Nevel

РАБОТА С ПЛАТФОРМОЙ

Модели и режимы работы

Как Авто выбирает модель, чем отличается от режима инструмента, как работает резервный выбор и где выполняется обработка.

Облако / В контуре

https://docs.nevel.sh/platform/models/

В Nevel отдельно выбираются модель обработки и режим задачи. Модель задаёт способ получения ответа; режим помогает начать конкретный сценарий — поиск, исследование, работу с документами, таблицами, презентациями или медиа. Для обычного запроса начните с **«Авто»**, а инструментальный режим включайте под требуемый результат.

Выбор модели

Рядом с полем ввода находится кнопка с названием выбранной модели. Нажмите её: откроется меню **«Модель»**. Выберите доступный пункт; галочка показывает текущий выбор. Названия конкретных провайдерских моделей берутся из каталога установки, поэтому не ориентируйтесь на старый список из презентации.

Конкретные семейства моделей и порядок обновления — в каталоге моделей.

Вариант	Подсказка интерфейса	Когда попробовать
Авто	Баланс качества и скорости, подходит для большинства задач	Первый запуск сценария, смешанные задачи
Быстрый	Отвечает быстро, подходит для простых вопросов	Короткий ответ, небольшой пересказ, простая правка
Думающий	Думает дольше для лучших ответов	Сложное сравнение, анализ условий, последовательное решение
Пишущий	Для создания текстов	Черновик письма, статья, редактура формулировок

Эти варианты помогают выбрать характер работы. Они не обещают одинаковые затраты, точность или время ответа для любых исходных данных. Длинный файл остаётся длинной задачей и в быстром режиме. Сложность результата зависит также от необходимых инструментов: создание файла включает дополнительные операции помимо написания текста.

Как работает «Авто»

«Авто» поручает Nevel выбрать модель для текущего обращения. Пользователю не нужно заранее сопоставлять каждую задачу с названием провайдера. В автоматическом сценарии работает отдельный маршрутизатор — Router. Он анализирует задачу, выбирает один вариант из доступного набора и передаёт выполнение выбранной модели.

Выбор происходит перед основным ответом. Он не означает одновременную отправку одного вопроса всем моделям или голосование между их ответами. Router получает текущее обращение и ограниченный настройкой фрагмент последних сообщений пользователя и ассистента. Поэтому существенные требования полезно повторить в самой задаче: тему, ожидаемый результат, необходимость краткости или глубокого разбора. Нельзя рассчитывать, что условие из начала длинного диалога обязательно попадёт именно в контекст выбора.

Порядок обработки:

1. Platform определяет доступные для выбора модели и действующую конфигурацию автоматического сценария.

2. Router получает недавний контекст и описание вариантов. Он оценивает область задачи, её сложность и специальные требования.
3. Router возвращает выбранную модель и краткое обоснование решения.
4. Platform проверяет, что выбранный идентификатор входит в предложенный набор. Придуманное название не становится новым подключением.
5. Основная модель выполняет задачу. Если ей нужны разрешённые инструменты, начинается их отдельная работа.

СХЕМА 1

Автоматический выбор модели



В отдельных конфигурациях «Авто» может быть связано с заранее заданной моделью без отдельного вызова Router. Поэтому само название в переключателе ещё не доказывает, что в каждом запросе выполнялся динамический выбор. Для технического разбора нужны сведения о фактически выбранной модели и конфигурации сценария.

По каким признакам выбирается модель

В первую очередь учитывается **область задачи**, затем сложность и дополнительные требования. В настройке маршрутизатора сопоставлены направления: программирование, математика и логика, общие задачи разной сложности. Один и тот же выбранный провайдер не закрепляется за человеком навсегда: следующее обращение может требовать другого способа обработки.

Признак запроса	Как влияет на выбор
Простое объяснение или короткая правка	Подходит вариант для несложных задач
Сравнение нескольких условий, последовательный разбор	Учитывается необходимость более глубокого анализа
Программирование	Различаются обычные изменения и сложные архитектурные задачи
Математика или логическая задача	Используется назначение для соответствующего направления
Просьба ответить быстро или кратко	Учитывается предпочтение скорости
Просьба подробно обосновать вывод	Учитывается требование глубины ответа

Фраза «ответь быстро» не отменяет сложность исходного материала и не гарантирует определённое время. Обоснование выбора — объяснение Router, а не доказательство, что модель объективно лучшая для любой такой задачи. Если результат важен, проверьте выводы по источникам и повторите задачу с явно выбранной моделью для сравнения.

Router, модель ответа и инструменты

У сложной задачи могут быть разные обработчики на разных этапах. Например, поиск получает материалы, основная модель связывает сведения в ответ, а генератор документа создаёт файл. Для внутренних инструментов могут быть настроены собственные модели. Выбор «Авто» относится к автоматическому сценарию ответа; он не означает, что все эти операции выполняет одна и та же модель.

Различайте три решения:

- **Подбор модели задачи:** какой вариант подходит запросу по области и сложности.
- **Маршрут по политике:** кому разрешено получить эти данные и где допустима обработка.
- **Модель инструмента:** какой обработчик выполняет поиск, создание изображения или другую специализированную операцию.

Предпочтение Router не расширяет права на данные или разрешённые направления передачи. Для корпоративного запроса проверяется также путь самого Router, поскольку он получает содержание, достаточное для выбора. Разрешение основной модели не делает автоматически разрешёнными маршрутизатор, поиск или генератор файла. Правила получателей описаны в архитектуре и границах доверия.

У маршрутизатора есть собственный вызов модели, время работы и учёт потребления. Поэтому автоматический выбор не следует считать бесплатной вычислительной операцией или гарантией уменьшения общей стоимости. Для оценки эффективности сравнивают завершённые задачи одинакового качества с учётом выбора модели, инструментов и повторов. Сам факт использования разных моделей на этапах не является измерением экономии.

Резервный выбор и диагностика

Если Router не может выбрать модель — например, недоступен, не успел ответить или вернул вариант вне доступного набора, — используется настроенный резервный выбор, **fallback**. Это возврат к предусмотренной модели, а не поиск произвольного провайдера. Резервный маршрут подчиняется ограничениям организации на получателей и данные.

Для разбора запроса администратору нужны выбранный пользователем режим, фактическая модель ответа, время обращения, признак резервного выбора и причина. В обработке автоматического выбора также учитываются модель самого Router и её потребление. Краткое обоснование решения и сведения о fallback относятся к технической диагностике; их наличие не означает отдельную кнопку просмотра в каждом пользовательском интерфейсе.

Если выбор кажется неудачным, действуйте последовательно:

1. Сохраните формулировку запроса и отметьте, что именно не подошло: глубина ответа, формат, работа с файлом или доступность инструмента.
2. Проверьте, был ли выбран «Авто» и какой результат фактически получен. Ошибка генерации файла отличается от ошибки выбора модели.
3. В новой ветке повторите запрос с конкретной разрешённой моделью и теми же материалами.
4. Для обращения в поддержку приложите время и идентификатор запроса, если он показан. Содержание закрытых документов передавайте только по принятому организацией каналу.

Неудачный ответ сам по себе не доказывает, что сработал fallback. И наоборот, резервный выбор не означает потерю диалога: это изменение обработчика текущего обращения. При изменении доступного каталога или правил администратор повторяет проверку типичных задач и резервного пути.

Изменение модели в диалоге

Можно выбрать другую модель для следующего сообщения. Уже написанный ответ от этого не изменится. Чтобы получить сравнимый вариант, повторите задачу с теми же источниками или создайте ветку чата до спорного ответа. Иначе вы сравните не только модели, но и разный накопленный контекст.

Удобный контрольный запрос для сравнения:

> По приложенному регламенту перечисли пять условий возврата. Для каждого укажи раздел документа. Если условие не найдено, напиши «в источнике не найдено». Не дополняй ответ общими знаниями.

Оцените полноту, точность ссылок и наличие неподтверждённых деталей. Не выбирайте модель только по уверенности тона: уверенная формулировка не является доказательством правильности ответа.

Режим задачи

Нажмите **«+»** у поля ввода. Кроме **«Загрузить файлы»**, доступны режимы:

- **«Поиск в Интернете»** — запрос с поиском внешних сведений.
- **«Детальное исследование»** — многоэтапное исследование с подготовкой отчёта.
- **«Документы»** — создание и изменение текстовых документов.
- **«Таблицы»** — работа с табличным файлом.
- **«Презентации»** — подготовка и изменение слайдов.
- **«Изображения»** и **«Видео»** — генерация соответствующих материалов.

На стартовом экране часть режимов также показана отдельными кнопками; оставшиеся находятся под **«Ещё»**. Выбранный режим отображается рядом с вводом. Уберите его нажатием на элемент режима, если задача изменилась. Одновременно выбирается один такой режим.

После отправки выбор режима сбрасывается. Для следующего сообщения проверьте поле ввода: если снова нужен явный инструментальный режим, выберите его повторно. При этом историю и уже созданные файлы диалог сохраняет. Сброс кнопки не означает исчезновение предыдущего документа.

Модель не заменяет постановку задачи

Выбор **«Пишущий»** сам по себе не задаёт формат DOCX, аудиторию или длину текста. Напишите это в запросе. Аналогично **«Думающий»** не включает автоматически доступ к закрытой базе, а **«Таблицы»** не определяет, какой лист нужно изменить.

Сформулируйте задачу четырьмя частями: источник, операция, результат и ограничения. Например: «По @План.xlsx на листе Продажи посчитай квартальные итоги. Добавь лист Сводка с формулами, исходный лист не меняй». Для презентации укажите аудиторию, число слайдов как желаемый объём и основную мысль; для поиска — период и тип источников.

Доступность моделей и инструментов

Значок замка означает недоступный выбор. Если коллега видит другой набор, сравните выбранную команду, роль и конфигурацию доступа. Модель, инструмент и источник данных проверяются отдельно: наличие модели для ответа не даёт автоматически доступ к интеграции или чужому файлу.

Текст запроса не выдаёт новые права. Разрешения сотрудника и его дирекции настраиваются через администрирование команды и действующую матрицу доступа. Если нужен конкретный инструмент, сначала проверьте его доступность, а затем сравнивайте модели, которые могут с ним работать.

Размещение платформы и маршруты моделей

Сначала определите, где работает Platform: в облаке Nevel или в инфраструктуре заказчика. Затем отдельно выберите, куда отправляется запрос модели. Внутренний адрес платформы не означает, что все модели локальные; облачный интерфейс не означает обязательное использование глобального провайдера.

Маршрут модели	Где выполняется обработка	Что учитывать
Собственные модели заказчика	В инфраструктуре заказчика	Адрес модели, доступность из Platform, поддерживаемые операции и вычислительную мощность
Российские провайдеры, например Яндекс и Сбер	В инфраструктуре выбранного российского провайдера	Разрешённый сервис, договорные условия и данные, передаваемые за пределы платформы
Глобальные модели	У выбранного глобального провайдера	Политику компании, допустимые данные и фактический маршрут подключения
Self-hosted DeepSeek Nevel в облаке	На серверах Nevel в РФ	Это модель, размещённая Nevel; она не становится собственной моделью заказчика

И облачная Platform, и Platform в контуре могут использовать разрешённые для них маршруты. Для подключения модели заказчика из облака нужен настроенный канал к её точке доступа; для внешнего провайдера из контура — разрешённое исходящее соединение. Выбор в интерфейсе не открывает сетевой доступ автоматически.

Название модели и её маршрут — разные сведения. Например, DeepSeek может означать модель в инфраструктуре заказчика, self-hosted модель Nevel или предложение провайдера. При выборе корпоративного сценария уточните именно точку выполнения и условия обработки, а не только семейство модели.

Изолированный профиль

Полностью закрытая установка — частный случай on-prem. В этом профиле модели, инструменты и источники работают внутри разрешённого периметра; внешние обращения отключены. Для веб-поиска, генерации медиа и интеграций нужны внутренние средства либо эти операции будут недоступны в изолированном сценарии.

При обычной установке в контуре российские и глобальные провайдеры могут использоваться, если их разрешает политика организации. Размещение платформы у заказчика не означает, что запрос целиком обработан у заказчика: для этого все компоненты маршрута должны работать в его инфраструктуре.

Проверьте модель на своём сценарии

Не все модели одинаково поддерживают изображения, инструменты и структурированный вывод. Если текстовый запрос работает, а запрос с файлом — нет, уменьшите пример до одного вложения и проверьте возможности выбранной модели. Смена провайдера может изменить формат ответа, доступный контекст и поддерживаемые операции.

После подключения или замены модели выполните короткий текстовый запрос, запрос по файлу и нужное инструментальное действие. Проверьте не только текст ответа, но и фактически созданный документ или выполненную операцию. Размер поддерживаемого контекста и возможности обработки определяются выбранной моделью и конфигурацией сценария.

Типичные затруднения

«Модель не выбрана» — дождитесь загрузки каталога и выберите доступный вариант. Если список не появился, обновите страницу и проверьте доступ к установке. **«Модель не найдена»** — выбранное имя могло измениться; используйте текущий каталог. При длинном ожидании посмотрите, идёт ли создание файла или работа инструмента: это отдельный этап. Если видна ошибка, сначала исправьте её причину и только затем сравнивайте модели.