

ЧЕК-ЛИСТ · БЕЗОПАСНОСТЬ ИИ

10 проверок безопасности LLM-системы

Короткий рабочий чек-лист для CISO и архитекторов ИБ: что проверить в LLM-системе перед запуском в закрытом контуре — от prompt injection и прав на RAG-индекс до контроля над агентами и аудита.

prompt injection

RAG security

агенты

КИИ

Пройдите по пунктам перед запуском. Если по любому ответ «не уверены» — это место для аудита. Чек-лист не заменяет модель угроз, но показывает, где она нужнее всего.

1

Данные

Все данные, доступные модели и RAG, классифицированы по уровню конфиденциальности; известно, что и откуда попало в индекс.



2

Доступы

Действует единая ролевая модель (RBAC/ABAC), применяемая до retrieval; модель видит только то, к чему есть права у конкретного пользователя.



3

RAG

Контуры разной чувствительности изолированы; права на индекс описаны; загрузка документов проходит контроль источника.



4

Prompt injection

Есть санитизация ввода, policy layer и контекстные границы; внешний контент подаётся модели как недоверенный.



5

Tool use

Инструменты ограничены allowlist; всё лишнее запрещено по умолчанию; вызовы изолированы.



6

Агенты

Чувствительные и необратимые действия проходят через approval workflow; агент действует в правах пользователя, а не под общей учёткой.



7

Логирование

Запросы, ответы и вызовы инструментов журналируются; логи сами защищены как чувствительные данные.



8

Eval

Поведение системы регулярно проверяется на наборе тестов, включающих попытки инъекций и обхода ограничений.



9

Инциденты

Есть процедура реагирования: как обнаруживается, кто реагирует, как откатывается действие агента.



10

Эксплуатация

Три принципа, на которых держатся все десять пунктов. Если держать их в голове, чек-лист перестаёт быть формальностью.

1. Граница «данные / инструкции» размыта

Модель не отличает описание факта от команды. Вредоносная инструкция может прийти внутри документа, письма или ответа внешнего сервиса — то есть уже за сетевым периметром. Поэтому пункты 3–5 (RAG, prompt injection, tool use) важнее, чем кажется на схеме.

2. Защита рассчитывается на сбой фильтра

Надёжного «полного» метода предотвращения prompt injection не существует. Проектируйте так, будто часть инъекций пройдёт: главный вопрос — не «как не пропустить», а «что максимум сделает модель, если инъекция сработала». Ответ задают пункты 2, 5, 6 (права, allowlist, агенты).

3. Системный промпт — не средство защиты

Он стохастический и не может быть аудируемой границей безопасности. Важные ограничения выносятся за пределы модели — в детерминированную проверяемую среду: права доступа, allowlist, подтверждение человеком. Это пункты 2, 6, 9.

Нужно проверить конкретную систему?

AZONE AI проводит ИБ-аудит ИИ-систем: строим модель угроз для LLM-части, проверяем слои защиты от RAG-контуров до прав агентов и даём приоритезированный план по слабым местам — под закрытый контур и требования КИИ.

[Запросить ИБ-аудит ИИ-системы](#)

azoneai.ru/contact · услуга «ИИ для КИИ»