Инструкция SciBox Москва

Инструкция по использованию моделей LLM-сервиса SciBox

В этой инструкции показано, как получить список моделей и выполнить запросы к ним с помощью `curl`. Вместо реального токена используйте переменную окружения или подставляйте свой токен в заголовке `Authorization`.

0. Swagger и базовые URL

- Swagger: по домену `https://llm.t1v.scibox.tech/` или по IP `http://45.145.191.148:4000/`.
- Endpoint списка моделей (пример по IP): `http://45.145.191.148:4000/v1/models`.

Вы можете использовать либо домен без порта, либо IP с портом 4000. В командах ниже оставлен IP с портом, но доменное имя также будет работать без указания порта.

Доступные модели и RPS

- `bge-m3` — 7 RPS. Эмбеддинг-модель для поиска и ранжирования.
- `qwen3-coder-30b-a3b-instruct-fp8` — 2 RPS. Инструкционная кодовая модель.
- `qwen3-32b-awq` — 2 RPS. Универсальная чат-модель.

> Ограничение по RPS распространяется на одну команду (workspace). Если несколько членов команды шлют запросы параллельно, они делят общий лимит, поэтому при нагрузочных сценариях синхронизируйте отправку или ставьте очереди/ретраи.

1. Получение списка моделей

```
curl -H "Authorization: Bearer <YOUR_TOKEN>" \
     https://llm.t1v.scibox.tech/v1/models
```

Пример ответа: