Nevel

БЕЗОПАСНОСТЬ

Модель угроз и негативные проверки

Сценарии утечки, prompt injection, злоупотребления инструментами, ошибки маршрута, изоляция и остаточный риск.

Облако / В контуре

https://docs.nevel.sh/security/threat-model/

Модель угроз Nevel начинается с последствий для бизнеса: раскрытие корпоративных данных, изменение источников и результатов, выполнение неразрешённого действия, потеря доступности и невозможность расследовать событие. Для каждого последствия определяется путь атаки, применяемая мера и проверка, способная показать отказ защиты.

Состав активов, нарушителей и обязательных мер в модели угроз зависит от внедрения. Подключение новой базы, внешнего провайдера, инструмента записи или публичного ассистента меняет поверхность атаки и требует пересмотра соответствующих сценариев.

Активы и границы

К защищаемым активам относятся исходные документы, диалоги, сведения о сотрудниках, коммерческие условия, секреты, таблицы соответствий, базы знаний, журналы и способность управлять политиками. Отдельный актив — полномочия подключённого инструмента: модельный ответ может причинить вред через действие в системе, даже если текст ответа не содержит секретов.

Границы проходят между пользователем и сервером, организациями, разрешённым контекстом и недоступными документами, данными и управляющими инструкциями, приложением и модельным получателем, рабочими процессами и эксплуатацией. Каждая граница должна иметь проверяемое правило. Доверие к сотруднику не означает доверие к каждому файлу, который он загрузил из внешней переписки.

Возможные нарушители

Источник риска	Начальные возможности	Что ограничивает последствия
Внешний человек без учётной записи	Попытки входа, публичные URL, подбор ключей	Аутентификация, лимиты, отсутствие публичных хранилищ
Обычный участник	Свои документы, разрешённые модели и инструменты	Серверные права, изоляция пространств, обязательные политики
Скомпрометированный пользователь	Действующая сессия и доступ сотрудника	Минимальные права, отзыв, ограничения экспорта и маршрута
Автор вредоносного документа	Текст, который может попасть в RAG	Разделение данных и инструкций, контроль источников и инструментов
Администратор	Изменение части настроек	Разделение полномочий, аудит изменений и ограничение доступа к содержимому
Компрометированный сервис	Сервисная идентичность и её данные	Узкая область доступа, сегментация, отдельные секреты
Внешний обработчик	Полученный по разрешённому маршруту контекст	Минимизация, маскирование, условия обработки и запрет неподходящих категорий

Ошибки добросовестного пользователя рассматриваются наряду с намеренной атакой: загрузка полной кадровой таблицы, публикация ассистента с личным подключением или выбор внешней модели для закрытого документа могут привести к тем же последствиям.

Матрица основных угроз

Угроза	Контроль	Проверка отказа защиты
Чтение чужого файла по известному ID	Проверка организации и прав на сервере	Запрос участника другого пространства
Утечка через RAG	Права до выдачи фрагмента и актуализация доступа	Поиск контрольной фразы чужого или удалённого документа
Пропуск чувствительного значения	DLP, классификация целого документа, локальный маршрут	OCR, транслитерация, разрыв номера, косвенные признаки
Обход защищённого API	Обязательное серверное и сетевое правило маршрута	Альтернативный endpoint, фоновые вызовы и прямой адрес модели
Утечка при fallback	Резервный маршрут того же разрешённого класса	Отказ локальной модели при документе «только в контуре»
Кража исходных значений масок	Изоляция соответствий и проверка восстановления	Поддельная маска другого запроса или пространства
Компрометация секрета	Ограниченная область, защищённое хранение, отзыв	Старый ключ после ротации; поиск секрета в логах
Неавторизованное действие инструмента	Серверная проверка операции и полномочий	Инструкция в документе выполнить запись или внешнюю отправку
Истощение ресурсов	Ограничение размера, частоты, параллелизма и повторов	Архив с большим раскрытием и серия конкурентных задач
Скрытое ослабление защиты	Версии политик, административный аудит	Изменение маршрута или логирования перед запросом
Возвращение удалённых данных	Очистка производных объектов и контроль восстановления	Восстановление старой копии после удаления

Меры сочетаются. DLP не заменяет права на источник, сетевой запрет не обнаруживает злоупотребление разрешённым инструментом, а аудит не предотвращает уже состоявшуюся утечку. Каждая мера должна иметь собственный наблюдаемый результат.

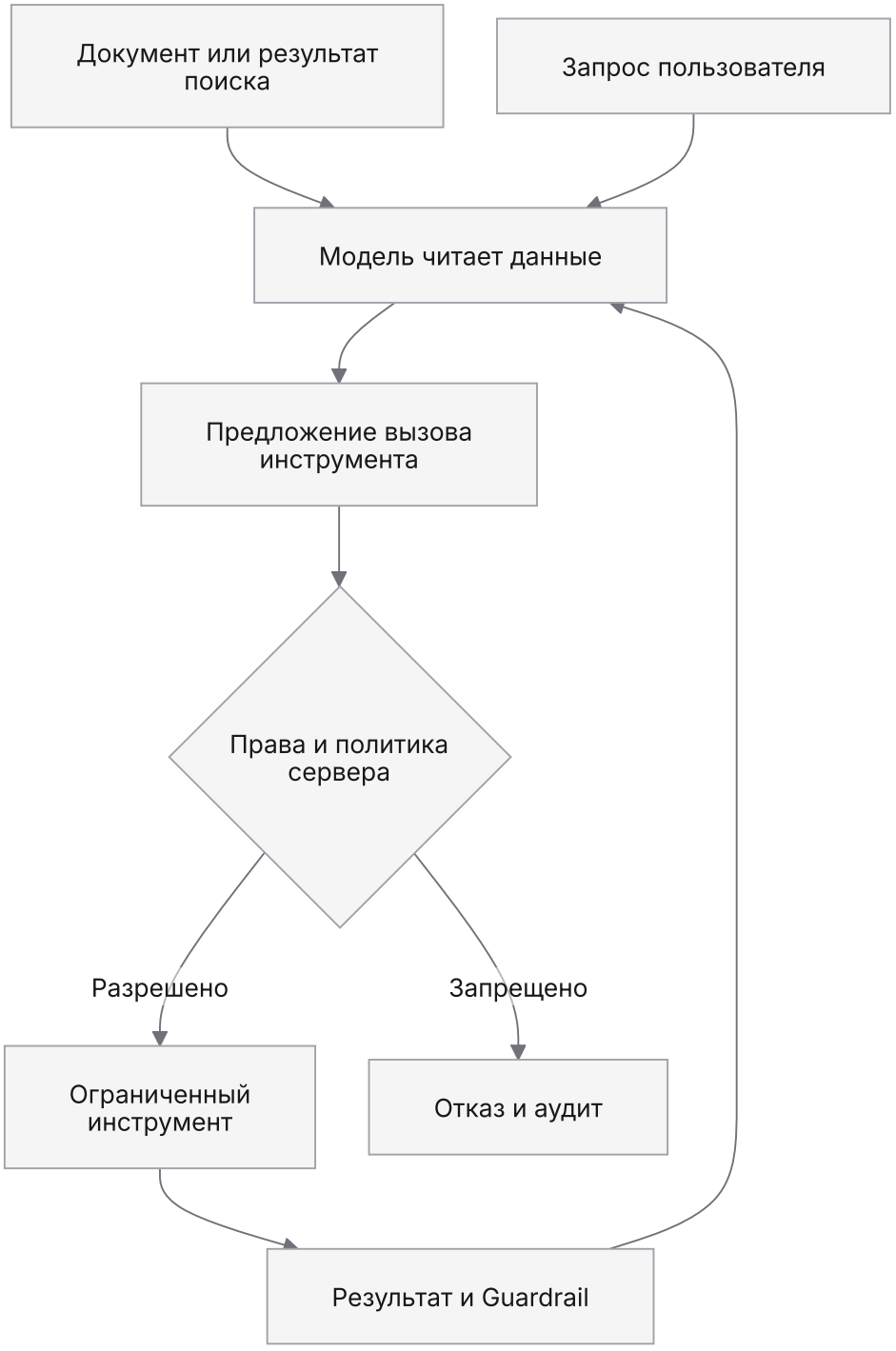
Prompt injection

Подробный процесс проверки контекста и инструментов, пример профиля реакции и программа испытаний вынесены в защиту от prompt injection. Допуск источников, карантин и смена версий описаны в защите файлов и баз знаний.

Prompt injection — попытка заставить модель принять недоверенное содержание за управляющую инструкцию. Источником может быть документ, веб-страница, результат поиска, письмо или ответ подключённого сервиса. Типичные последствия — раскрытие контекста, обращение к другому источнику, изменение маршрута или выполнение действия от имени пользователя.

СХЕМА 1

Недоверенный текст и полномочия



Защита опирается на то, что найденный текст не обладает полномочиями. Он не меняет серверный список моделей, не выдаёт секрет, не предоставляет доступ к чужому файлу и не разрешает сетевой адрес. Обнаружение характерной инструкции служит дополнительным сигналом, но блокировка одной фразы не является универсальным решением: атака может быть косвенной, многошаговой или скрытой в структуре документа.

Для проверки используется синтетический документ с просьбой раскрыть контрольную фразу недоступного источника или вызвать запрещённый инструмент. Успех — отсутствие раскрытия и действия, а не обязательное употребление моделью конкретной фразы отказа. Модель может пересказать вредоносный текст как содержание документа, не исполняя его.

Ошибки модели и целостность результатов

Модель может неверно интерпретировать таблицу, придумать факт, ссылку или реквизит. Это риск целостности, даже если конфиденциальность сохранена. Для аналитического отчёта ответ связывается с источниками, а существенное действие проверяется по исходным данным. Создание черновика отделяется от отправки письма, изменения записи или утверждения документа.

Особенно важны реквизиты, расчёты, юридические формулировки и инструкции с последствиями для людей. Наличие ссылки в ответе не доказывает правильность вывода: проверяется, что источник существует, доступен проверяющему и подтверждает утверждение. Автоматическая цепочка из нескольких моделей не отменяет этот риск и не получает больше прав только из-за количества этапов.

Обновление модели может изменить качество, поведение инструментов и восприимчивость к инструкциям. Поэтому новая модель проходит регрессионный набор для разрешённых сценариев. Автоматическое появление модели в каталоге и разрешение использовать её для закрытых данных — отдельные решения.

Остаточный риск маскирования

DLP может пропустить сущность или заменить лишний текст. Внешний получатель может определить человека по совокупности немаскированных признаков. Поэтому правило «при отсутствии масок отправлять куда угодно» недостаточно для данных с обязательным локальным режимом. При таком требовании локальный маршрут задаётся категорией и политикой независимо от результата детектора.

Обратимое соответствие дополнительно сохраняет возможность восстановления исходных значений. Его изоляция и короткий срок снижают последствия компрометации, но не дают автоматического юридического вывода об отсутствии персональных данных. Техническая оценка качества и правовая допустимость внешней передачи проводятся отдельно; см. 152-ФЗ.

Размещение платформы и угрозы маршрута

В облаке отдельно оцениваются изоляция организаций и доступ эксплуатации Nevel. В on-prem оцениваются управление инфраструктурой заказчика и её сервисные полномочия. В обоих случаях внешняя обработка зависит от модели и политики: российские провайдеры, self-hosted Nevel в РФ и глобальные API могут быть разрешёнными отдельными маршрутами.

Для каждого маршрута проверяются конечные и промежуточные получатели. В API через OpenRouter этот сервис включается в цепочку наряду с конечной моделью. Self-hosted DeepSeek у Nevel в РФ оценивается по месту выполнения этого экземпляра; API внешнего владельца модели с тем же названием имеет другую границу доверия.

Изолированный профиль исключает внешнюю рабочую обработку, но сохраняет угрозы инсайдера, компрометации сервера, ошибок прав, вредоносных документов и утраты копий. Внутренняя модель может ошибаться и предлагать опасные действия. Локальное размещение Platform не отменяет Guardrail и контроль инструментов, а обычный on-prem с внешними моделями дополнительно сохраняет риски выбранных получателей.

В тесте отказа проверяется именно политика данных: заранее разрешённый резерв может использоваться с событием аудита; запрещённый маршрут и скрытый fallback не допускаются ни в облаке, ни в on-prem.

Программа испытаний и доказательства

Набор делится на обычные рабочие сценарии и негативные проверки. Первые подтверждают, что защита сохраняет полезность: файл читается, ответ восстанавливается, разрешённый инструмент выполняет задачу. Вторые проверяют границу: чужой документ, отозванный ключ, запрещённая модель, недоступный контроль, поддельная маска и повреждённый файл.

Для каждого теста записываются исходные условия, версия политики, синтетические данные, ожидаемый результат, фактические события и сетевой след при необходимости. Проверка не должна использовать реальные секреты и чужие персональные данные только ради демонстрации утечки. Если тест не проведён, результат отмечается как непроверенный, а не как успешный на основании архитектурного описания.

При обнаружении отклонения сначала ограничивается опасный маршрут или возможность, затем воспроизводится причина и выполняется исправление. Возврат возможности требует повторения конкретного проваленного теста и затронутых соседних сценариев. Дальнейшие действия изложены в инцидентах и эксплуатации.