

AI-АГЕНТЫ В БИЗНЕСЕ

Ваш AI-агент не падает с ошибкой. Но уверенно нейрослопит

Где агент ломается незаметно, и три уровня защиты: обвязка, гейты, эвал.



Илья Утов · @gorilla_under_hood

AI-агент редко падает с критической ошибкой и встаёт. Чаще он уверенно выдаёт то, что выглядит правильно, а на деле черновик или нейрослоп. И вот эти тихие провалы дороже всего: на демке незаметно, а в реальной работе рушится весь смысл затеи с агентами.

Первое, на что кидаются: это всё модель, тупая попалась. Был бы рад, если бы так работало. Но лечится это не моделью подороже и поумнее, а качественной обвязкой вокруг агента плюс тотальная проверка и эвал. Короче, погнали разбираться.

В ЭТОМ ВЫПУСКЕ

- 01** Где агент ломается тихо · четыре места
- 02** Проверка: ловим ошибку поштучно · гейты
- 03** Эвал: ловим ошибку системно · свой набор задач

Уровень 1. Где агент ломается тихо

В бизнесе агент спотыкается не о мощность модели, а об обязательку вокруг неё. Четыре типичных места.

ЧЕТЫРЕ МЕСТА ТИХОГО ПРОВАЛА

- **Состояние** · решает на устаревших данных
- **План** · испаряется при сжатии контекста
- **Гигиена** · сливает чужие данные не туда
- **Диагностика** · лог больше контекстного окна

34,7

пункта теряет успех, если
забыть план

68%

утечка чужих данных в
стресс-тестах

4

бага вылезли только в
бою

ИЗ ПРАКТИКИ · КОННЕКТОР WB И OZON



Молча ноль строк

у каждого метода свой путь к данным в ответе



Пагинация тремя способами

агент знал только один



Параметр не туда

сервер тихо его игнорировал

Офлайн-тесты были зелёные. На тесте выглядит правильно, в бою совсем не то.

И ни одно из этого не чинится сменой модели. Всё про обязательку.

Уровень 2. Проверка: ловим ошибку поштучно

Те же тихие провалы лезут и в обычной работе человека с AI. Вот гейты, которые мы себе поставили. Честно, со своими шишками.

ГЕЙТЫ, КОТОРЫЕ МЫ СЕБЕ ПОСТАВИЛИ

- 1 «Прогнал» не значит «сделал»**
запуск родил черновик, не результат
- 2 «Выглядит правильно» не значит «верно»**
проверь данные под красивым выводом
- 3 Непроверенное число или факт, это черновик**
пересчитай сам, размечай доверие
- 4 Гейт на стыке, самый важный**
проверь прошлый шаг, реши вперёд или назад

Маленькая ошибка с шага два доезжает до шага шесть, где всплывает как «непонятно, почему всё развалилось». А чинить дорого: причина осталась четыре шага назад.

Поэтому на каждом стыке короткая проверка прошлого шага и явное решение: вперёд или назад на переделку. Особенно если шаг делал другой агент, его результат тоже черновик.

15-30%

на столько AI завышает свою уверенность

−15%

вычитаем из любой своей оценки по умолчанию

Уровень 3. Эвал: ловим ошибку системно

Ручные гейты хороши, но гонять их руками на каждое изменение дорого. Систематизированная версия называется скучным словом эвал. Десять-двадцать типичных задач, для которых вы заранее знаете правильный ответ. Поменяли модель или промпт, прогнали весь набор, посмотрели, на скольких сошлось.

0 / 8

цитат советника реально
из источника

10-20

своих задач достаточно
для эвала

7 / 10

вместо «вроде работает»

На нашем примере: делали AI-советника, который отвечает цитатами реального деятеля. На глаз убедительно. Прогнали эвал на одну вещь, правда ли каждая цитата есть в первоисточнике. Ноль из восьми. Красивые, но по атрибуции выдуманные. Глаз бы пропустил, эвал поймал тихую фабрикацию до того, как она ушла бы людям.

НА ГЛАЗ

Убедительно, гладко, в духе

Автор сам себя не проверит

Тихую фабрикацию пропустишь

«Агент вроде справился»

ЭВАЛ

Число, а не ощущение

Прогон при каждом изменении

Ловит то, что глаз пропустил

«18 из 20, упал на возвратах»

И не верьте бенчмаркам вендора: это их задачи, которые они сами и выбрали. Честный бенчмарк один, собранный на ваших кейсах.

КАК НАЧАТЬ ДЁШЕВО

- 01 Выпишите 10 задач, которые агент делает чаще всего
- 02 Для каждой запишите, каким должен быть правильный ответ
- 03 Прогоняйте весь набор при каждом изменении модели или промпта

Модель в бизнесе не узкое место. Узкое место это проверка. Кто строит её на каждом этапе, тот и выигрывает.

Под капотом · разбираем, что реально работает с AI

Не хайп про автономных сотрудников, а инженерия: где агент ломается тихо, какие гейты ставить и как мерить, что он правда работает. Без воды.



Подпишись: [@gorilla_under_hood](https://t.me/gorilla_under_hood)



наведи камеру