Чек-лист архитектуры безопасного ИИ

Докладчик — Гарик Давтян. Ведущий — Денис Батранков.

10 вопросов к своей ИИ-системе. Заполняйте вместе с владельцем системы и архитектором. «Да» ставьте только при наличии подтверждения; «Не знаю» считайте за «Нет».

01 / ЧТО ЗАЩИЩАЕМ: ВЛАДЕЛЕЦ, ДАННЫЕ, СИСТЕМА

01 РЕЕСТР

У ИИ-системы есть владелец, назначение и перечень разрешённых данных и действий?

Проверка. В реестре указаны модель, инструменты, уровень риска, ответственный и дата пересмотра.

ОТВЕТ
☐ Да
☐ Нет

02 ДАННЫЕ

Разрешена ли передача этих данных выбранному сервису именно для вашей задачи?

Проверка. Есть согласованный маршрут данных. Лишнее удаляется до отправки; маскирование само по себе не разрешает передачу.

ОТВЕТ
☐ Да
☐ Нет

03 МОДЕЛЬ И ПОСТАВЩИК

Проверены обработка и хранение запросов, обучение на них и порядок смены модели?

Проверка. Условия и версия учтены. Контрольные задачи повторяются после изменений; в работе отслеживаются ошибки и запрещённые вызовы (дрифт поведения).

ОТВЕТ
☐ Да
☐ Нет

04 ПОИСК ПО ДОКУМЕНТАМ / RAG

Поиск передаёт модели только документы, доступные текущему пользователю?

Проверка. Тестовый документ не попадает в контекст без права доступа. Пользователь с доступом его получает: поиск действительно работает.

ОТВЕТ
☐ Да
☐ Нет

05 ИНФРАСТРУКТУРА

Эксперименты изолированы от рабочей среды, а секреты недоступны модели?

Проверка. Отдельный аккаунт, синтетические данные, разрешённые MCP-серверы и инструменты. Ключ из хранилища использует инструмент, а не модель.

ОТВЕТ
☐ Да
☐ Нет

02 / ЧЕМ ЗАЩИЩАЕМ: ЧЕТЫРЕ РУБЕЖА И РЕАГИРОВАНИЕ

06 ВХОД

Учтены пользователь и агент, ограничены доступ, число вызовов и бюджет задачи?

Проверка. Прямой обход шлюза закрыт; у задачи есть идентификатор. Превышение тестового лимита блокирует новые вызовы.

ОТВЕТ ☐ Да ☐ Нет

07 КОНТЕКСТ / ПРОМПТ

Внешние материалы отделены от доверенных инструкций и проверяются на инъекции?

Проверка. Есть тесты с инструкцией в письме, документе и ответе инструмента. При пропуске фильтра запрет действий из пункта 8 сохраняется.

ОТВЕТ ☐ Да ☐ Нет

08 ДЕЙСТВИЯ

Права и параметры проверяются кодом, а рискованные операции подтверждает человек?

Проверка. Разрешены только нужные операции. Подмена адресата, суммы или вложения требует нового согласования; модель не может подтвердить его сама.

ОТВЕТ ☐ Да ☐ Нет

09 ВЫХОД

Проверяется всё, что система выдаёт или отправляет наружу: текст, вложения и ссылки?

Проверка. Тестовые секреты не уходят по запрещённым каналам. Проверка не ограничена ответом в чате и работает перед отправкой.

ОТВЕТ ☐ Да ☐ Нет

10 ИНЦИДЕНТЫ

На тесте проверены остановка агента, отзыв прав и восстановление цепочки событий?

Проверка. Новый вызов после отзыва прав отклонён. Журнал связывает инициатора, инструмент и решение проверки; есть план восстановления и повторных тестов.

ОТВЕТ ☐ Да ☐ Нет

Как читать результат

Каждое «Нет» — задача с ответственным и сроком. Сначала закройте передачу запрещённых данных и опасные действия. Ограничение «только чтение» не предотвращает утечку. Фильтр промпта дополняет проверки доступа и операций.

Проверьте механизм отдельно от модели. Даже если модель отказалась выполнять учебную инъекцию, подайте запрещённую операцию напрямую на проверку инструмента. Ожидаемый результат — отказ и запись причины в журнале.