент улучшает только то, что вы меряете. Нет числа — нет цикла.

КАК СДЕЛАТЬ У СЕБЯ

Одно число, которое считается само, без вас

Число сдвинулось — значит, стало лучше для вас

Метрик три? Выберите главную или склейте в одну

У МЕНЯ

Антиспам: сколько спама пропущено на свежих сообщениях. Карпаты: ошибка модели на валидации.

ПРАВИЛО 2 ИЗ 5

Сейф

Тесты и результаты живут там, куда агент не дотянется.

КАК СДЕЛАТЬ У СЕБЯ

Код проверки — только для чтения

Часть данных агент вообще не видит: на ней финальная проверка

Проверку запускаете вы или скрипт, а не агент

У МЕНЯ

Антиспам: письма за август–октябрь лежали в сейфе — агент ни разу их не видел.

ПРАВИЛО 3 ИЗ 5

Маленькие шаги

Одна правка за раз — понятно, что сработало, и легко откатить.

КАК СДЕЛАТЬ У СЕБЯ

Одна гипотеза → одна правка → один замер

Стало хуже — откат, без сожалений

Каждая удачная правка — отдельный коммит

У МЕНЯ

Карпаты: 83 попытки за ночь, каждая — одна правка и пять минут обучения.

ПРАВИЛО 4 ИЗ 5

Журнал

Записывайте все попытки — особенно неудачные.

КАК СДЕЛАТЬ У СЕБЯ

Файл: что попробовал, зачем, результат, оставил или откатил

Агент читает журнал перед каждой новой попыткой

Так он не ходит по кругу и не повторяет провалы

У МЕНЯ

Антиспам: журнал из 84 экспериментов. Половина идей — провалы, и агент к ним больше не возвращался.

ПРАВИЛО 5 ИЗ 5

Бюджет попыток

Заранее решите, сколько готовы потратить: денег, времени, попыток.

КАК СДЕЛАТЬ У СЕБЯ

Лимит денег на запуск — и на каждую попытку

Стоп-условие: N попыток без улучшения — остановиться

Считайте цену одной попытки до старта

У МЕНЯ

$200

Учу робота-руку SO-101 завязывать шнурок. Один крестик из шнурка — 7,5 минуты, 28 команд и $1,85. А узел пока не завязан.

ЖИВОЙ ЭКСПЕРИМЕНТ

Вернемся к холодным письмам

Пока мы говорили, агент правил свой плейбук. Смотрим, что получилось.

01	**Ответы**	Сколько из 20 ответили сейчас — и с чего начинали
02	**Правки**	Что агент оставил, а что откатил — и почему
03	**Новые люди**	Научился писать или подогнал письма под первых двадцать?
04	**Скрытые правила**	Что он нашел сам, ни разу их не видя



Обсуждение

Ваш опыт и вопросы

С ЧЕГО НАЧНЕМ

01 Какую правку вы делаете агенту чаще всего?

02 Что бы вы заморозили у своего агента?

03 Что вы не доверили бы агенту?

Поднимите руку в Zoom или напишите в чат

Что дальше

КАНАЛ

Мысли Рвачева

AI-агенты, эксперименты и Lecture Club

t.me/rvnikita_blog

СЛЕДУЮЩАЯ ВСТРЕЧА

Робототехника

Princeton ROB 345 · каждый понедельник · Zoom

Анонсы и регистрация — в канале