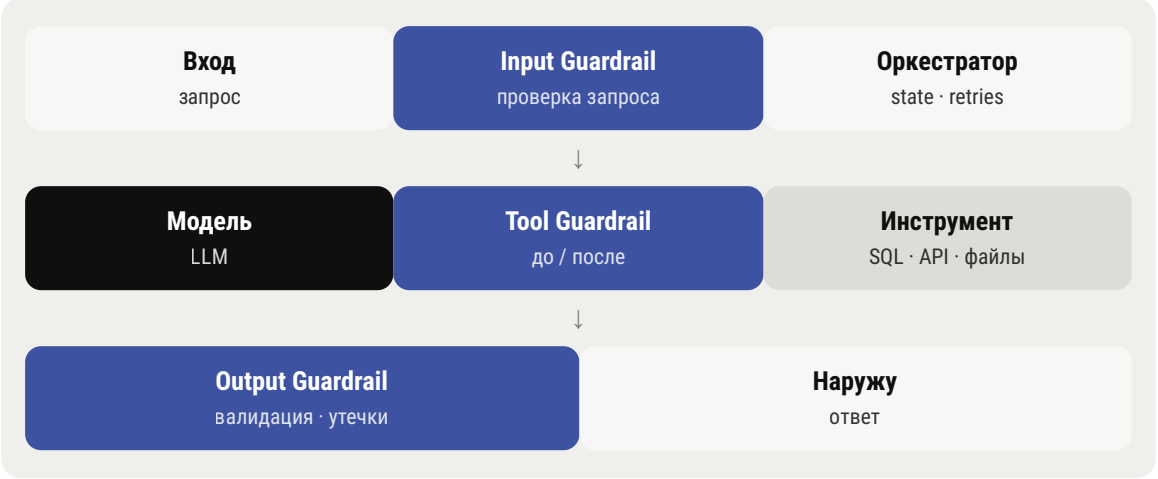
Григорий Добряков AI-DRIVEN HEAD OF ENGINEERING

Чек-лист готовности AI-агента к продакшену

Агент отправил письмо не тому адресату. Другой выполнил опасный SQL — потому что «пример команды на удаление» лежал в тексте тикета. Третий через RAG подтянул секрет и отдал его наружу. Ни одна модель не «сломалась» — каждая отработала штатно. **Проблема не в ошибке модели, а в отсутствии контроля последствий.** Этот чек-лист проверяет, есть ли вокруг вашей модели два слоя, без которых она — просто API: **guardrails** (проверки на границах доверия) и **harness** (инженерная обвязка эксплуатации).

КАРТА ПУТИ ЗАПРОСА

Где стоят guardrails — и что они проверяют



Слева направо доверие падает. На каждой стрелке — своя проверка. Guardrails живут на границах, а не «вокруг модели».

ГРАНИЦЫ ДОВЕРИЯ

Что доверенное, а что нет — определите до промптов

Системный промпт	Инструкции приложения — в системном сообщении или отдельном пользовательском сообщении.
Внешний контент	RAG, PDF, письма, GitHub README, результаты поиска — передаются как `tool_result`, не как инструкция.

!	**Пользовательский ввод**	Недоверенный по умолчанию. Фильтрация: классификация, поиск признаков атаки, PII.

ОСНОВНОЙ АРТЕФАКТ

8 проверок перед запуском агента в прод

- [ ] 01 **Границы доверия нарисованы явно**

  На каждой стоит своя проверка — на входе, на контексте, на ответах инструментов, на выходе.

- [ ] 02 **Вход, контекст и ответы инструментов фильтруются как недоверенные**

  Включая письма, PDF, результаты поиска, README и SQL-выдачу. Любой внешний источник — недоверенный.

- [ ] 03 **Policy layer закрыт правами, а не промптом**

  Класс опасных действий — деньги, удаление, production — недоступен по RBAC. Самый аккуратный выходной фильтр не спасёт, если агенту выдали право дропнуть таблицу.

- [ ] 04 **Необратимое проходит через человека**

  Движение денег, удаление данных, публикация наружу, изменение инфраструктуры. Подтверждение — только там, где действие необратимо или дорого. На каждый шаг — обесценивается.

- [ ] 05 **Промпты версионируются, поведение ловится регрессией**

  Промпты не в коде — версионируются, тестируются, проходят A/B. Перед выпуском — evals и регрессионный прогон. Версия хуже предыдущей → не в продакшен.

- [ ] 06 **Есть трейсинг, реплей сессий и мониторинг стоимости**

  Тексты запросов, ответы, трейсы по шагам, токены, стоимость, ошибки, вызовы инструментов. Счёт за токены растёт тихо — без мониторинга вскрывается уже в счёте.

- [ ] 07 **Состояние живёт в системе, а не «в голове модели»**

  История диалога, результаты инструментов, артефакты, статусы согласований, счётчики повторов — хранятся отдельно и контролируются оркестратором.

☐ 08 **На отказ модели есть план**

Retry, смена параметров генерации, роутинг на другую модель по цене и сложности, деградация вместо падения. Feature flags — на части трафика, с быстрым откатом.

САМООЦЕНКА

Оцените зрелость системы — 0–2 балла по каждому критерию

КРИТЕРИЙ	0 — НЕТ	1 — ЧАСТИЧНО	2 — ЕСТЬ
Границы доверия нарисованы	☐	☐	☐
Внешний контент фильтруется	☐	☐	☐
Policy layer = RBAC, не промпт	☐	☐	☐
Human-in-the-loop на необратимом	☐	☐	☐
Evals + регрессия перед выкаткой	☐	☐	☐
Трейсинг + мониторинг токенов	☐	☐	☐
State вне модели	☐	☐	☐
План отказа модели	☐	☐	☐

0–5 баллов

Системы контроля нет. Агент в проде = неконтролируемый blast radius.

6–10 баллов

Базовые проверки есть, но дыры остаются. Закройте критические пункты 03 и 04.

11–16 баллов

Промышленная готовность. Harness и guardrails работают как независимые слои.

Быстрый результат за 10 минут

Возьмите один реальный агент из вашего проекта и ответьте на три вопроса:

1. Какие инструменты он вызывает? Выпишите каждый.
2. Какие права у его роли? Может ли он удалить данные, отправить письмо, дропнуть таблицу?

3. Какой внешний контент попадает в контекст без фильтрации — письма, PDF, README, SQL-выдача?

Если на пункте 2 вы обнаружили право, которого агент не должен иметь — это первая дыра. Закройте её через RBAC, а не промптом.

> **Что делать дальше.** Пройдите чек-лист по каждому агенту в проде или перед запуском. Заведите список guardrails как журнал инцидентов — не пишите раз и навсегда, добавляйте по мере реальных случаев. Если пункт 03 (policy layer) не закрыт правами — это первый приоритет: никакой выходной фильтр не компенсирует агенту право на необратимые действия. LLM может предлагать действия, но выполнять их или нет — решает система, а не модель.

Григорий Добряков

AI-DRIVEN HEAD OF ENGINEERING

dobryakov.com

grigoriydobryakov@gmail.com

t.me/grigoriydobryakov

linkedin.com/in/dobryakov

youtube.com/@IT-Head
