



FABLE 5 · СУТКИ СПУСТЯ

Сутки с Claude Fable 5

Тихий нерф, отказы на обычных задачах и проверка на себе. Что системная карта прячет от анонса и что с этим делать бизнесу.



@gorilla_under_hood

Вчера я разобрал релиз Claude Fable 5, первой публичной модели топового класса Mythos. Бенчмарки взрывные, цена вдвое выше Orus, доступ в подписке до 22 июня. За сутки картина усложнилась.

В системной карте нашлась деталь, которой нет в анонсе. Сообщество разделилось на два лагеря. А сегодня Fable отказался работать посреди нашей собственной, совершенно обычной задачи, прямо на открытой лекции. Собрал всё в один разбор: факты, спектр мнений, личный опыт и выводы для бизнеса.

В ЭТОМ РАЗБОРЕ

- **Деталь, которой нет в анонсе**
- **Нерф измерим: цифры**
- **Спектр мнений: четыре позиции**
- **Проверка на себе: кейс на лекции**
- **За что Fable ставит на паузу**
- **Штрих: заморозки аккаунтов**
- **Позиция: четвёртое измерение риска**
- **Что делать бизнесу: 5 шагов**

Деталь, которой нет в анонсе

В системной карте Fable 5 описаны два контура защиты. Первый прозрачный. Второй работает иначе, и про него в анонсе ни слова.

КОНТУР 1 · ПРОЗРАЧНЫЙ

Биология, химия, кибербез,
дистилляция

Запрос уходит на модель попроще,
Opus 4.8

Пользователь это видит

КОНТУР 2 · СКРЫТЫЙ

«Разработка фронтальных LLM» по
мнению фильтра

Ответ намеренно хуже:
переписывание промпта, steering-
векторы, точно в веса

Пользователю не сообщают

Ключевое про второй контур: ответ приходит обычный на вид, просто хуже, а пользователю об этом не сообщают. Технически это переписывание промпта, steering-векторы и точечное вмешательство в веса. Anthropic оценивает охват меры в 0.03% трафика.

Нерф измерим

Это не теоретическая страшилка. На бенчмарке PostTrainBench, где модель пишет код для дообучения других моделей, Fable 5 показывает результат на уровне Sonnet 4.6, средней модели линейки.

НЕРФ ИЗМЕРИМ, НЕ СТРАШИЛКА

0.03%

трафика под мерой
(оценка Anthropic)

→ 4.6

на PostTrainBench
Fable
падает до уровня
Sonnet 4.6

×2

цена выхода против
Opus
\$50 за миллион
токенов

На остальных задачах Fable бьёт всё, что Anthropic выпускала. Деградация реальна, велика и воспроизводима, просто включается избирательно.

То есть деградация под защитой реальна, велика и воспроизводима. Просто включается она избирательно, и заранее вы не знаете, попал ваш запрос под неё или нет.

Что говорят первые сутки

Реакция раскладывается на четыре позиции. Я свожу их без имён и каналов, только суть аргумента.

ПЕРВЫЕ СУТКИ · ЧЕТЫРЕ ПОЗИЦИИ

1 Критики

Модель тихо глупеет без уведомления и ломает саму возможность доверять инструменту. Часть исследователей уже ушла с Fable.

2 Защитники

Использование Claude для конкурирующих моделей и так нарушает ToS. Скрытость не даёт нарушителю переформулировать запрос. Механизм описан в карте добровольно.

3 Прагматики

Модель сильная, но «аудит системы» и проверка проекта на уязвимости роняют её в отказ или Orus. Это повседневная работа любой IT-команды, не только тех, кто учит модели.

4 Про данные

Все запросы и ответы хранятся минимум 30 дней и просматриваются. Обучать без разрешения обещают не на них. Но то, что вы загружаете, будут читать.

Проверка на себе, вживую на лекции

Сегодня мы наткнулись на это сами, и не в лаборатории. Это было на открытой лекции про ИИ, в прямом эфире перед аудиторией. Задача наша учебная, мы прогоняем её из раза в раз на занятиях и для себя с друзьями: конкурентный анализ по открытым источникам. Стандартное задание аналитика, найти сильные стороны, слабые места, оценить, где предложение уязвимо. Никакой разработки LLM, никакого взлома, это упражнение мы делали десятки раз.

И ровно посреди демонстрации Fable собрал данные, а на шаге анализа встал. Приложение предложило перейти на Opus, модель попроще.

Chat paused

Fable has safety measures that flag messages on most cybersecurity or biology topics. They may flag safe, normal content as well. These measures let us bring you Mythos-level capability in other areas sooner, and we're working to refine them. Continue with Opus 4.8, [send feedback](#), or [learn more](#).

Так выглядит отказ: приложение ставит чат на паузу и предлагает Opus 4.8. Ровно это мы увидели посреди обычного анализа.

Потом мы разобрали логи. Самое примечательное: дословной формулировки отказа в них нет, переключение происходит на уровне приложения и для самой модели прозрачно. Наиболее правдоподобная реконструкция: слова «слабые места», «точки атаки», «уязвимость предложения» модель прочитала как намерение навредить чужому бизнесу. Хотя это словарь любого учебника по маркетингу, и упражнение это совершенно стандартное. Мы делали его десятки раз, а сегодня модель встала, и объяснить аудитории, почему, за секунду не вышло.

За что Fable ставит на паузу

Граница «чего модель не станет делать» проходит не там, где написано в анонсе. Её формулирует фильтр, а не здравый смысл, и заранее её не видно. Узнаёшь посреди задачи, в нашем случае посреди лекции.

ПОДТВЕРЖДЕНО · ВИДНО В ПРИЛОЖЕНИИ



Кибербезопасность

сюда же аудиты и проверки на уязвимости



Биология и химия

прозрачный контур: понижение до Orus 4.8



Разработка и дистилляция LLM

скрытый контур: ответ молча хуже

НАБЛЮДАЕТСЯ · НИЖЕ ДОСТОВЕРНОСТЬ



Конкурентный анализ

«слабые места», «точки атаки» как красный флаг. Наш кейс



Сбор и скрейпинг данных

массовый сбор по источникам



Здоровье и медицина

смежно с биологией, в карте не названо

Фильтр реагирует на лексику, а не на намерение: «уязвимость», «атака», «аудит», «сбор данных» это рабочий словарь маркетинга, безопасности и аналитики. Поэтому под раздачу попадает обычная работа. Зелёное подтверждено картой и видно в интерфейсе, янтарное собрано из тестов и нашего опыта, это гипотезы, а не официальный список.

Штрих к картине: заморозки аккаунтов

В тот же день нам прислали скриншот: аккаунт Claude уходит в заморозку с формулировкой «unusual activity». Восстановление через ручное ревью, которое занимает около 10 дней.



Your account is on hold

We put your account on hold on **Jun 10, 2026 at 3:22 PM** because of unusual activity. Your chats and data are safe.

If you think this hold is an error, you can request an account review.

What happens next:

- 1 Request a review**
Tell us more about what happened.
- 2 We'll review your account**
A team member will review your request and account activity together.
- 3 We'll email you the outcome**
Reviews take about 10 days.

[Request a review](#)

See our [Terms of Service](#) and [Usage Policy](#) for what Claude can be used for.

Скриншот аккаунт-холда: «unusual activity», ревью около 10 дней. Личных данных на нём нет.

ШТРИХ · ЗАМОРОЗКА АККАУНТА



«Unusual activity»

Аккаунт уходит в заморозку. Восстановление через ручное ревью.



Дней ревью

До десяти дней без инструмента, на который завязаны процессы.

Связь с Fable не доказана и по одному скрину волну не установить. Но политика усиленного просмотра запросов Mythos-класса делает ложные срабатывания на уровне аккаунта вероятнее. Это риск с конкретной ценой в днях.

Десять дней ревью стоит прочитать глазами бизнеса: это до десяти дней без инструмента, на который завязаны процессы. Не как прогноз, а как риск, у которого теперь есть конкретная цена в днях.

Заморозка и пауза Fable это разные системы, но направление у них общее: контроль за пользователем плотнее, а значит и ложных срабатываний больше. Вот за что обычно прилетает именно заморозка.

ОТДЕЛЬНАЯ ОСЬ · ЗА ЧТО ПРИЛЕТАЕТ ЗАМОРОЗКА АККАУНТА

- **Платёжка и IP из разных стран**
- **VPN скачет между странами внутри сессии**
- **Сторонние обёртки и скрипты поверх подписки**
- **Холодный аккаунт сразу в тяжёлый продакшн**
- **Апгрейд тарифа, новая карта и устройство разом**
- **Запрос ручной верификации личности**

Это не про контент, а про доступ. Заморозка аккаунта и пауза Fable это две разные системы. Связь конкретной заморозки с Fable не доказана, но усиленный просмотр запросов Mythos-класса делает ложные срабатывания на уровне аккаунта вероятнее. Список собран по нашему опыту и сообщениям пользователей, это ориентир, а не регламент.

Моя позиция

Скажу честно, и тут две честности сразу. Первая: Anthropic в этой истории я считаю честнее остальных. Механизм описан в системной карте, а не спрятан. Другие вендоры могут делать ровно то же самое и не писать об этом нигде. Проверить это снаружи нельзя было ни вчера, ни год назад.

Вторая честность: поймать блокировку посреди живой работы это нервно. Одно дело читать про 0.03% в спокойной обстановке. Другое дело, когда инструмент встаёт на стандартном упражнении, на глазах у аудитории, и ты за секунду не понимаешь, это твоя ошибка, сбой или политика вендора. Спокойно объяснить это «прозрачностью в системной карте» в моменте не выходит.

Но прецедент создан, и он важнее конкретной модели.

До сих пор риск AI-вендора бизнес считал в трёх измерениях: цена, лимиты, доступность. Эта неделя добавила четвёртое: поведение модели может измениться без уведомления. Раньше «инструмент работает сегодня так же, как вчера» было допущением по умолчанию. Теперь это гипотеза. А гипотезы проверяют.

Что делать бизнесу

Ответ инженерный, а не панический. Пять шагов, два из них можно сделать сегодня вечером.

ЧТО ДЕЛАТЬ БИЗНЕСУ · 5 ШАГОВ

- 01 Регрессионный набор.** 10 реальных задач компании с эталонными ответами: договор, отчёт, письмо, разбор таблицы.
- 02 Прогон после каждого обновления.** Сравнение занимает минуты и ловит деградацию за один прогон, а не через месяц испорченных результатов.
- 03 Запасной маршрут.** Вторая модель, на которую переключаетесь без переписывания процессов. Страхует и от деградации, и от потери доступа к аккаунту.
- 04 Чувствительные данные в Fable не грузить,** пока действует политика 30-дневного хранения с просмотром.
- 05 Сроки.** В подписке Fable до 22 июня, дальше только API: \$10 вход / \$50 выход за миллион токенов.

Качество AI-инструмента проверяется на ваших задачах, а не на бенчмарках вендора. На этой неделе у этого тезиса появился первый именованный прецедент. И, как выяснилось сегодня, проверка на себе занимает один рабочий день.

Разбираю AI для бизнеса без хайпа

Что реально работает с продуктами Anthropic, что произошло за неделю и на что смотреть дальше. Без инфоцыганства, с проверкой на своих задачах.



Подпишись: @gorilla_under_hood



наведи камеру