

1 — Наблюдаемость (нельзя эксплуатировать то, чего не видишь)

Trace = вся задача; span = один шаг (вызов LLM или инструмента) с токенами и временем.
Одно обращение — обычно 40-75 спанов. Логируете только финал — отказ не переиграть.
OpenTelemetry GenAI: `invoke_agent -> chat + execute_tool; gen_ai.usage.input/output_tokens`.

2 — Оценка (хорош ли он на самом деле? — столп, который пропускают)

Golden set фиксированные задачи с эталоном; перепрогон на каждое изменение.
Три уровня end-to-end (успех) - trajectory (путь) - component (какой инструмент упал).
LLM-as-judge сильная модель судит по рубрике; шкала 0-5 ближе всего к людям (~0.89).
Надёжность мерьте `pass^k`, не `pass^1`: 0.9 на шаг = $0.9^8 = 43\%$ за восемь шагов.

3 — Стоимость (доллары за УСПЕШНУЮ задачу, не за токен)

Агент в цикле: 3-8 вызовов на задачу; 50k-200k токенов на одну; вызовы = 70-85% затрат.
Пример: 5 вызовов x 30 000 = 150 000 токенов; x \$5/M = \$0,75/задача; x10 000/день = \$7 500.
Агент за \$0,75 с провалом в половине случаев и ретраем = \$1,50 за успех. Маршрут/кеш/batch.

4 — Безопасность (агент ДЕЙСТВУЕТ — компрометация = плохое действие, не фраза)

Справочник: OWASP Top 10 для агентов (ASI01-ASI10, дек 2025) + NIST AI RMF GenAI Profile.
Риски видео-агента: угон цели (ASI01), злоупотребление инструментами (ASI02), память (ASI06).
Сигнатурная атака: КОСВЕННАЯ инъекция — команда спрятана в субтитре/транскрипте для агента.
Защиты: входные guardrails (данные != инструкции) + права least-agency + шлюз с человеком.

5 — Перед выпуском агента спросите:

- ☐ Трассируется ли каждый запуск (каждый вызов LLM и инструмента как span), чтобы переиграть отказ?
- ☐ Есть ли golden set, по которому агента оценивают на каждое изменение — задача, траектория, компонент?
- ☐ Отчитывается ли надёжность как `pass^k` (повтор успеха), а не только средний `pass^1` «у меня сработало»?
- ☐ Считается ли стоимость как доллары-за-успешную-задачу, лёгкие задачи — на дешёвой модели, контекст — в кеше?
- ☐ Держит ли агент права least-agency — минимум инструментов и score, привязанных к конкретному пользователю?
- ☐ Стоят ли входные guardrails, чтобы команда в субтитре или транскрипте не стала инструкцией?
- ☐ Одобрят ли человек каждое дорогое, необратимое или высокоставочное действие до того, как агент его сделает?
- ☐ Может ли команда показать контроли против названной рамки (OWASP ASI / NIST AI RMF / EU AI Act)?