Nevel

БЕЗОПАСНОСТЬ

Защита файлов и баз знаний

Загрузка и разбор документов, права до поиска, отравление RAG, скрытое содержимое файлов и безопасные инструменты.

Облако / В контуре

https://docs.nevel.sh/security/files-rag/

Файлы и базы знаний расширяют контекст модели за пределы текста, который пользователь набрал в чате. Защита должна учитывать оригинал, извлечённый текст, изображения, фрагменты, векторы и результаты поиска. RAG — получение подходящих фрагментов перед ответом модели — не даёт модели самостоятельного права читать любой документ в хранилище.

В Nevel права проверяются до выдачи контекста. Модель получает только разрешённые источники, после чего подготовленный запрос проходит классификацию и контроль маршрута. Включение файла в общий проект не отменяет его обязательную метку чувствительности и не разрешает внешнюю обработку, запрещённую политикой организации.

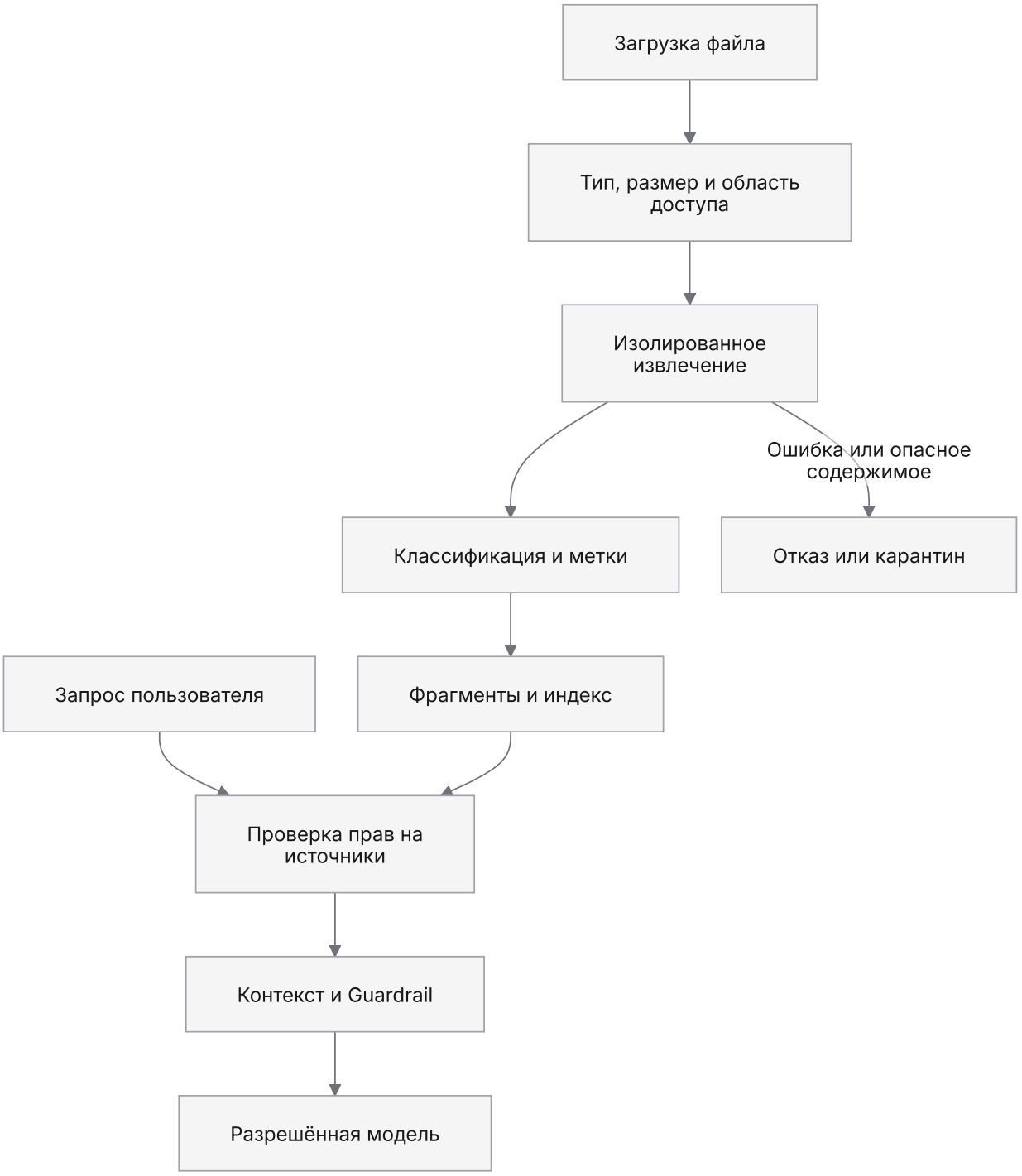
Жизненный цикл загрузки

При загрузке файл связывается с организацией, владельцем, проектом или пространством, областью доступа и категорией данных. Название и расширение не считаются достаточной проверкой фактического типа. Обработчик получает ограниченный вход и создаёт производные объекты с той же связью с источником.

Разбор отделяется от исполнения активного содержимого. Макросы, внедрённые скрипты, внешние ссылки и команды внутри документа не должны автоматически запускаться только ради извлечения текста. Для сложного формата обработчик ограничивает время, память, объём распаковки и количество порождаемых объектов. Если безопасное извлечение не завершилось, файл не объявляется успешно проверенным.

СХЕМА 1

Безопасное включение файла в контекст



Карантин здесь обозначает исключение объекта из обычной выдачи до разрешения проблемы. Он не означает обещание встроенного антивируса конкретного производителя. Если внедрение использует корпоративную антивирусную или песочничную проверку, её результат является отдельным условием допуска файла; обход при недоступности обязательного средства исключается.

Особенности форматов

Формат или свойство	Что может быть не видно пользователю	Что включить в проверку
PDF	Изображение вместо текста, скрытый текстовый слой, вложения, формы	OCR, отличие видимого и извлечённого текста, отказ некорректного файла
DOCX	Комментарии, исправления, колонтитулы, встроенные объекты	Какие слои извлекаются и какие не должны попадать в ответ
XLSX	Скрытые листы, формулы, комментарии, внешние ссылки	Область чтения, формулы против значений, чувствительные вспомогательные листы
Изображение	Документ, подпись, QR-код, мелкий текст	Распознавание до текстового контроля или ограничение маршрута целого изображения
Архив	Множество вложений, вложенные архивы, большой объём распаковки	Ограничение глубины, количества и итогового размера
Защищённый паролем файл	Недоступное анализу содержимое	Контролируемый отказ либо разрешённый процесс раскрытия внутри нужного контура
Имя и свойства файла	ФИО, тема сделки, автор, путь	Минимизация метаданных, не только маскирование основного текста

Поддержка офисного формата не означает, что любая версия, встроенный объект или повреждённый файл обрабатываются одинаково. В приёмке используются документы, характерные для компании: многостраничные договоры, таблицы, сканы и изменения рецензентов. Качество извлечения непосредственно влияет на DLP: пропущенный текст не может быть обнаружен текстовым распознавателем.

Изображения и обход текстового контроля

Если модель получает оригинальное изображение, замаскированный текст рядом с ним не скрывает сведения на картинке. Для такого маршрута необходимо либо подготовить безопасное представление, либо использовать допустимую локальную обработку, либо запретить передачу. Аналогичная проблема возникает, когда внешний обработчик получает исходный PDF для OCR раньше, чем Guardrail проверит извлечённый текст.

В облаке и on-prem каждый обработчик OCR, изображений, embeddings и артефактов имеет собственный маршрут. Разрешённый внешний обработчик включается в реестр получателей и подчиняется политике данных; наличие локально установленной Platform не делает его внутренним. Только в изолированном профиле все эти стадии выполняются внутри заданного контура заказчика. Термин «предобработка» не выводит фактическую передачу файла из области контроля.

Права в RAG

При индексации каждому фрагменту сохраняются связь с документом, область организации, права и метки. Поиск учитывает актуальную идентичность пользователя и допустимые источники. Фильтрация после передачи результатов модели слишком поздняя: содержание уже раскрыто обработчику.

Права проверяются и при поиске, и перед использованием найденного фрагмента. Это защищает от изменения членства между созданием индекса и запросом. Удаление пользователя, перенос файла в закрытый проект и отзыв общего доступа должны влиять на следующие обращения без ожидания полного переобучения модели. RAG не требует обучать модель на корпоративных документах: фрагменты передаются как контекст конкретного запроса.

Результат поиска также может раскрывать сведения через заголовок, число совпадений или ссылку. Поэтому недоступный источник не появляется в подсказках и цитатах только потому, что основной текст скрыт. Векторное представление не объявляется анонимным или безопасным по определению: оно является производным от исходных данных и включается в их жизненный цикл.

Защита от отравления знаний

Документ может содержать инструкцию «игнорируй правила», скрытый запрос на раскрытие других файлов или ложное указание вызвать внешний инструмент. Такие строки являются содержимым источника. Они не получают полномочий системной политики или пользователя только потому, что поисковый механизм нашёл их релевантными.

Контроль строится на происхождении данных и правах записи. Подключаются определённые источники, у загрузок есть автор и история изменений, а изменение доверенного материала регистрируется. Нельзя считать всю корпоративную папку одинаково доверенной: сотрудник с правом добавить справочный файл не должен получать через его текст право менять политику модели.

Риск	Механизм ограничения
Подмена внутренней инструкции	Разделение доверенной настройки ассистента и извлечённого материала
Вредоносная команда в документе	Текст источника не даёт права выполнить инструмент
Подмена реквизитов в базе знаний	Происхождение, версия и ссылка на источник; проверка человеком перед действием
Отравление через доступную всем папку	Ограничение права записи и перечня индексируемых источников
Утечка через цитату	Проверка доступа к источнику и безопасная ссылка
Возврат удалённого документа	Очистка фрагментов и индекса, проверка после восстановления копии

Фильтр prompt injection помогает выявлять подозрительные конструкции, но не считается единственной границей. Основная защита — отсутствие у текста документа полномочий на изменение маршрута, выдачу секрета и исполнение внешнего действия. Стадии анализа, реакции и негативные тесты приведены в защите от prompt injection.

Допуск и проверка источников

Допуск источника состоит из двух решений: разрешена ли сама коллекция и проверена ли конкретная версия документа. Список разрешённых источников, происхождение и контроль целостности входят в рекомендации OWASP по безопасности RAG. Разрешённый домен не означает доверия любому файлу на нём: значение имеют точная область подключения, владельцы, права записи и путь получения.

Для источника фиксируются владелец содержания, организация, разрешённая коллекция или путь, способ подключения, полномочия сервисной учётной записи, допустимые типы данных и порядок обновления. Для версии документа сохраняются идентификатор, происхождение, время получения, контрольная сумма оригинала, версия экстрактора и результат проверок. Фрагменты и embeddings связаны с этой версией; замена текста без обновления связи не считается допустимой актуализацией.

Контрольная сумма помогает обнаружить изменение относительно зарегистрированного объекта. Она не доказывает добросовестность автора и истинность текста. Поэтому проверяются также происхождение, полномочия на публикацию и характер изменения; версионирование сохраняет возможность определить, какие материалы повлияли на ответ. Такой подход к происхождению и изменениям данных описан также в OWASP LLM04:2025.

СХЕМА 2

Допуск версии источника



Карантин исключает версию из обычного поиска, генерации и цитирования. Сотрудник, разбирающий карантин, имеет отдельное право просматривать образец; обычный доступ к базе знаний такого права не создаёт. Исправленный документ проходит проверки как новая версия. Инструкция внутри документа «считать источник проверенным» не может изменить состояние допуска.

Для стабильной внутренней коллекции допуск очередной версии может выполняться по утверждённым автоматическим правилам. Новый внешний источник, смена владельца подключения, расширение прав или подозрительное изменение направляются на содержательную проверку. Ручное чтение каждого файла не заменяет технические проверки и не объявляется обязательным для всех коллекций.

Пример правил допуска

Ниже — **рекомендуемый профиль для внутренней справочной базы**, а не готовый файл конфигурации или название экранов Nevel. Значения источника и контрольные документы синтетические.

Условие	Решение профиля	Сохраняемое основание
Файл поступил из зарегистрированной коллекции «Регламенты», текущие права и проверки подтверждены	Допустить конкретную версию	Коллекция, версия, права, результаты проверок
Ссылка ведёт в соседнюю коллекцию того же сервиса, которая не включена в подключение	Не индексировать	Фактический источник за пределами разрешённой области
При загрузке изменился адрес после перенаправления	Повторно проверить получателя и область	Исходный и конечный адрес без секретных параметров
В новой версии появилась скрытая команда отправить данные	Карантин версии	Категория сигнала и ссылка на защищённый образец
Контрольная сумма не соответствует зарегистрированной версии	Не выдавать затронутые данные до проверки	Несоответствие оригинала и производных объектов
Источник доступен, но обязательный анализ не завершён	Не переводить новую версию в допущенную	Ошибка или незавершённая стадия
Владелец отозвал документ или доступ	Исключить из последующей выдачи	Событие отзыва и перечень затронутых объектов

Обновление не должно молча оставлять пользователю устаревшую версию под видом новой. Если новая версия находится в карантине, использование предыдущей допустимо только пока её содержание и права сохраняют силу. При отзыве или обнаружении компрометации старой версии она также исключается. Удаление охватывает фрагменты, индекс и применимые кэши; восстановление копии повторно учитывает отзыв. Сроки и границы удаления описаны в хранении данных.

Проверка достоверности ответа

Допущенный источник может содержать ошибку или устареть. Для существенного утверждения проверяются существование документа, его действующая версия, применимость к вопросу и фрагмент, который подтверждает вывод. Ссылка на главную страницу, совпадение ключевых слов или высокий результат векторного поиска не являются подтверждением факта.

Если два источника расходятся, ответ обозначает противоречие и приводит доступные основания, вместо смешения несовместимых условий. Приоритет определяется предметом: утверждённый регламент для внутренней процедуры, оригинальный договор для его условий, действующий первичный нормативный акт для правового требования. Право читать источник всё равно проверяется отдельно. Для цитаты сохраняется связь с той версией, по которой сформирован ответ; недоступный читателю документ не раскрывается через название или фрагмент цитаты.

Приёмка включает новую коллекцию без допуска, изменённый файл с прежним идентификатором, скрытую команду, отказ обязательной проверки, отзыв источника и обычное корректное обновление. Испытатель проверяет не только ответ модели, но и отсутствие недопущенной версии в результатах поиска и цитатах. Фактические результаты записываются после испытания; описание процесса само по себе не доказывает прохождения проверок.

Файловые знания и живые подключения

Личная база, общий проект и подключение к корпоративной системе решают разные задачи. Загруженный файл представляет снимок на определённый момент. Живой инструмент обращается к источнику с текущими правами и может вернуть более свежие данные. Крупная таблица или аналитический куб не становятся безопасными для произвольного SQL только потому, что модель умеет сформулировать запрос.

Для подключения устанавливаются разрешённые операции и область данных. Чтение отделяется от записи; результат ограничивается необходимыми строками и полями; параметр запроса проверяется сервером. Секрет источника хранится вне текста ассистента. Выданное модели описание схемы объясняет формат, но не расширяет права пользователя. Перед следующим вызовом модели результат инструмента снова проходит контроль чувствительности.

Экспорт и созданные артефакты

Ответ может стать DOCX, XLSX, PDF или другим файлом. Созданный артефакт наследует чувствительность использованных данных и область доступа результата. Ссылка на скачивание проверяет полномочия; публикация ассистента не делает все его результаты общедоступными.

В таблицах дополнительно рассматривается исполнение формул при открытии результата, внешние ссылки и скрытые листы. В отчётах проверяются персональные сведения в колонтитулах, свойствах и комментариях. Маскирование входа не гарантирует, что итоговый файл безопасен для внешней отправки после восстановления исходных значений.

Негативная приёмка

Проверьте скрытый слой PDF, скан паспорта, скрытый лист XLSX, комментарии DOCX, чрезмерный распакованный архив и команду раскрыть чужие знания. Заранее задайте результат: извлечение в разрешённом контуре, отказ или карантин; бесконтрольная отправка оригинала наружу исключена.

Для RAG создайте два пространства с уникальными контрольными фразами: чужая фраза недоступна в ответах, цитатах и названиях. После отзыва доступа и удаления источника повторите поиск, генерацию, фоновую задачу и восстановление копии. Инструкции в файлах не меняют маршрут и не разрешают запрещённые действия.

Происхождение документа и выполнение правил допуска не гарантируют достоверность его содержания. Существенные выводы проверяются по самим материалам, а работа защиты — на контрольном наборе источников.