Nevel

БЕЗОПАСНОСТЬ

Guardrail: маскирование и маршрутизация

15 типов сущностей, режимы распознавания, классификация документов, восстановление масок и поведение при ошибках.

Облако / В контуре

https://docs.nevel.sh/security/guardrail/

Guardrail контролирует подготовленный запрос до передачи модели. Он объединяет обнаружение чувствительных сущностей, замену значений, классификацию документа и выбор допустимого маршрута. Эти операции решают разные задачи: замена имени сокращает раскрытие идентификатора, а классификация финансового отчёта может запретить передачу всего отчёта независимо от найденных имён.

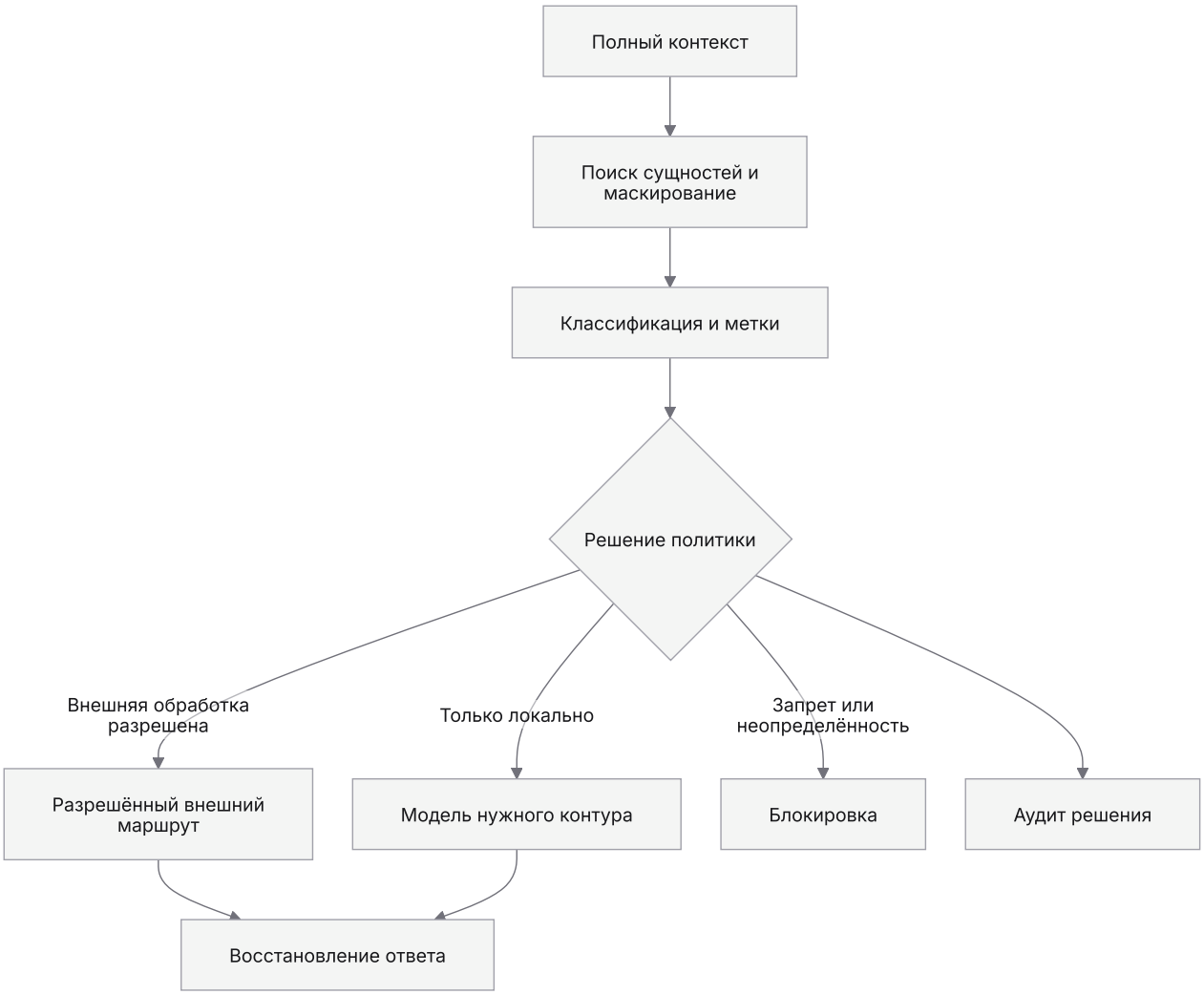
Защищённый API-маршрут использует префикс `/secure/v1`. Наличие обычного совместимого API-маршрута не означает, что к нему автоматически применяется тот же набор DLP-проверок. Для обязательной защиты администратор закрепляет защищённый маршрут в клиентах и закрывает обходные пути сетевыми и серверными правилами. Доступность конкретных методов, параметров и потокового режима проверяется по документации API, а не по одному сходству URL.

Порядок обработки

Gateway проверяет доступ и ограничения запроса. DLP анализирует текст, заменяет обнаруженные сущности и сохраняет соответствия. Классификатор оценивает содержание целиком и обязательные метки, после чего маршрутизатор выбирает разрешённый периметр и модель. После ответа выполняется восстановление допустимых масок и запись события обработки. История, извлечённый текст и результаты инструментов включаются в эту последовательность перед соответствующим модельным вызовом.

СХЕМА 1

Решение Guardrail



Как обнаруживаются сведения

Обнаружение сочетает правила для структурированных последовательностей и модели распознавания именованных сущностей, адаптированные к русскоязычным данным. Правила полезны для форматов документов, счетов и контактов; контекстное распознавание — для имён, адресов и неоднозначных фрагментов. Более строгая обработка использует дополнительный анализ сложного контекста.

В базовом текстовом обработчике используются Microsoft Presidio, NLP-обработка spaCy и собственные распознаватели русскоязычных форматов, включая телефоны, электронную почту и имена с инициалами. Распознаватель сочетает шаблон, контекстные слова и оценку уверенности. Этот состав объясняет принцип обработки; полный профиль Guardrail определяется также расширенным набором сущностей, классификацией и маршрутизацией. Наличие одной библиотеки само по себе не подтверждает весь профиль защиты.

Ни одна из этих стадий не является доказательством отсутствия всех персональных данных. Текст с ошибками OCR, транслитерацией, вставленными пробелами, нетипичным сокращением или косвенным описанием человека может потребовать другого правила или локального маршрута. Название конкретной библиотеки не определяет качество защиты: проверяется настроенный набор распознавателей на реальных форматах организации и синтетических примерах.

Тип сущности	Что распознаётся	Существенное ограничение проверки
PERSON	Имена, фамилии, инициалы	Совпадение с названием организации или термином может давать ложное срабатывание
PHONE_NUMBER	Телефонные номера, в том числе российские формы записи	Добавочные номера, разделители и ошибки распознавания требуют примеров
EMAIL_ADDRESS	Адреса электронной почты, включая кириллические варианты	Косвенная запись словами и разорванный адрес сложнее обычного формата
CREDIT_CARD	Номера платёжных карт	Номер в изображении становится текстом только после распознавания
IP_ADDRESS	IP-адреса	Адрес может быть техническими данными, а его чувствительность зависит от контекста
INN	ИНН физических и юридических лиц	Сам по себе ИНН организации не равен персональным данным физического лица
SNILS	СНИЛС	Пробелы, дефисы и ошибки OCR проверяются отдельно
PASSPORT	Данные внутреннего паспорта	Контекст серии и номера важен для отличия от других последовательностей
ADDRESS	Почтовые и фактические адреса	Неполные адреса и свободная форма повышают неопределённость
DATE_OF_BIRTH	Даты рождения	Обычная дата события не должна автоматически трактоваться как дата рождения
BANK_ACCOUNT	Банковские счета	Счёт, БИК и другие реквизиты различаются по назначению
DRIVER_LICENSE	Водительские удостоверения	Проверяются используемые варианты записи
FOREIGN_PASSPORT	Заграничные паспорта	Форматы документа и контекст отличаются от внутреннего паспорта
VEHICLE_PLATE	Регистрационные номера автомобилей	Нужны проверки регионов, похожих символов и OCR
MEDICAL_POLICY	Номера полисов ОМС и ДМС	Вид номера не раскрывает автоматически всю медицинскую информацию в документе

Перечень показывает покрываемые классы, а не исчерпывающее юридическое определение персональных данных. Сведения о здоровье, зарплате, поведении или трудовой оценке могут оставаться чувствительными без любого из этих идентификаторов. Специальные категории оцениваются по содержанию и цели обработки, см. 152-ФЗ.

Режимы распознавания

Режим	Поведение	Подходящий контрольный сценарий
Balanced	Баланс полноты обнаружения и числа лишних замен; базовый режим	Обычные деловые тексты с контактами, договорами и перепиской
Strict	Более агрессивная обработка неоднозначных фрагментов	Повышенная чувствительность, допустимость дополнительных масок и большей задержки

Режим	Поведение	Подходящий контрольный сценарий
Permissive	Снижение лишних замен в техническом тексте	Код, технические адреса и числовые значения при отдельно определённой допустимости маршрута

Режим влияет на компромисс между пропуском и лишней заменой. Permissive не является способом разрешить запрещённую обработку. Strict не означает стопроцентное обнаружение. Политика может независимо запретить внешний маршрут для класса данных, даже если пользователь предпочитает более мягкое распознавание.

При оценке качества отдельно считают полноту по каждому типу сущности, ложные срабатывания, ошибки восстановления и долю документов, отправленных в неверный класс. Сводная «точность» без состава выборки скрывает важные различия: редкие паспортные данные могут распознаваться хуже частых адресов электронной почты.

Классификация целого документа

Классификатор определяет допустимость обработки документа как целого. Примеры категорий — общедоступный материал, внутренний рабочий документ, коммерческое предложение, финансовый отчёт, кадровые сведения. Организация использует свои определения и обязательные метки. Категория не привязана только к расширению файла: финансовая таблица может быть вставлена в обычное сообщение.

Место Platform не является единственным условием решения. В облаке и on-prem используются разрешённые локальные, российские, self-hosted Nevel в РФ и глобальные маршруты, доступные в конфигурации. Например, self-hosted DeepSeek в РФ и API внешнего провайдера с той же моделью различаются получателем и географией. Подробное сравнение — в вариантах маршрутизации.

Решение учитывает несколько источников ограничений: административную политику, метку исходного документа, выбранный пользователем повышенный уровень чувствительности, результат обнаружения и классификации. Пользовательская отметка «только локально» усиливает ограничения; текст внутри файла «можно отправлять всем» не снимает обязательный запрет. При конфликте действует более ограничительный допустимый маршрут.

Условие	Решение
Допущены внешние модели, содержимое соответствует разрешённой категории	Маскирование и отправка по разрешённой цепочке получателей
Разрешена обработка в РФ у определённых получателей	Только выбранный допустимый российский маршрут: локальная модель, self-hosted Nevel или российский провайдер
Изолированный профиль или метка «только в контуре заказчика»	Только модель внутри заданного контура
Категория запрещена для любых моделей	Блокировка и событие аудита
Для категории требуется локальная обработка, но нужный обработчик недоступен	Отказ или другой разрешённый локальный обработчик
Проверка не завершилась	Защищённый маршрут не передаёт непроверенный текст наружу

Приоритеты, разрешённые и запрещённые категории и резервные маршруты изменяются административной политикой. Изменение применяется к новым решениям без необходимости разрешать обход проверки на время обновления. Версия политики сохраняется вместе с событием, чтобы объяснить прошлое решение после изменения настроек.

Маски и обратное восстановление

Каждой обнаруженной сущности назначается заменитель, а исходное значение помещается в таблицу соответствий. Область соответствий ограничена организацией и текущим запросом или разрешённым контекстом обработки. Модель не получает таблицу и не может запросить произвольное раскрытие значения через текст ответа.

При восстановлении проверяются принадлежность маски, срок жизни соответствия и право получить результат. Совпадение текста с шаблоном маски само по себе недостаточно. Поддельный идентификатор из другого диалога не должен открывать чужое значение. Просроченное соответствие не восстанавливается эвристически. Повреждённая моделью маска остаётся нераскрытой либо приводит к контролируемой ошибке обработки.

Маски могут влиять на качество: модель не знает склонение фамилии, свойства удалённого адреса или часть контекста. Поэтому приёмка проверяет не только отсутствие исходных значений на внешнем выходе, но и полезность результата, сохранение структуры документа и корректность восстановления. Сроки таблицы соответствий разобраны в хранении.

Ограничения по времени обработки

В базовом описании Guardrail версии 1.2 приведён профиль добавочной задержки шлюза: p50 до 150 мс и p99 до 500 мс без времени ответа LLM. В обсуждении практического сценария Balanced также использовался ориентир около 500 мс дополнительной обработки. Это разные контексты оценки, которые нельзя превратить в одно безусловное обещание для любого текста, файла и нагрузки.

Для испытания отдельно измеряются приём, DLP, классификация, ожидание модели и восстановление. Фиксируются размер входа, наличие OCR, режим, параллелизм и оборудование. Процентили рассчитываются по достаточной серии запросов, а не по одному успешному ответу. Обязательные показатели конкретной поставки определяются её SLA и программой испытаний.

Негативные проверки

Проверочный корпус содержит все 15 классов, несколько сущностей одного типа, кириллицу, транслитерацию, пробелы внутри номера, ошибку OCR, ложные совпадения в коде и чувствительные сведения только в истории. Отдельный набор содержит документ без явных идентификаторов, который всё равно запрещён к внешней передаче.

Далее проверяются: поддельная и разделённая потоком маска; повтор чужого идентификатора; истёкшее соответствие; отказ DLP; отказ локальной модели; попытка вызвать обычный API вместо защищённого; смена модели автоматическим режимом; запрещённый результат инструмента. Успех означает отсутствие обхода и понятный след решения, а не только красивый ответ в интерфейсе.

Смежные меры: потоки, AI firewall и сеть, аудит. Обезличивание и правовой статус передачи рассматриваются отдельно в 152-ФЗ.