Nevel

ВАРИАНТЫ УСТАНОВКИ

On-prem в инфраструктуре компании

Установка Platform у заказчика, подключение локальных и внешних моделей по политике и отдельный изолированный профиль.

В контуре

https://docs.nevel.sh/deployment/onprem/

On-prem — установка Platform в инфраструктуре заказчика. Платформа может использовать локальные модели, российские провайдеры, self-hosted модели Nevel на серверах в РФ и глобальные модели по разрешённым маршрутам. Полностью внутренняя обработка является отдельным изолированным профилем, а не обязательным значением слова on-prem.

Состав установки

Внутри инфраструктуры заказчика размещаются компоненты поставки: пользовательский интерфейс, прикладные сервисы, рабочие хранилища и защитная обработка. Местоположение моделей и вспомогательных обработчиков задаётся отдельно. Для каждого OCR, векторизатора, инструмента или модельного endpoint указываются исполнитель, сеть, данные и разрешённая область обработки.

Комплект установки фиксирует версии, ресурсы, конфигурацию связей, порядок развёртывания, обновления и восстановления. Команды берутся из комплекта конкретной версии. Наличие on-prem не означает одинаковых требований к GPU: при внешних модельных маршрутах вычисления модели выполняет соответствующий получатель, а при локальных — инфраструктура заказчика.

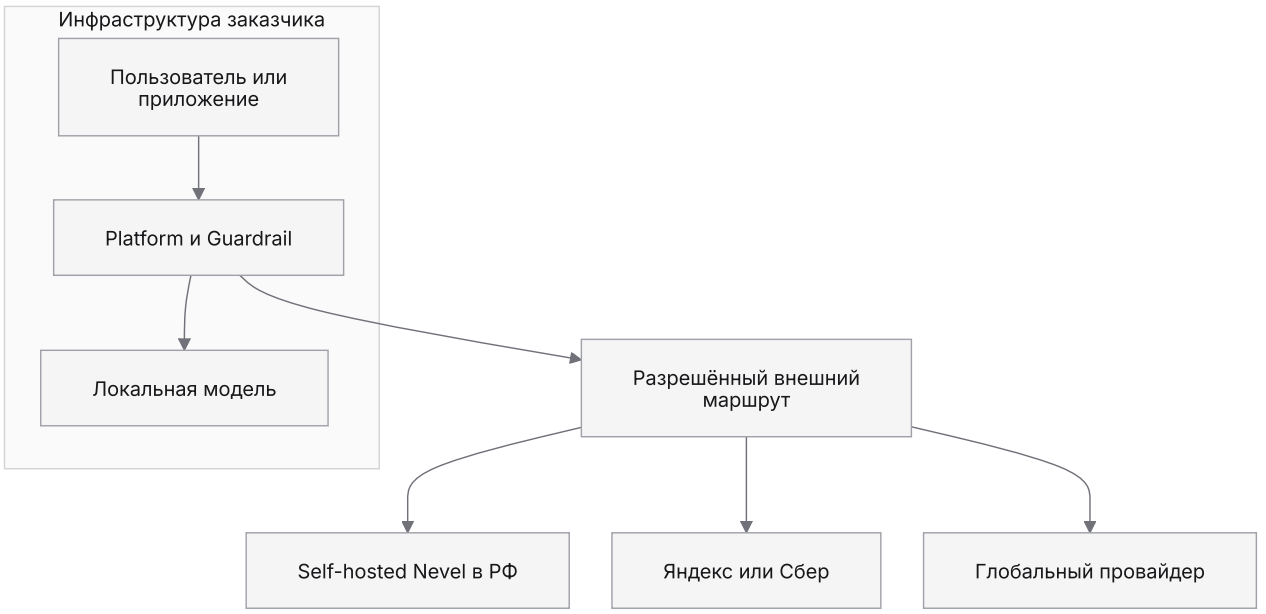
Выберите маршруты моделей

Вариант	Подключение из on-prem	Ключевой контроль
Локальная модель заказчика	Внутренний endpoint	Доступность нужных форматов и производительность внутренних ресурсов
Российский провайдер, например Яндекс или Сбер	Разрешённое исходящее API-соединение	Получатель вне инфраструктуры заказчика, категории данных и условия обработки
Self-hosted Nevel в РФ, например DeepSeek	Разрешённый endpoint инфраструктуры Nevel	Отличие этого экземпляра от API внешнего владельца модели
Глобальная модель	Разрешённая цепочка API, включая промежуточные сервисы при их наличии	Все получатели и география, политика Guardrail и отсутствие обхода

Выбор фиксируется для класса данных, пространства или приложения в пределах политики. Например, закрытый кадровый документ может обрабатываться только локально, общедоступный текст — глобальной моделью, а внутренний рабочий материал — разрешённым российским маршрутом. Названия категорий и конкретные разрешения задаёт организация; этот пример не является готовым юридическим разрешением.

СХЕМА 1

Маршруты модели из on-prem



Передача к Яндексу, Сберу или Nevel пересекает границу инфраструктуры заказчика. География фактической обработки проверяется отдельно. Место установки Platform само по себе не определяет, возникает ли трансграничная передача.

Подготовьте инфраструктуру

Определите приоритетные сценарии, одновременную нагрузку, размеры файлов и требования к ожиданию. Для локальной модели отдельно оцените вычисления и память; для внешней — соединения, лимиты и доступность провайдера. Поддержка метода API не делает возможности разных моделей одинаковыми: текст, изображения, инструменты и форматы ответа проверяются на выбранном экземпляре.

Подготовьте адреса сервисов, сертификаты, секреты и доступ эксплуатационной команды. Исходящие соединения разрешаются по конкретным назначениям. Базы и хранилище соответствий не открываются внешним модельным провайдерам: им передаётся подготовленный контекст по разрешённому API-маршруту.

Журналы, резервные копии, каналы поддержки и обновлений включаются в схему отдельно. On-prem не даёт автоматически разрешение на любой внешний мониторинг или выгрузку диагностического архива.

Порядок развёртывания

1. Зафиксировать место Platform, компоненты, модельные маршруты и матрицу ответственности.
2. Получить комплект установки для выбранной версии и подготовить инфраструктуру.
3. Развернуть компоненты и проверить их состояние, сертификаты и сервисные права.
4. Подключить выбранные локальные и внешние модели по разрешённым адресам.
5. Настроить категории данных, Guardrail, основные и резервные маршруты.
6. Создать пространства, участников и отдельные ключи тестовых приложений.
7. Принять рабочие сценарии, запреты, аудит и восстановление до переноса рабочих данных.

Конфигурация сохраняется в контролируемом процессе эксплуатации. Секреты не включаются в открытые инструкции. Для изменения маршрута фиксируются инициатор, прежнее и новое состояние и затронутые категории данных.

Приёмка маршрутизации

Проверьте один запрос для каждого разрешённого маршрута и по журналу подтвердите конечного и промежуточных получателей. Повторите сценарий с документом, которому этот маршрут запрещён. Наличие ключа глобального провайдера не должно обходить локальную метку.

Затем отключите основную модель. Запрос может использовать только заранее разрешённый для этой категории резерв; выбор отражается в аудите. Если подходящего резерва нет, возвращается контролируемый отказ. Клиентский SDK также проверяется на отсутствие собственного скрытого переключения.

Включите загрузку файла, OCR, поиск по знаниям, инструмент и сохранение артефакта: каждый этап может обращаться к отдельному обработчику. Отдельно проверьте отзыв доступа, защиту секретов и восстановление тестовой копии.

Изолированный профиль on-prem

Если организация требует, чтобы вся рабочая обработка оставалась внутри контура, используются только внутренние модели и вспомогательные сервисы. Внутри также находятся OCR, embeddings, инструменты, хранилища, журналы и резервные копии. Внешний поиск и облачная модель для таких задач исключаются политикой и сетевыми правилами.

Приёмка этого профиля проводится при запрещённой внешней обработке и охватывает обычные задачи и отказы. Недоступность локальной модели не открывает внешнюю альтернативу. Сетевая проверка подтверждает внутреннюю цепочку, а не только успешный ответ интерфейса.

Эксплуатация

Ответственность за Platform и внутренние компоненты отделяется от ответственности внешнего модельного провайдера. Изменение endpoint, модели, получателя, ключа или региона рассматривается как изменение маршрута и проходит затронутые проверки.

В тестовой среде проверяются обновления приложения, моделей, коннекторов и политик. Для изолированного профиля дополнительно определяется перенос пакетов внутрь контура. Порядок доступа поддержки, диагностические данные, каналы и откат фиксируются в регламенте. Подробнее — эксплуатация и сетевая защита.