РУКОВОДСТВО ПОЛЬЗОВАТЕЛЯ

LLMaaS платформа ВПРОД

КОНСОЛЬ АДМИНИСТРАТОРА

Обзор

Инференс
Модели
Хранилище
Рабочие нагрузки
Парк GPU
Бенчмарки

Тенанты
Организации
Федерация

Биллинг
Потребление
Цены

TA Вы вошли как
tester-admin@example.com
platform-admin

Обзор

↻ Обновить

Модели	Рабочие нагрузки	GPU-узлы	Организации
1	1	2	5
1 в витрине · 0 загружается	1 опубликованы · 0 с проблемами	1 из 2 карт заняты	0 заморожено

Обслуживание · 1 из 24

Платформа одним взглядом

Четыре счётчика показывают, что сейчас работает: модели, рабочие нагрузки, GPU-узлы и организации. Ниже — список того, что требует внимания; каждая строка ведёт туда, где это можно исправить.

Пропустить тур Назад **Далее**



Токены всего · 14 дней
190 596

Расходы RUB
0,778 ₽

Занятость GPU по узлам

занято 1 из 2 GPU

k8s-wrk-gpu-0	0 / 1
k8s-wrk-gpu-1	1 / 1

Топ организаций за 30 дней

Открыть раздел

Benchmark	127 783	0,471 ₽
Test	34 441	0,168 ₽
ООО «Демо»	27 899	0,139 ₽
VPROD	194	0,00 ₽

Обзор

Инференс
Модели
Хранилище
Рабочие нагрузки
Парк GPU
Бенчмарки

Тенанты
Организации
Федерация

Биллинг
Потребление
Цены

Модели 1

Модели, доступные для развёртывания и для витрины клиентов

+ Добавить модель

Имя, отображаемое имя или репозиторий Hugging Face

все модели

Модель	Репозиторий Hugging Face	Тип	Витрина	Веса
qwen3-0-6b Qwen3-0.6B	Qwen/Qwen3-0.6B	llm	◉ в витрине	✓ Готова 1,5 GB

Обслуживание · 2 из 24

Добавить модель

Здесь вы добавляете новую модель в каталог: название, откуда взять веса на Hugging Face и движок. Пока модели нет в каталоге, запустить её нельзя.

Пропустить тур | Назад | Далее

Обзор

Инференс
Модели
Хранилище
Рабочие нагрузки
Парк GPU
Бенчмарки

Тенанты
Организации
Федерация

Биллинг
Потребление
Цены

Модели 1

Модели, доступные для развёртывания и для витрины клиентов

+ Добавить модель

Имя, отображаемое имя или репозиторий Hugging Face

все модели

1 из 1

Модель	Репозиторий Hugging Face	Тип	Витрина	Веса	Запущенные экземпляры	Изменена
qwen3-0-6b Qwen3-0.6B	Qwen/Qwen3-0.6B	llm	◉ в витрине	✓ Готова 1,5 GB	1	29.09.2026 …

Обслуживание · 3 из 24

Статус модели

В этой колонке видно, скачаны ли файлы модели. Запустить можно только модель со статусом «Готова», остальные ещё загружаются.

Пропустить тур Назад Далее

Хранилище моделей 1

Тома, на которых лежат скачанные веса моделей

PVC	Том (PV)	Веса моделей	Размер тома
platform-store-pvc	—	1,5 GB	—

Обслуживание · 4 из 24

Где лежат веса

Скачанные модели делят один том: модель загружается один раз, а все нагрузки подключают её только на чтение. Хранилище, которое ещё ждёт кластер, показано как «ожидает».

Пропустить тур | Назад | Далее

Обзор

Инференс
Модели
Хранилище
Рабочие нагрузки
Парк GPU
Бенчмарки

Тенанты
Организации
Федерация

Биллинг
Потребление
Цены

Рабочие нагрузки 1

Инференс: состояние, публикация, рантайм-метрики

+ Новая рабочая нагрузка

Поиск по нагрузке или модели | все состояния | все | ↻ Обновить

vprod-demo
k8s-wrk-gpu-1
● готова

Модель	Публикация	Рестарты
Qwen3-0.6B	опубликован	0

Текущие метрики
KV 0% · в работе / в очереди 0 / 0
TTFT p95 — мс · ITL p95 — мс · 0 токенов/с

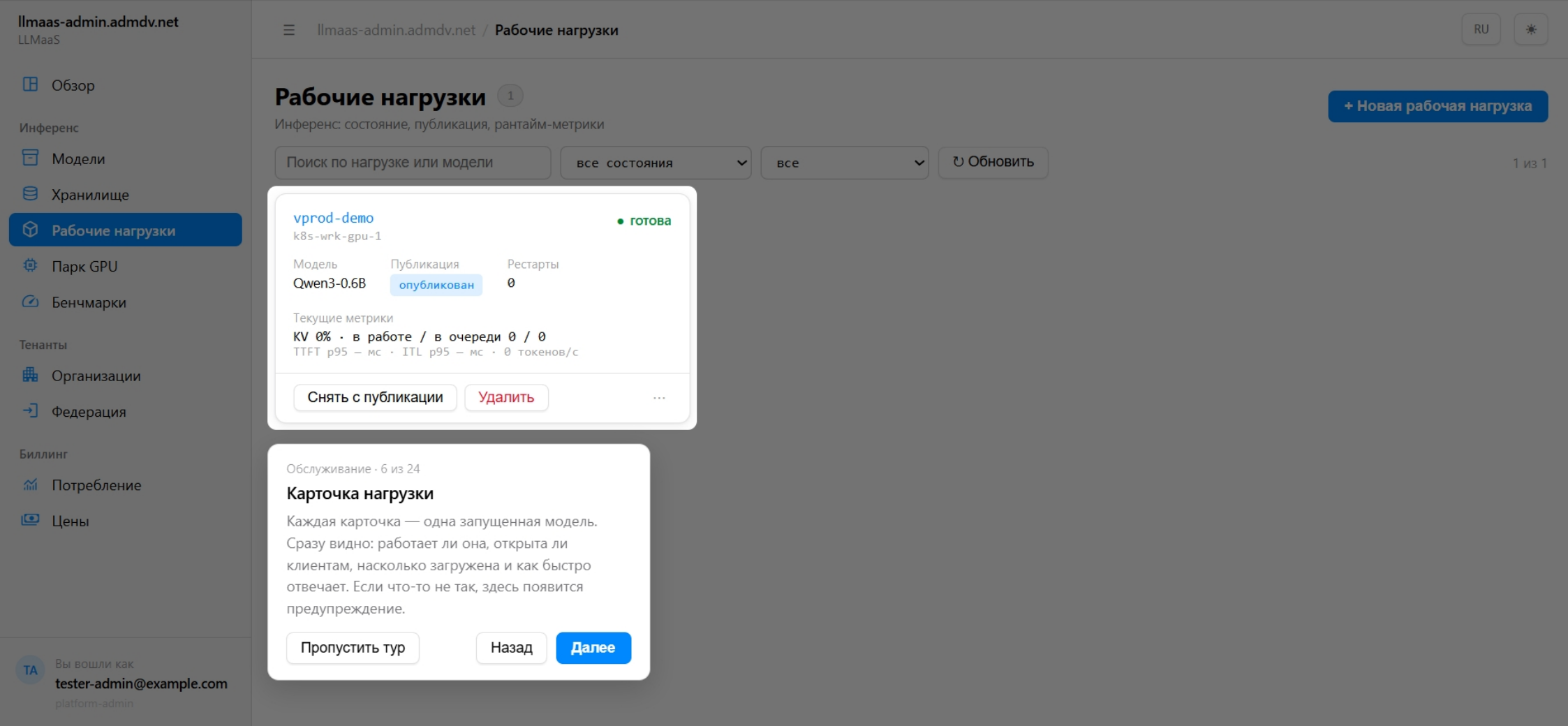
Снять с публикации | Удалить | ⋯

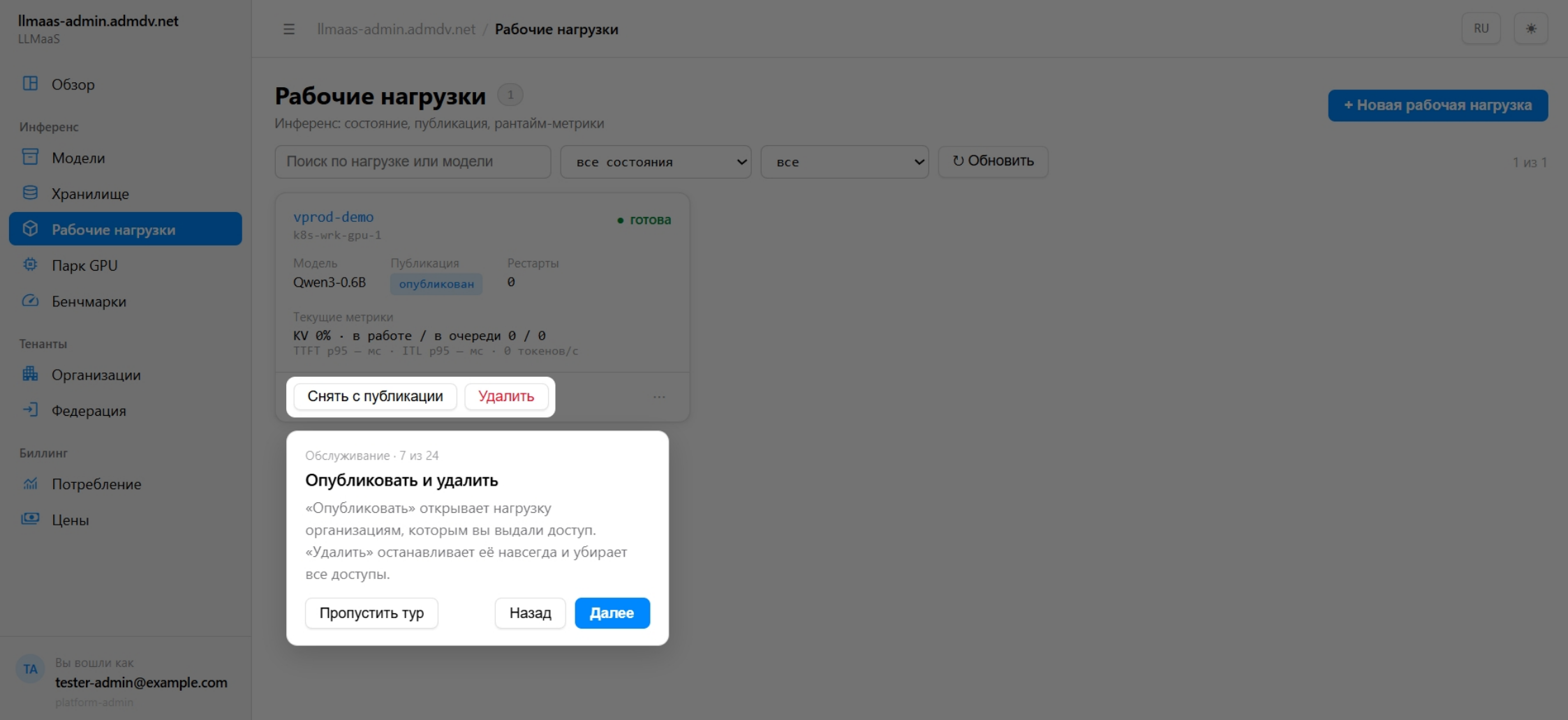
Обслуживание · 5 из 24

Запустить модель

Рабочая нагрузка — это запущенная копия модели на ваших GPU. Нажмите сюда, чтобы запустить новую: выберите модель, дайте ей имя и укажите, где она будет работать.

Пропустить тур | Назад | Далее


llmaas-admin.admdv.net
LLMaaS
Обзор
Инференс
Модели
Хранилище
Рабочие нагрузки
Парк GPU
Бенчмарки
Тенанты
Организации
Федерация
Биллинг
Потребление
Цены
TA
Вы вошли как
tester-admin@example.com
platform-admin
llmaas-admin.admdv.net / Рабочие нагрузки
RU
Рабочие нагрузки
1
Инференс: состояние, публикация, рантайм-метрики
+ Новая рабочая нагрузка
Поиск по нагрузке или модели
все состояния
все
↻ Обновить
1 из 1
vprod-demo
k8s-wrk-gpu-1
готова
Модель
Qwen3-0.6B
Публикация
опубликован
Рестарты
0
Текущие метрики
KV 0% · в работе / в очереди 0 / 0
TTFT p95 — мс · ITL p95 — мс · 0 токенов/с
Снять с публикации
Удалить
Обслуживание · 6 из 24
Карточка нагрузки
Каждая карточка — одна запущенная модель. Сразу видно: работает ли она, открыта ли клиентам, насколько загружена и как быстро отвечает. Если что-то не так, здесь появится предупреждение.
Пропустить тур
Назад
Далее


llmaas-admin.admdv.net
LLMaaS
Обзор
Инференс
Модели
Хранилище
Рабочие нагрузки
Парк GPU
Бенчмарки
Тенанты
Организации
Федерация
Биллинг
Потребление
Цены
TA
Вы вошли как
tester-admin@example.com
platform-admin
llmaas-admin.admdv.net / Рабочие нагрузки
RU
Рабочие нагрузки
1
Инференс: состояние, публикация, рантайм-метрики
+ Новая рабочая нагрузка
Поиск по нагрузке или модели
все состояния
все
Обновить
1 из 1
vprod-demo
k8s-wrk-gpu-1
готова
Модель
Qwen3-0.6B
Публикация
опубликован
Рестарты
0
Текущие метрики
KV 0% · в работе / в очереди 0 / 0
TTFT p95 — мс · ITL p95 — мс · 0 токенов/с
Снять с публикации
Удалить
Обслуживание · 7 из 24
Опубликовать и удалить
«Опубликовать» открывает нагрузку организациям, которым вы выдали доступ. «Удалить» останавливает её навсегда и убирает все доступы.
Пропустить тур
Назад
Далее

Обзор

Инференс
Модели
Хранилище
Рабочие нагрузки
Парк GPU
Бенчмарки

Тенанты
Организации
Федерация

Биллинг
Потребление
Цены

TA Вы вошли как
tester-admin@example.com
platform-admin

vprod-demo • готова

Снять с публикации | Переконфигурировать | Бенчмарк | ⊗ Удалить

Qwen3-0.6B · k8s-wrk-gpu-1 · http://vprod-demo-predictor.llmaas.svc.cluster.local

Обзор | Метрики | Недавние события | Логи | Недавние события кластера

Цепочка развёртывания

CR → KServe → маршрут → под, по графу объектов кластера

- ✓ **CR** ок
- ✓ **KServe** ок
- ✓ **Маршрут** ок
- ✓ **Под** Работает

Фаза пода	Работает
Контейнер	running
Наблюдение	29.09.2026

Обслуживание · 8 из 24

Как она работает

Здесь вся цепочка, от которой зависит нагрузка: от запроса до узла, на котором она запущена. Сломанное звено подсвечено. На других вкладках — метрики, события и логи.

Пропустить тур | Назад | Далее

Рантайм

Движок	vllm
GPU	1
Grace period	300s
Создана	24.09.2026


```
{
  "engine": "vllm",
  "gpuCount": 1,
  "tolerations": [
    {
      "effect": "NoSchedule",
      "key": "usage",
      "value": "gpu"
    }
  ],
  "vllm": {
    "maxModelLen": 2048
  }
}
```

Обзор

Инференс
Модели
Хранилище
Рабочие нагрузки
Парк GPU
Бенчмарки

Тенанты
Организации
Федерация

Биллинг
Потребление
Цены

TA Вы вошли как
tester-admin@example.com
platform-admin

Обслуживание · 9 из 24

Ваши GPU-узлы

Одна карточка — один GPU-узел: занятые GPU, занятая память, загрузка и температура. «Заполнен» — свободных GPU для новых нагрузок нет.

Пропустить тур | Назад | Далее

↻ Обновить инвентарь

одной VRAM по узлам, та же оценка, что в предпроверке при создании нагрузки.

все вердикты

k8s-wrk-gpu-0 ○ свободен
NVIDIA-RTX-PRO-2000-Blackwell × 1

Занято GPU	Занято VRAM
0 / 1	0 / 15 GB
Загрузка	**Температура**
0%	35°C

Sharing: none usage=gpu:NoSchedule

k8s-wrk-gpu-1 ● заполнен
NVIDIA-RTX-PRO-2000-Blackwell × 1

Свободных GPU нет — новые нагрузки сюда не встанут

Занято GPU	Занято VRAM
1 / 1	15 / 15 GB
Загрузка	**Температура**
0%	32°C

Sharing: none usage=gpu:NoSchedule

Парк GPU 2

Узлы, карты и оценка размещения модели · занято 1 / 2 GPU

↻ Обновить инвентарь

Оценщик размещения

Где может запуститься модель каталога: вес против свободной VRAM по узлам, та же оценка, что в предпроверке при создании нагрузки.

выберите модель

все вердикты

Обслуживание · 10 из 24

Поместится ли модель?

Выберите модель и увидите по каждому узлу, поместится ли она, ещё до запуска. Носит информационный характер.

Пропустить тур | Назад | Далее

k8s-wrk-gpu-1 • заполнен

NVIDIA-RTX-PRO-2000-Blackwell × 1

Свободных GPU нет — новые нагрузки сюда не встанут

Занято GPU	Занято VRAM
1 / 1	15 / 15 GB
Загрузка	Температура
0%	32°C

Sharing: none | usage=gpu:NoSchedule

Обзор

Инференс
Модели
Хранилище
Рабочие нагрузки
Парк GPU
Бенчмарки

Тенанты
Организации
Федерация

Биллинг
Потребление
Цены

Бенчмарки 1

Эталонные замеры производительности по моделям и железу

▶Запустить бенчмарк

Обслуживание · 11 из 24

Измерить скорость

Запустите бенчмарк, чтобы узнать, как быстро модель отвечает на вашем оборудовании под разной нагрузкой. Удачный прогон можно сохранить как эталон и сравнивать с ним следующие.

Пропустить тур Назад Далее

всё железо

Рантайм	Промпт, токенов	Пик, токенов/с	Под SLO	TTFT p95	Обновлён	Заметка
vllm-0.25.1	512	611 c=8	611 · c=8	253 мс	24.09.2026	demo

Запущен	Модель	Профиль нагрузки	Статус	Уровень	Запросы
2026-09-29 11:40	qwen3-0-6b	быстрый	завершён	4 / 4	80 / 80
2026-09-24 15:49	qwen3-0-6b	быстрый	завершён Эталон	4 / 4	80 / 80
2026-09-18 11:40	qwen3-0-6b-291	быстрый	завершён	4 / 4	80 / 80

Организации 5

Тенанты и их потребление

Непривязанные пользователи 1

+ Новая организация

	Люди	Команды	Состояние	Токены за 30 дней	Траты за 30 дней
ООО «Демо» demo	4	3	Активна		
audit-tesst audit-tesst	0	2	Активна		
Test test	1	3	Активна		
Benchmark platform-benchmark	0	1	Активна	127 783	0,471 ₽
VPROD vprod	2	3	Активна	194	0,00 ₽

Организации и доступ · 12 из 24

Создать организацию

Организация — это ваш клиент. Задайте название, при желании бюджет и лимиты запросов. У каждой организации сразу есть команда по умолчанию, куда попадают её люди.

Пропустить тур | Назад | Далее

Обзор

Инференс
Модели
Хранилище
Рабочие нагрузки
Парк GPU
Бенчмарки

Тенанты
Организации
Федерация

Биллинг
Потребление
Цены

Организации 5

Тенанты и их потребление

Непривязанные пользователи 1

+ Новая организация

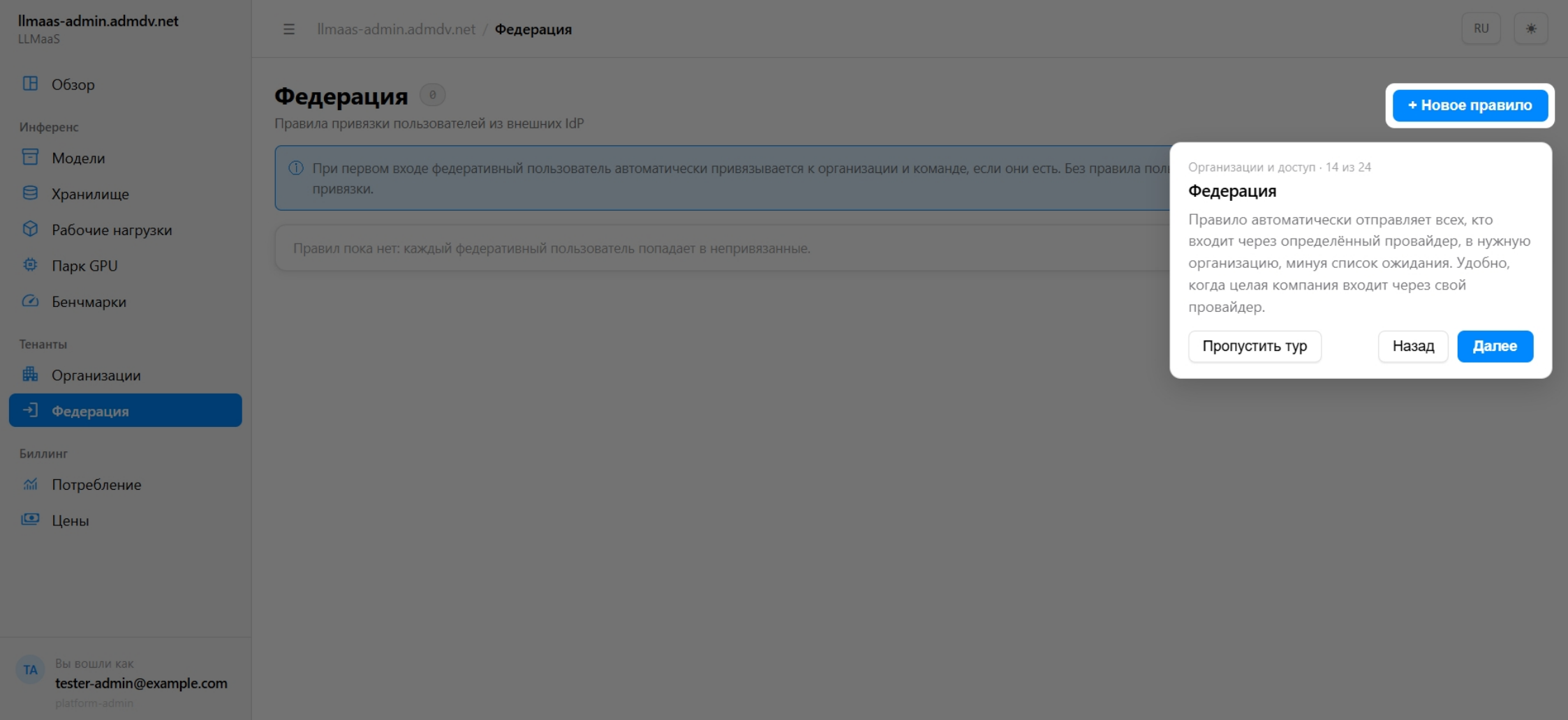
ООО «Демо» `demo`	Люди 4	Команды 3	Состояние • Активна		9 ₽
audit-tesst `audit-tesst`	Люди 0	Команды 2	Состояние • Активна		–
Test `test`	Люди 1	Команды 3	Состояние • Активна		8 ₽
Benchmark `platform-benchmark`	Люди 0	Команды 1	Состояние • Активна	127 783	0,471 ₽
VPROD `vprod`	Люди 2	Команды 3	Состояние • Активна	Токены за 30 дней 194	Траты за 30 дней 0,00 ₽

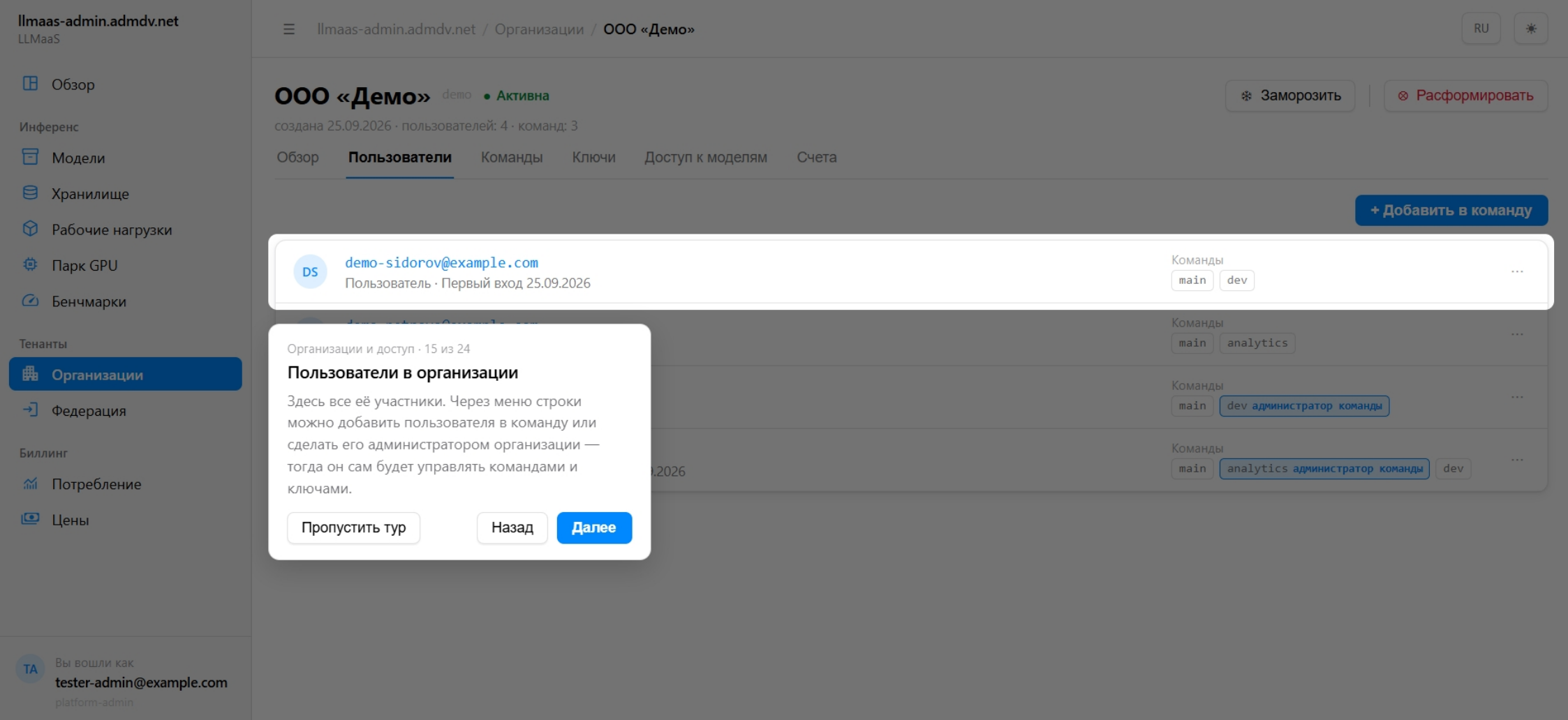
Организации и доступ · 13 из 24

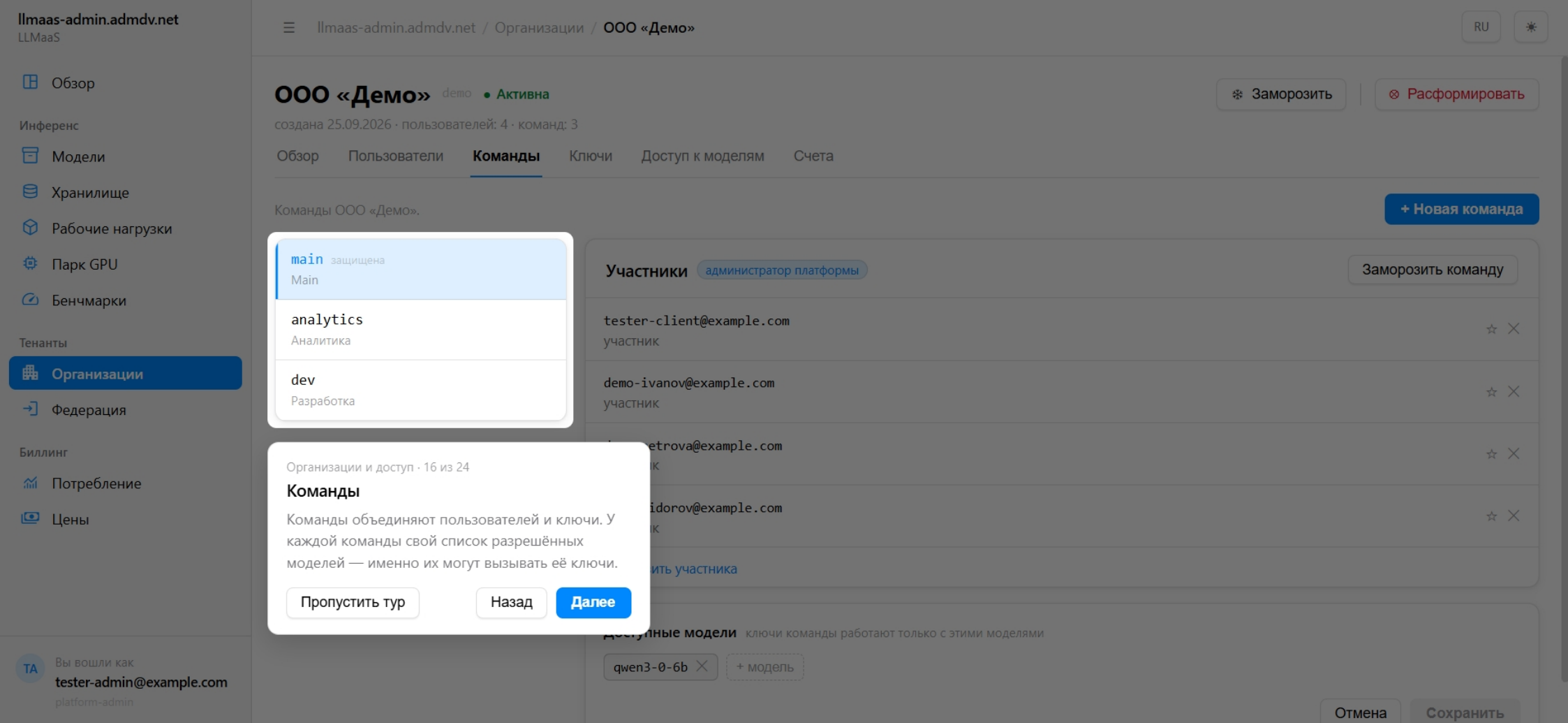
Кто ждёт назначения

Здесь те, кто вошёл в систему, но ещё не относится ни к одной организации. Назначьте пользователя в организацию, и он сразу получит доступ.

Пропустить тур | Назад | Далее


llmaas-admin.admdv.net
LLMaaS
Обзор
Инференс
Модели
Хранилище
Рабочие нагрузки
Парк GPU
Бенчмарки
Тенанты
Организации
Федерация
Биллинг
Потребление
Цены
TA
Вы вошли как
tester-admin@example.com
platform-admin
llmaas-admin.admdv.net / Федерация
RU
Федерация
0
Правила привязки пользователей из внешних IdP
+ Новое правило
При первом входе федеративный пользователь автоматически привязывается к организации и команде, если они есть. Без правила пол
привязки.
Правил пока нет: каждый федеративный пользователь попадает в непривязанные.
Организации и доступ · 14 из 24
Федерация
Правило автоматически отправляет всех, кто входит через определённый провайдер, в нужную организацию, минуя список ожидания. Удобно, когда целая компания входит через свой провайдер.
Пропустить тур
Назад
Далее


llmaas-admin.admdv.net
LLMaaS
Обзор
Инференс
Модели
Хранилище
Рабочие нагрузки
Парк GPU
Бенчмарки
Тенанты
Организации
Федерация
Биллинг
Потребление
Цены
TA
Вы вошли как
tester-admin@example.com
platform-admin
llmaas-admin.admdv.net / Организации / ООО «Демо»
RU
ООО «Демо»
demo
Активна
создана 25.09.2026 · пользователей: 4 · команд: 3
Заморозить
Расформировать
Обзор
Пользователи
Команды
Ключи
Доступ к моделям
Счета
+ Добавить в команду
DS
demo-sidorov@example.com
Пользователь · Первый вход 25.09.2026
Команды
main
dev
Команды
main
analytics
Команды
main
dev администратор команды
Команды
main
analytics администратор команды
dev
Организации и доступ · 15 из 24
Пользователи в организации
Здесь все её участники. Через меню строки можно добавить пользователя в команду или сделать его администратором организации — тогда он сам будет управлять командами и ключами.
Пропустить тур
Назад
Далее


llmaas-admin.admdv.net
LLMaaS
Обзор
Инференс
Модели
Хранилище
Рабочие нагрузки
Парк GPU
Бенчмарки
Тенанты
Организации
Федерация
Биллинг
Потребление
Цены
TA
Вы вошли как
tester-admin@example.com
platform-admin
llmaas-admin.admdv.net / Организации / ООО «Демо»
RU
ООО «Демо»
demo
Активна
создана 25.09.2026 · пользователей: 4 · команд: 3
Заморозить
Расформировать
Обзор
Пользователи
Команды
Ключи
Доступ к моделям
Счета
Команды ООО «Демо».
+ Новая команда
main
защищена
Main
analytics
Аналитика
dev
Разработка
Участники
администратор платформы
Заморозить команду
tester-client@example.com
участник
demo-ivanov@example.com
участник
Организации и доступ · 16 из 24
Команды
Команды объединяют пользователей и ключи. У каждой команды свой список разрешённых моделей — именно их могут вызывать её ключи.
Пропустить тур
Назад
Далее
qwen3-0-6b
+ модель
Отмена
Сохранить

Обзор

Инференс
Модели
Хранилище
Рабочие нагрузки
Парк GPU
Бенчмарки

Тенанты
Организации
Федерация

Биллинг
Потребление
Цены

ООО «Демо» demo • Активна

❄ Заморозить | ⊗ Расформировать

создана 25.09.2026 · пользователей: 4 · команд: 3

Обзор | Пользователи | Команды | **Ключи** | Доступ к моделям | Счета

Ключ принадлежит команде и наследует её список моделей. Владелец — участник, к которому ключ привязан, либо создавший его.

все команды ▾ + Создать ключ

Название	Ключ	Команда	Владелец	RPM / TPM	Бюджет	Истекает	Состояние
backend-service командный	…ENGg	dev	demo-ivanov@example.com	– / –	без ограничения	бессрочно	• Активен ⋯
petrova-jupyter личный	…yWUg	analytics	demo-petrova@example.com	– / –	без ограничения	бессрочно	• Активен ⋯
reports-pipeline командный	…vVTQ	analytics	tester-client@example.com	– / –	без ограничения	бессрочно	• Активен ⋯
sidorov-notebook			demo-sidorov@example.com	– / –	без ограничения	бессрочно	• Активен ⋯

…пуска отзовите ключ и создайте заново. 4 из 4

Организации и доступ · 17 из 24

API-ключи

Все ключи организации в одном месте. Ключ принадлежит команде или пользователю, у него свои лимиты и срок действия. Секрет показывается один раз при создании; здесь виден только префикс, и ключ можно отозвать.

Пропустить тур | Назад | **Далее**

Обзор

Инференс
Модели
Хранилище
Рабочие нагрузки
Парк GPU
Бенчмарки

Тенанты
Организации
Федерация

Биллинг
Потребление
Цены

ООО «Демо» demo • Активна

❄ Заморозить | ⊗ Расформировать

создана 25.09.2026 · пользователей: 4 · команд: 3

Обзор Пользователи Команды Ключи **Доступ к моделям** Счета

ⓘ Доступы в два слоя: слой 1 — модель выдана организации, слой 2 — командам внутри неё. Команда не может получить больше, чем выдано организации.

Модель	Организация	main	analytics	dev
qwen3-0-6b llm · Qwen3-0.6B	✓	✓	✓	✓

✓ выдан · — не выдан · · нет доступа организации

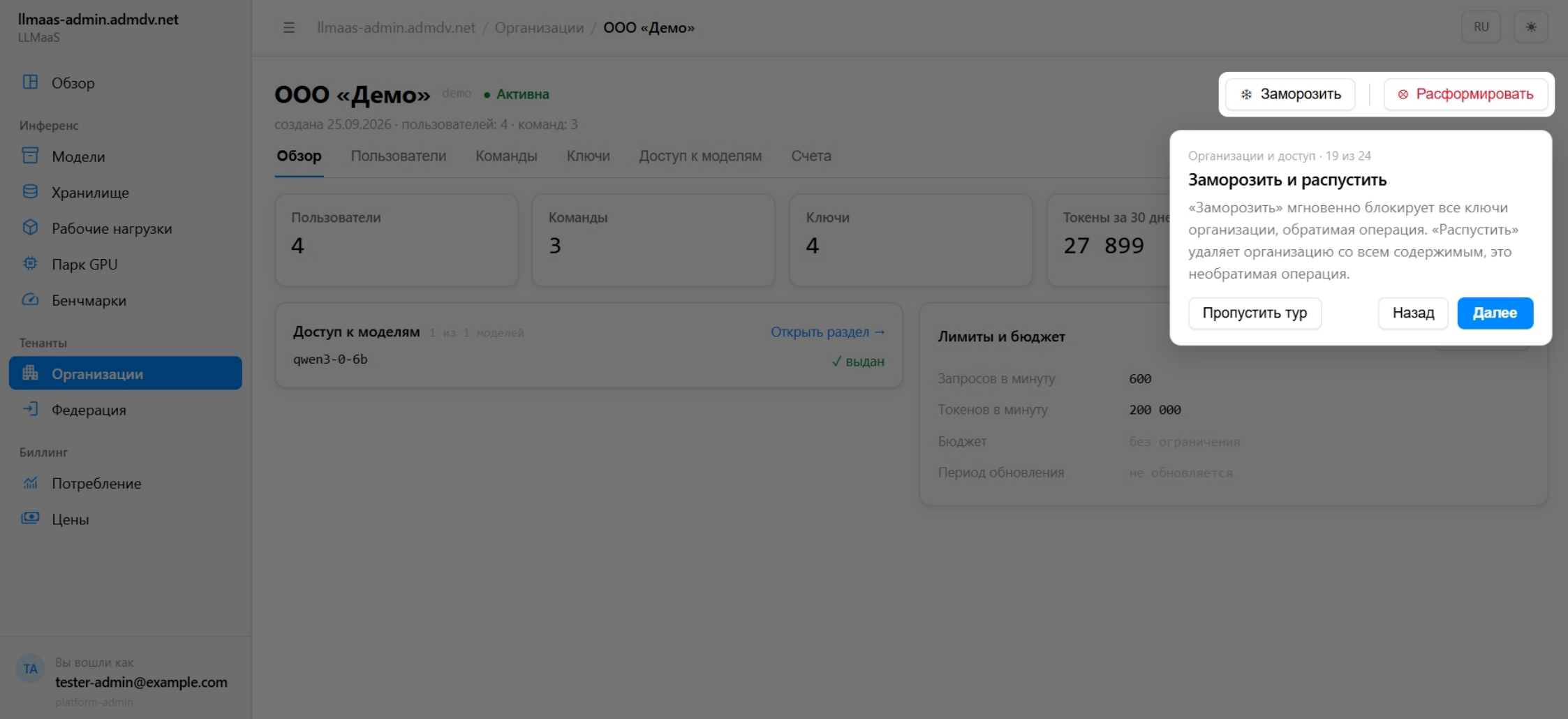
1 из 1 моделей

Организации и доступ · 18 из 24

Кому какие модели доступны

Первая колонка — какие модели вообще доступны организации. Остальные — что из них получает каждая команда. Команда не может получить больше, чем её организация.

Пропустить тур | Назад | **Далее**


llmaas-admin.admdv.net
LLMaaS
Обзор
Инференс
Модели
Хранилище
Рабочие нагрузки
Парк GPU
Бенчмарки
Тенанты
Организации
Федерация
Биллинг
Потребление
Цены
TA
Вы вошли как
tester-admin@example.com
platform-admin
llmaas-admin.admdv.net / Организации / ООО «Демо»
RU
ООО «Демо»
demo
Активна
создана 25.09.2026 · пользователей: 4 · команд: 3
Заморозить
Расформировать
Обзор
Пользователи
Команды
Ключи
Доступ к моделям
Счета
Пользователи
4
Команды
3
Ключи
4
Токены за 30 дне
27 899
Доступ к моделям
1 из 1 моделей
Открыть раздел →
qwen3-0-6b
✓ выдан
Лимиты и бюджет
Запросов в минуту
600
Токенов в минуту
200 000
Бюджет
без ограничения
Период обновления
не обновляется
Организации и доступ · 19 из 24
Заморозить и распустить
«Заморозить» мгновенно блокирует все ключи организации, обратимая операция. «Распустить» удаляет организацию со всем содержимым, это необратимая операция.
Пропустить тур
Назад
Далее

Потребление

Потребление в разрезе организаций, команд, ключей и моделей

все организации | С 16.09.2026 | По 29.09.2026 | по дням | ↓ CSV

Группировать по: Организация, Модель, Команда, Пользователь, Ключ — Команда, пользователь и ключ доступны только когда выбрана конкретная организация.

Биллинг · 20 из 24

Что было использовано

Выберите период, организацию или модель и посмотрите, сколько токенов и запросов прошло и сколько это стоило. Таблицу можно выгрузить в CSV.

Пропустить тур | Назад | Далее

Расходы
0,778 ₽

Строк
18

Дата	Организация	Модель	Входные токены	Выходные токены	Всего токенов	Запросы	Расходы
		qwen3-0-6b	36 207	10 880	47 087	85	0,235 ₽
		qwen3-0-6b	10	228	238	1	0,001 ₽
25.09.2026	ООО «Демо»	qwen3-0-6b	4 185	23 714	27 899	150	0,139 ₽
24.09.2026	Benchmark	qwen3-0-6b	36 269	10 880	47 149	85	0,236 ₽
24.09.2026	Test	gpt-4o	0	0	0	1	–
24.09.2026	Test	qwen3-0-6b	3 633	29 634	33 267	232	0,166 ₽
18.09.2026	01a0b28c-7bb3-7c0a-8ba1-ae24d818e2d6	qwen3-0-6b-291	15	16	31	1	0,00 ₽

Обзор

Инференс
Модели
Хранилище
Рабочие нагрузки
Парк GPU
Бенчмарки

Тенанты
Организации
Федерация

Биллинг
Потребление
Цены

Цены

+ Задать цену

Цены | Валюты | Организации

Модель	Тип токенов	RUB за 1M токенов	Установлена	
qwen3-0-6b	Входные	5,00 ₽	22.09.2026	Изменить
qwen3-0-6b	Выходные	5,00 ₽	22.09.2026	Изменить
qwen3-0-6b-291	Входные	5,00 ₽	22.09.2026	Изменить
qwen3-0-6b-291	Выходные	5,00 ₽	22.09.2026	Изменить

Биллинг · 21 из 24

Цены

Цена каждой модели за миллион токенов, отдельно для входящих и исходящих, в каждой валюте. При изменении старая цена остаётся в истории, поэтому прошлые счета не меняются.

Пропустить тур | Назад | Далее

Обзор

Инференс
Модели
Хранилище
Рабочие нагрузки
Парк GPU
Бенчмарки

Тенанты
Организации
Федерация

Биллинг
Потребление
Цены

Цены

Цены | **Валюты** | Организации

+ Добавить валюту

	Добавлена
RUB	22.09.2026

Биллинг · 22 из 24

Валюты

Добавьте валюту, чтобы выставлять в ней счета. У всех цен сразу появится значение в новой валюте. Валюту, которая уже есть в счетах, удалить нельзя.

Пропустить тур | Назад | **Далее**

Цены

Цены Валюты **Организации**

Организация

— выберите организацию —

Биллинг не проверяет идентификатор организации и молча примет опечатку.

Биллинг · 23 из 24

Особые условия

Выберите организацию, чтобы задать ей собственные цены вместо общих. Её счета — на соседней вкладке.

Пропустить тур Назад **Далее**

Обзор

Инференс
Модели
Хранилище
Рабочие нагрузки
Парк GPU
Бенчмарки

Тенанты
Организации
Федерация

Биллинг

Обзор

↻ Обновить

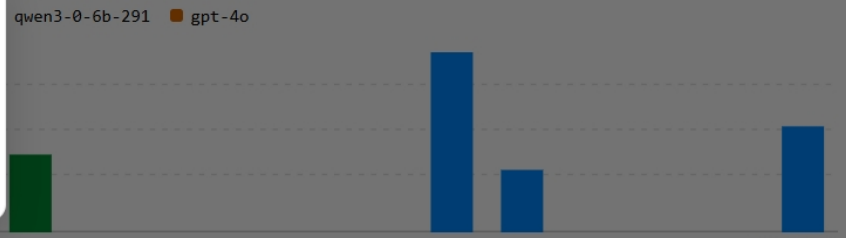
Модели	Рабочие нагрузки	GPU-узлы	Организации
1	1	2	5
1 в витрине · 0 загружается	1 опубликованы · 0 с проблемами	1 из 2 карт заняты	0 заморожено

Требуют внимания 0

Ничего не требует внимания

Токены по дням и моделям

Открыть раздел



Токены всего · 14 дней
190 596

Расходы RUB
0,778 ₽

Занятость GPU по узлам

занято 1 из 2 GPU

k8s-wrk-gpu-0	0 / 1
k8s-wrk-gpu-1	1 / 1

Топ организаций за 30 дней

Открыть раздел

Benchmark	127 783	0,471 ₽
Test	34 441	0,168 ₽
ООО «Демо»	27 899	0,139 ₽
VPROD	194	0,00 ₽

Биллинг · 24 из 24
Ваш аккаунт
Нажмите на своё имя, чтобы пройти тур заново или выйти.
Пропустить тур
Назад
Готово

КЛИЕНТСКИЙ ПОРТАЛ

ООО «Демо»
LLMaaS

Работа

- Обзор
- Песочница

Администрирование

- API-ключи
- Команды
- Потребление и биллинг

TE Вы вошли как
tester-client@example.com
customer

☰ ООО «Демо» / **Обзор**

RU ☀ ООО «Демо»

Обзор

Активные ключи
4 из 4 всего

Токены за 30 дней
27 899

Траты за 30 дней
0,139 ₽

Ваш портал · 1 из 6

Аккаунт одним взглядом

Сколько у вас ключей и — если включён биллинг — сколько токенов вы использовали и сколько это стоило за последние 30 дней.

Пропустить тур | Назад | **Далее**

Копировать

Модель
Qwen3-0.6B

Скопировать сниппет

```
curl https://llmaas-litellm.admdv.net/v1/chat/completions \
  -H "Authorization: Bearer $LLMAAS_API_KEY" \
  -H "Content-Type: application/json" \
  -d '{
    "model": "qwen3-0-6b",
    "messages": [{"role": "user", "content": "Hello"}],
    "stream": true
  }'
```

ⓘ Перед запуском сниппета экспортируйте ключ в LLMAAS_API_KEY.

ООО «Демо»
LLMaaS

Работа

- Обзор
- Песочница

Администрирование

- API-ключи
- Команды
- Потребление и биллинг

TE Вы вошли как
tester-client@example.com
customer

RU ООО «Демо»

Ваш портал · 2 из 6

Первый запрос

Всё для первого запроса: адрес, доступная модель и готовый пример кода. Подставьте свой API-ключ и запускайте.

Пропустить тур | Назад | Далее

Токены за 30 дней
27 899

Траты за 30 дней
0,139 ₽

Быстрый старт

OpenAI-совместимый эндпоинт

Базовый URL
https://llmaas-litellm.admdv.net Копировать

Модель
Qwen3-0.6B

curl | python

Скопировать сниппет

```
curl https://llmaas-litellm.admdv.net/v1/chat/completions \
  -H "Authorization: Bearer $LLMAAS_API_KEY" \
  -H "Content-Type: application/json" \
  -d '{
    "model": "qwen3-0-6b",
    "messages": [{"role": "user", "content": "Hello"}],
    "stream": true
  }'
```

ⓘ Перед запуском сниппета экспортируйте ключ в LLMAAS_API_KEY.

ООО «Демо»
LLMaaS

Работа
- Обзор
- Песочница

Администрирование
- API-ключи
- Команды
- Потребление и биллинг

Вы вошли как
tester-client@example.com
customer

☰ ООО «Демо» / API-ключи

RU ☀ ООО «Демо»

API-ключи 4

Ключи выдаются команде или пользователю и наследуют список моделей команды

+ Создать ключ

petrova-jupyter
личный · demo-petrova@example.com · команда: analytics

Модели	Токены	Расход	RPM
qwen3-0-6b	4 100 567 + 3 533 20 запр.	0,021 ₽	–

reports-pipeline
командный · команда: analytics

Модели	Токены	Расход	RPM	TPM	Бюджет	Истекает
qwen3-0-6b	8 381 1 251 + 7 130 45 запр.	0,042 ₽	–	–	без потолка	бессрочно

sidorov-notebook
личный · demo-sidorov@example.com · команда: dev

Модели	Токены	Расход	RPM	TPM	Бюджет	Истекает
qwen3-0-6b	4 552 678 + 3 874 25 запр.	0,023 ₽	–	–	без потолка	бессрочно

backend-service
командный · команда: dev

Модели	Токены	Расход	RPM	TPM	Бюджет	Истекает
qwen3-0-6b	10 866 1 689 + 9 177 60 запр.	0,054 ₽	–	–	без потолка	бессрочно

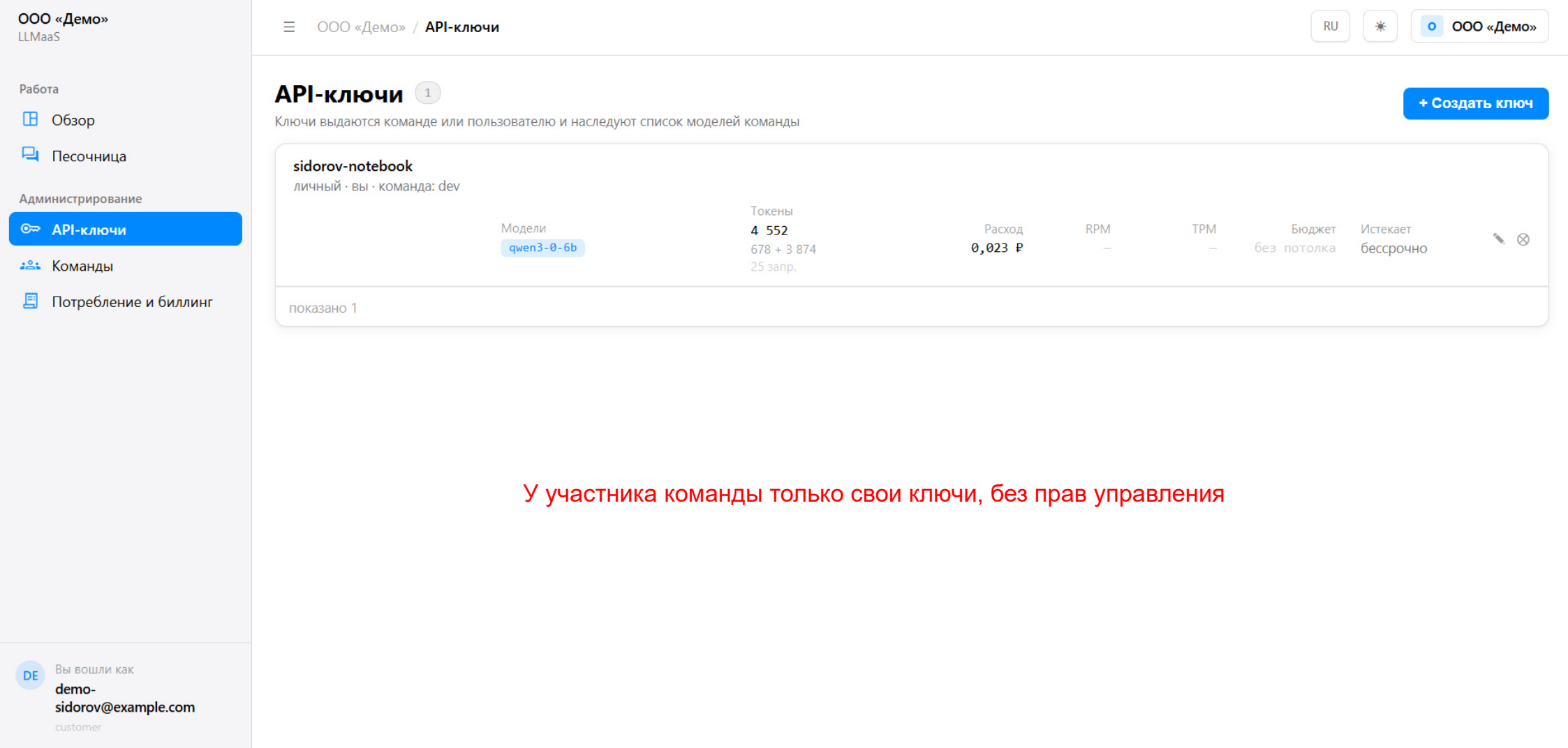
показано 4

Ваш портал · 3 из 6

Создать ключ

Ключ — это то, чем ваш код обращается к моделям. Создайте его здесь: секрет показывается только один раз, сразу скопируйте его.

Пропустить тур | Назад | Далее


ООО «Демо»
LLMaaS
Работа
Обзор
Песочница
Администрирование
API-ключи
Команды
Потребление и биллинг
DE
Вы вошли как
demo-sidorov@example.com
customer
ООО «Демо» / API-ключи
RU
ООО «Демо»
API-ключи
1
Ключи выдаются команде или пользователю и наследуют список моделей команды
+ Создать ключ
sidorov-notebook
личный · вы · команда: dev
Модели
qwen3-0-6b
Токены
4 552
678 + 3 874
25 запр.
Расход
0,023 ₽
RPM
–
TPM
–
Бюджет
без потолка
Истекает
бессрочно
показано 1
У участника команды только свои ключи, без прав управления

ООО «Демо»
LLMaaS

Работа
- Обзор
- Песочница

Администрирование
- API-ключи
- Команды
- Потребление и биллинг

Вы вошли как
tester-client@example.com
customer

ООО «Демо» / Песочница

RU

ООО «Демо»

Песочница

qwen3-0-6b

Очистить

Протестируйте модель

Выберите её выше и отправьте сообщение — ключ не нужен.

Ваш портал · 4 из 6

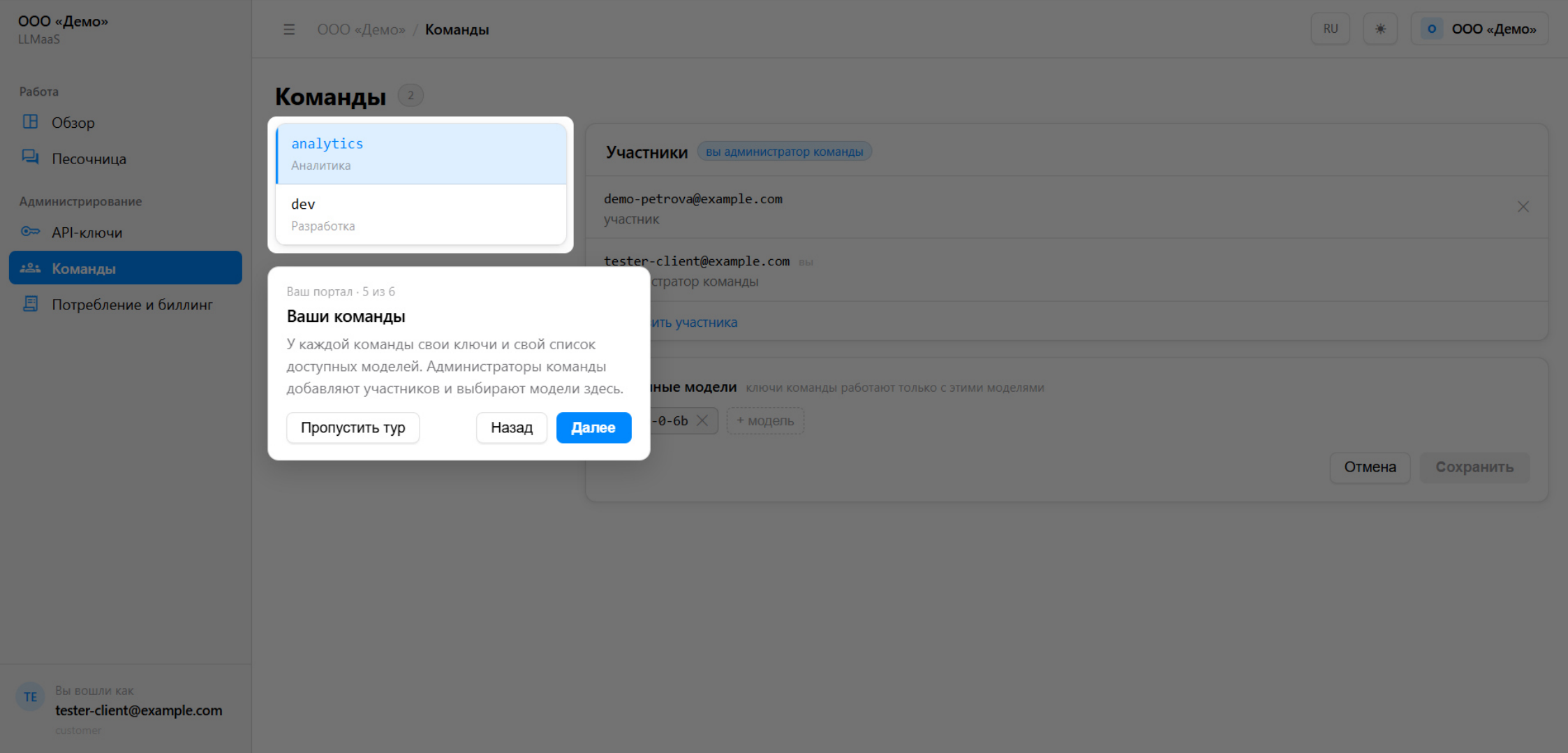
Попробовать модель здесь

Выберите модель и напишите сообщение. Ключ не нужен: этот чат работает прямо из браузера.

Пропустить тур | Назад | Далее

Напишите сообщение... Shift+Enter — перенос строки

Отправить


ООО «Демо»
LLMaaS
Работа
Обзор
Песочница
Администрирование
API-ключи
Команды
Потребление и биллинг
Вы вошли как
tester-client@example.com
customer
ООО «Демо» / Команды
RU
ООО «Демо»
Команды
2
analytics
Аналитика
dev
Разработка
Участники
вы администратор команды
demo-petrova@example.com
участник
tester-client@example.com вы
стратор команды
ить участника
ные модели
ключи команды работают только с этими моделями
-0-6b
+ модель
Отмена
Сохранить
Ваш портал · 5 из 6
Ваши команды
У каждой команды свои ключи и свой список доступных моделей. Администраторы команды добавляют участников и выбирают модели здесь.
Пропустить тур
Назад
Далее

ООО «Демо»
LLMaaS

Работа
- Обзор
- Песочница

Администрирование
- API-ключи
- Команды
- Потребление и биллинг

ООО «Демо» / Обзор

RU

ООО «Демо»

Обзор

Активные ключи
4 из 4 всего

Токены за 30 дней
27 899

Траты за 30 дней
0,139 ₽

Быстрый старт
OpenAI-совместимый эндпоинт

Базовый URL
https://llmaas-litellm.admdv.net
Копировать

Модель
Qwen3-0.6B

curl | python

Скопировать сниппет

```
curl https://llmaas-litellm.admdv.net/v1/chat/completions \
ization: Bearer $LLMAAS_API_KEY" \
t-Type: application/json" \

 "qwen3-0-6b",
s": [{"role": "user", "content": "Hello"}],
: true
```

Перед запуском сниппета экспортируйте ключ в LLMAAS_API_KEY.

Ваш портал · 6 из 6

Ваш аккаунт

Нажмите на своё имя, чтобы пройти тур заново или выйти.

Пропустить тур | Назад | Готово

TE
Вы вошли как
tester-client@example.com
customer