

HEISENBUG · 2026

01 / 31



МИХАИЛ МУРУНОВ · SCALA-ИНЖЕНЕР · GENPROOF.COM · FORMALIZED.SI · EX T-BANK

Баги, которые можно **доказать**: как агенты ищут ошибки в любых программах

9 багов в компиляторе Scala 3 · 302 репорта в 90 open-source проектах · 135 уже исправлены

ИИ научился находить настоящие баги

PROJECT GLASSWING
 •
 OPENSSE
 •
 DARPA AIXCC

27 лет

возраст бага в OpenBSD, найденного Claude Mythos Preview

> 10 000

high/critical-уязвимостей нашли Anthropic и ~50 партнеров к маю 2026

12 из 12

CVE OpenSSL в январе 2026 нашла AI-компания AISLE

54 из 63

внедренных уязвимостей нашли системы DARPA AIXCC, финал 2025

«Now it's limited by how quickly we can verify, disclose, and patch» — Anthropic, май 2026

Anthropic, янв 2026: PBT-агент на Hypothesis — 984 репорта в 100+ Python-пакетах, 56% валидны

...и завалил мейнтейнеров slop'ом

май 2025	curl	«We are effectively being DDoSed» — за AI slop банят сразу
январь 2026	curl	bug bounty закрыт: реальных уязвимостей среди репортов меньше 5%
апрель 2026	curl	«The slop situation is not a problem anymore» — репортов вдвое больше
май 2026	Linux	Торвальдс: из-за AI-репортов security-лист «почти неуправляемый»
октябрь 2026	Google	OSS VRP на паузе: «the vast majority of which are not valid»

«...we'll be in a kind of new golden era of discovery that [is] a kind of new renaissance»

Демис Хассабис · Fortune · 11.02.2026

Мы хотим ускорить этот прогресс.
Открыто, для всех и без границ.



Корректный код, битый байткод

```
def mixed: (erased Proof) ?=> Int ?=> String =  
  s"value: ${summon[Int]}"
```

—	scalac	→ компилируется, без предупреждений	—
	java	→ VerifyError: Bad local variable type	

-Ycheck:all проходит. Ошибку находит только JVM.

Отмена подтверждена. Команда ушла.

01

Запуск

Директива ждет в очереди с приоритетом 6

02

Отмена

CANCEL с приоритетом 8 обрабатывается первым

03

Простой

Секвенсор отвечает: последовательность отменена

04

Команда ушла

Директива из очереди выполняется без проверки состояния

Те же замеры — p95 ниже

LOCUST — OPEN-SOURCE ИНСТРУМЕНТ ДЛЯ НАГРУЗОЧНОГО ТЕСТИРОВАНИЯ НА PYTHON

100 ИЗМЕРЕННЫХ ЗАПРОСОВ

+ 1900 ЗАПРОСОВ БЕЗ ВРЕМЕНИ

96 мс

1 мс

Тест зеленый, данные неверные

KARATE — OPEN-SOURCE ФРЕЙМВОРК ДЛЯ АВТОТЕСТОВ API

ПРОВЕРКА	ФАКТИЧЕСКИЕ ДАННЫЕ	ДО v2.1.2	ДОЛЖНО БЫТЬ
contains only deep [admin, user, user]	[admin, admin, user]	PASS	FAIL
#!^part), part = {a: 1}	[{a: 1, b: 2}]	PASS	FAIL
== 9007199254740992	9007199254740993	PASS	FAIL



Selenium



Ajv



Appium



Cypress

goja

Goja



Gatling



k6

JsonPath

JsonPath



Karate



Locust



Playwright

REST
Assured

REST Assured



Robot Framework



Bruno



WireMock

Михаил Мурунов

GENPROOF.COM · FORMALIZED.SI · EX T-BANK

formalized.si — независимый некоммерческий проект
Исследуем ИИ на задачах, где ответ можно строго проверить.

Есть принятые баги в **FFmpeg**, **GitLab** и в разных протоколах децентрализованных финансов.



FORMALIZED.SI



Что уже сделали публично

302

репортов в 90 open-source проектах; 135 багов уже исправлены

236 → 71

блоков в сумматоре редстоуна; верность доказана для всех входов

891 / 917

блоков книги Symmetry формализованы в Narya; 16 утверждений книги опровергнуты

Топ-4

место в лидерборде lean-eval от Lean FRO. В нем 85 задач

Вы не айтишник, но у вас есть здравая идея из фундаментальной науки?
Приходите — сделаем вместе.

Контекст из прошлого

Четыре подхода к генерации входов

01 • LLM

Нет валидности

Агент пишет программы, большая часть умирает на парсере

02 • ФАЗЗИНГ

3,7% и 0 крашей

25,5 млн мутаций на TON: парсер прошли 948 тыс., крашей ядра — ноль

03 • PBT

Дорогой генератор

Генератор на месяцы, и его самого надо тестировать

04 • ТИПЫ

Генератор из типа

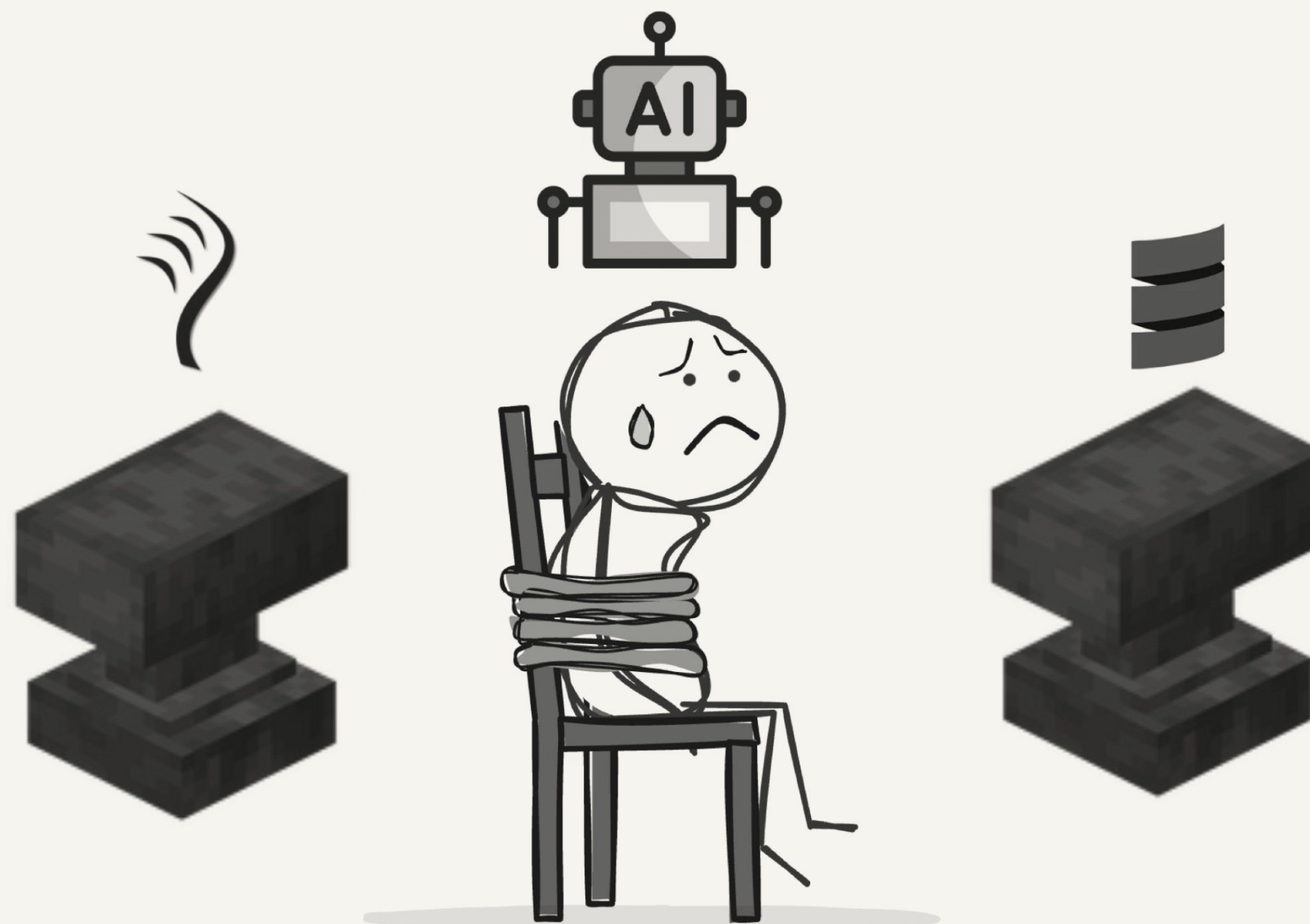
Idris 2 + DepTyCheck: невалидную программу нельзя построить

10 000 000 / 10 000 000 сгенерированных программ проходят scalac • генерация на типах, не на LLM

Хорошо протестировать нетестируемое и не сойти с ума

heisenbug.ru/archive/2024_Autumn/talks/20005186-test-untestable-well-still-staying-sane





Что это дало на Scala

9

issue в компиляторе Scala 3,
6 уже закрыты как
исправленные

204

тикета в 53 репозиториях
Scala-экосистемы

82

приняты и починены
мейнтейнерами (срез
05.06.2026)

7%

false positive: 14 из 204, сняли
сами с извинениями

От задачи — к гипотезам

Сначала понять объект, затем выбрать, что проверять.



Общая файловая система

задания · контекст · отчёты · результаты

— Дальше Striker берёт из очереди одну гипотезу.

Одна гипотеза — один эксперимент

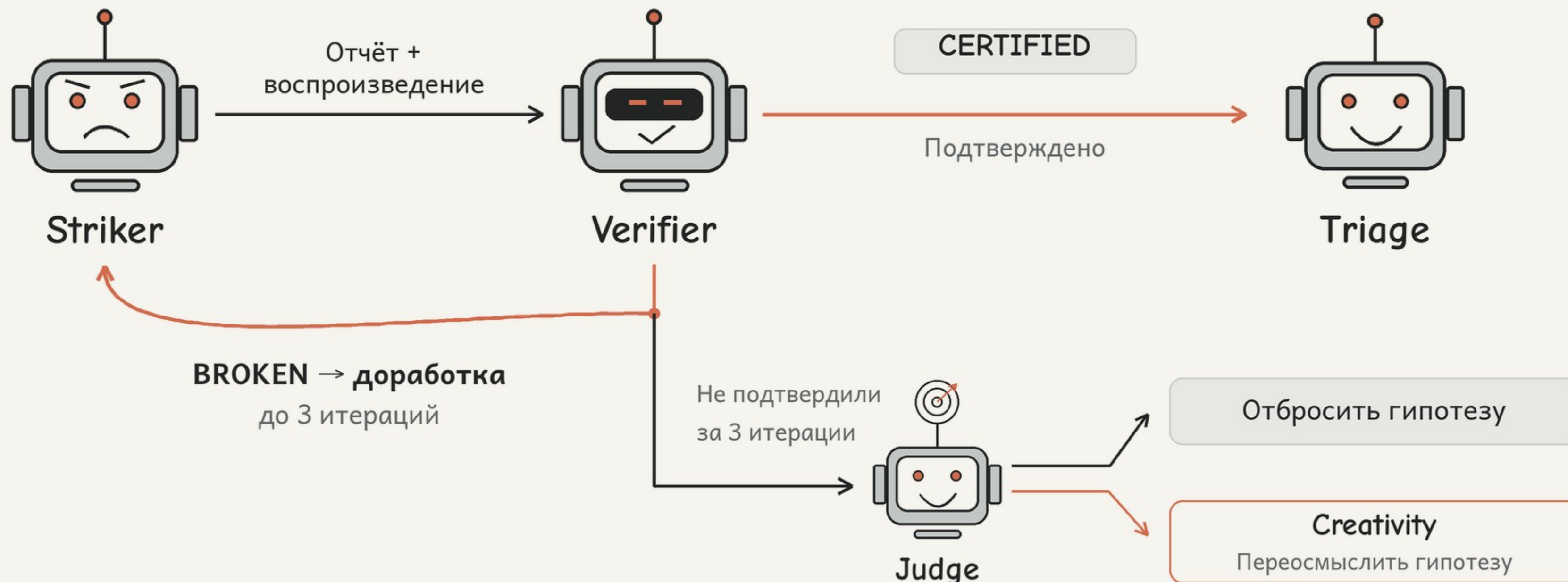
Striker проверяет идею и сохраняет то, что удалось воспроизвести.



— Это ещё кандидат. Подтверждение — следующий этап.

Находка — ещё не подтверждение

Verifier проверяет отчёт и воспроизведение; затем выдаёт статус.



BROKEN: подтверждение не прошло. Это не означает, что уязвимости нет.

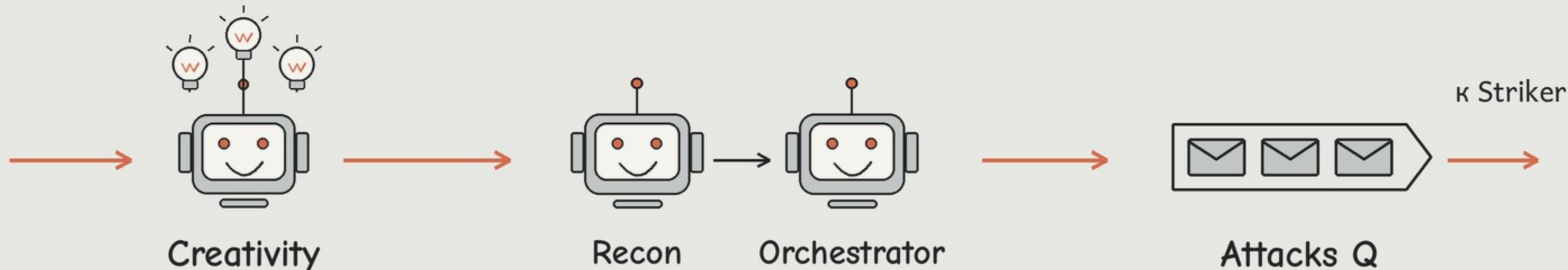
Из находки — в следующий шаг

Triage оценивает важность и помогает понять, куда двигаться дальше.



СЛЕДУЮЩИЙ ЦИКЛ

Новые направления
Пустая очередь
Решение Judge



— Задания, отчёты и результаты — в общей файловой системе.

Тикеты читают живые люди

297

репортов разобрали по реакциям, срез 06.10.2026

48%

issue исправлены: 133 из 279, медиана — 6 дней

32%

полная тишина: ни комментария, ни метки, ни закрытия

8

враждебных реакций в 4 проектах, из них 3 бана

ЗА

«Sounds like a proper bug! Can you make a PR, adding the test case as well as the fix?»

cyberw · Locust · PR смержили в тот же день

КОПАЕТ ВМЕСТЕ

«So, the bug is the mismatch between the type and the tree after erasure»

tanishiking · Scala 3 · фикс через день

PRS WELCOME

«Thanks for the report - PRs welcome!» × 12 за 32 секунды

adamw · sttp · потом сам починил 10 из 12 и спросил, как мы их находим

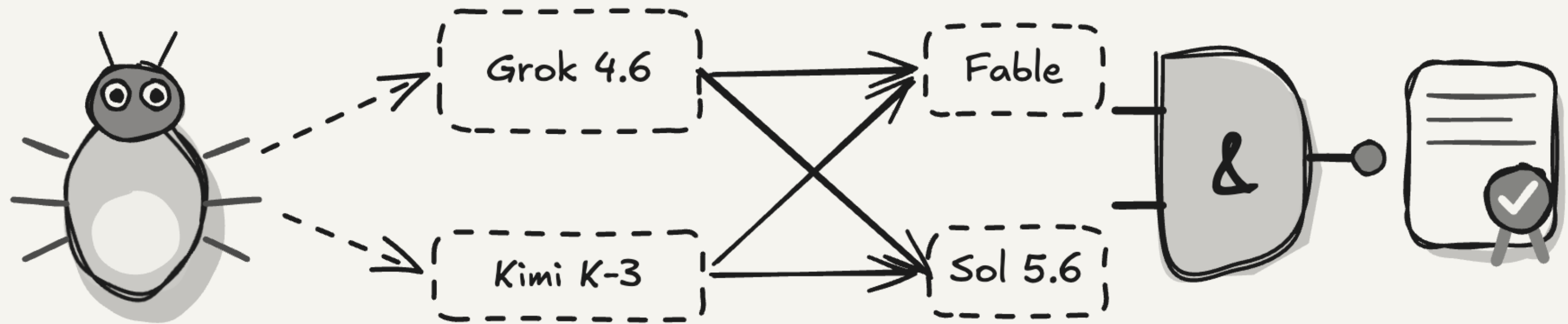
ПРОТИВ

«Nonsense AI-generated issue. Banned from project.»

AugustNagro · magnum · из 3 банов один сняли после письма мейнтейнеру

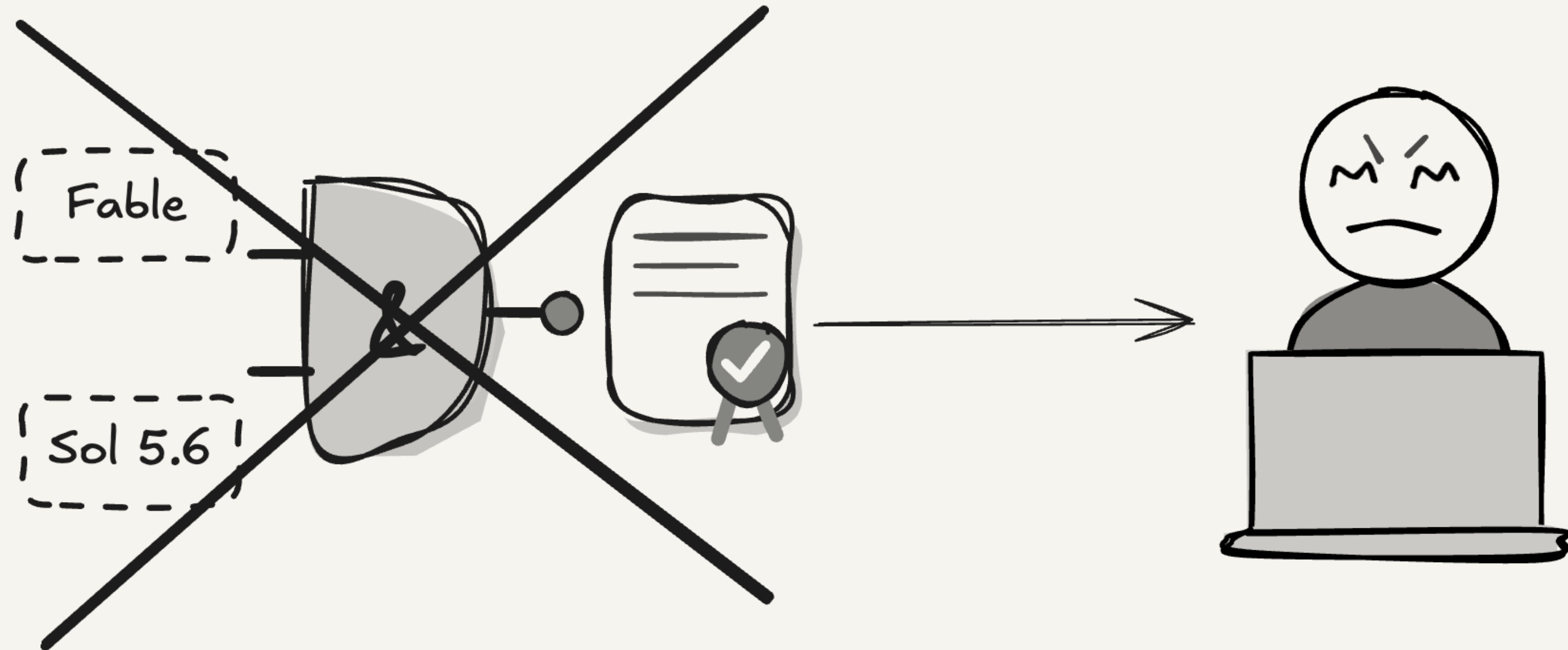
Жалуются на slor-флуд, старую версию и многословие → ≤ 3 тикета на проект, свежая версия, коротко, с патчем

Provider-based model tunnel vision



Модели одной семьи разделяют слепые зоны — репорт сертифицируется, только если баг независимо подтверждают модели разных провайдеров

Если модели не согласны — призывается человек



Расхождение оракулов — не провал, а сигнал: такой репорт не сертифицируется автоматически, его разбирает человек

Откуда берутся лиды

01



Оси спецификации

Модель берет из спеки 3–12 осей. У каждой оси два полюса.

ПРИМЕР

Кофемашина. Оси: «крепость» и «объем». Проверяем угол: очень крепко и очень много.

02



Механизмы

Кэш, fast path, ленивый агрегат. Быстрый ответ = ответ с нуля.

ПРИМЕР

Лифт помнит последний этаж. С памятью и без памяти он едет на тот же этаж.

03



Кузены бага

Свежий фикс показывает шаблон. Ищем соседей без фикса.

ПРИМЕР

В одной двери починили замок. Проверяем все двери с этого завода.

04



Известные issue

Issue — это семена, а не находки. Метки показывают горячие зоны.

ПРИМЕР

Много жалоб на кухню. Ищем на кухне, а не в спальне.

05



Соседние режимы

deer и обычный, X и not X. Два входа в один API дают один ответ.

ПРИМЕР

Пицца с сыром и пицца без сыра. Разница только в сыре.

Лид = обещание инструмента + способ его опровергнуть. Без path:line лид не стоит ничего.

Семантический поиск похожих лидов

gbrain: PGLite + pgvector · эмбединги nomic-embed-text (768) · индекс: исходники цели и референсов

01 · ПОИСК

Где еще живет механизм

2–3 запроса в gbrain на итерацию: тот же шаблон в соседних модулях и в других реализациях

02 · СКОРИНГ

+2 к лиду

если лид проходит через 2 и больше проиндексированных источника

03 · ДЕДУП

До тикета

локальный индекс багов → поиск по словам → семантика по issue трекера

04 · БАНАЛЬНОСТЬ

cos > 0,50 — в мусор

идея слишком близка к центроиду очевидных: generic ≈ 0,60, отобранные 0,28–0,37

Персоны: 10 линз, чтобы думать иначе

01	Конкретизация	только вещь ↔ вещь, без категорий
02	Слабый признак	группируем по несущественному свойству
03	Дрейф цепочки	пять прыжков, потом взгляд назад с пятого
04	Персеверация	прошлую рамку — на следующую вещь
05	Разноплановость	три плана, которые не сводятся

06	Обходной путь	назвать привычное решение и запретить его
07	Детский вопрос	«а почему вообще...?»
08	Палка-лошадь	заменить носитель главного признака
09	Опредмечивание	увидеть вещь в аморфном
10	Режим пятилетнего	вещь по ситуации, а не по назначению

2–3 линзы на застой · вес по приживаемости, пол 0,15 · агент на линзу → 3–5 зародышей

От зародыша к тикету

01

Линзы

3–5 недоделанных зародышей на агента

02

Кристаллизатор

режет резонерство, пишет гипотезы

03

Скоринг

+3 горячая ось, –5 исчерпано; 3 из 10 — в непокрытое

04

Свойство

генератор и оракул; без path:line лид не стоит ничего

05

Сертификат

репро, негативный контроль, фикс на копии

06

Гейт

разные провайдеры + человек; ≤ 3 тикета на проект

TON: 40 вспышек линз → 101 зародыш → 162 гипотезы → 21 подтвержденная находка · все 10 линз дали хотя бы одну

Работает ли это

~1000×

больше расхождений на вход: 3 линзы → 35 атак → 13 расхождений (TON)

21 / 162

гипотез стали подтвержденными находками (TON)

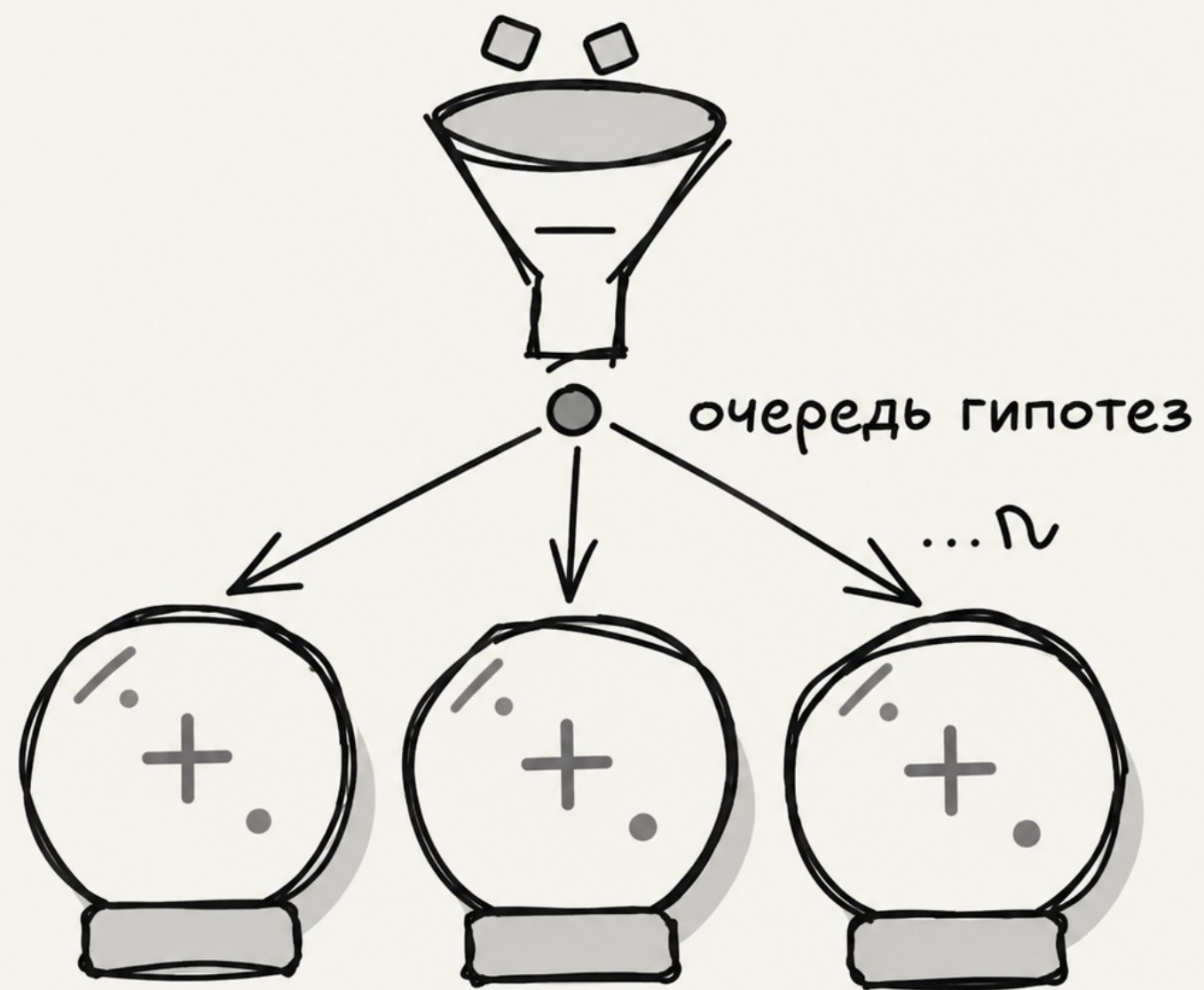
79 / 107

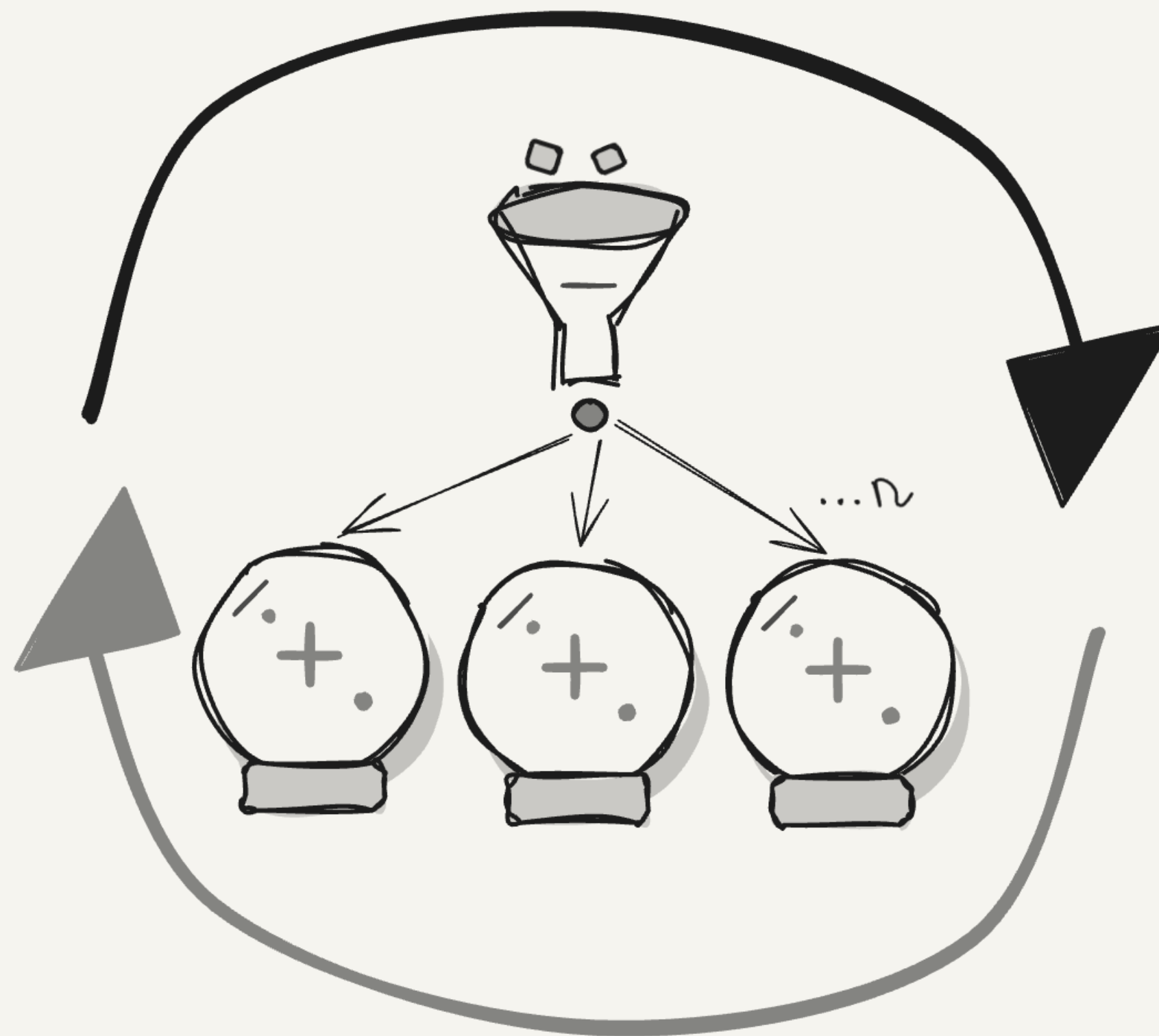
кандидатов QA-кампании прошли гейт → 36 тикетов в 15 проектах

0 / 768

агентов лобового перебора «линза × файл» — ни одной находки

Линзы без доказательств не работают. Каждый лид должен принести path:line и способ себя опровергнуть.





Агенты делают подконтрольную работу

РАСТЯТ МОДЕЛЬ

Одна правка, один файл

ДЕДУПЯТ

Двадцать клонов одного класса — один тикет

МИНИМИЗИРУЮТ

Текстовая дельта, потому что шринкера нет

НЕ РЕШАЮТ

Валиден ли сэмпл. Это решает тип

Итоги в цифрах

302

репорта в 90 open-source проектах

135

багов уже исправлены мейнтейнерами

4 из 4

флагманских бага подтверждены: Scala 3, NASA F', Locust, Karate

1 из 3

банов сняли, когда мы написали мейнтейнеру

Выводы

- 01 Зеленый тест — тоже утверждение. Его нужно проверять.
- 02 Валидность — в генератор, не в фильтр: корректность входа держат типы.
- 03 Лиды масштабируются через оси, кузенов и линзы, но каждый приносит path:line и способ опровергнуть.
- 04 Чем больше независимых валидаторов, тем лучше: разные провайдеры, движки и человек in the loop.
- 05 Неожиданно: агенты лучше рассуждают, когда спецификация абстрактна и выразительна. Возможно, декларативна.



Anthropic Research — Property-based testing

anthropic.com/research/property-based-testing

Шпаргалка для вашего агента

Цель: найти баг, который можно доказать. Сфотографируйте слайд и дайте его агенту.



ФОРМУЛЫ

- лид = обещание + способ опровергнуть
- баг = репро + негативный контроль
- 2 бага + 1 фикс = 1 баг
- гейт = модель A и модель B
- быстрый ответ == ответ с нуля
- оракул: без фикса падает, с фиксом нет

ТИКЕТ · ЧЕК-ЛИСТ

- ✓ последняя версия
- ✓ минимальный пример
- ✓ тест падает без фикса
- ✓ короткий текст
- ✓ патч, если можешь
- ✓ не больше 3 тикетов на проект

НЕ ДЕЛАЙ

- × LLM генерирует входы → нет валидности
- × перебор «линза × файл» → 0 / 768
- × одна семья моделей → общие слепые зоны
- × 15 тикетов за вечер → бан
- × репорт со старой версии → дубль

ССЫЛКИ

30 / 31

formalized

Спасибо

formalized.si · genproof.com

github.com/haskiindahouse · X [@mike_genproof](https://twitter.com/mike_genproof)



github.com/haskiindahouse

204 бага по природе дефекта



~55% — тихие баги без краша ~25% — краши ~15% — low security-находки конкурентность 4% — слепая зона метода