

ОПИСАНИЕ API

ИИ Фабрика API

API сервиса LLMaaS совместим с OpenAI API и доступен по адресу:

```
https://api.aillm.ru
```

Аутентификация

Все запросы требуют API-ключ в заголовке `Authorization:`

```
Authorization: Bearer sk-<ваш-ключ>
```

API-ключи выдаются администратором платформы. Каждый ключ привязан к команде и имеет лимиты по бюджету, RPM и TPM.

Доступные модели

Модель	Контекст	Описание
<code>gpt-oss-120b</code>	128K токенов	RedHatAI/gpt-oss-120b, 4x GPU

Получить список моделей

```
curl https://api.aillm.ru/v1/models \
  -H "Authorization: Bearer sk-<ваш-ключ>"
```

Основные эндпоинты

Chat Completions

```
curl https://api.aillm.ru/v1/chat/completions \
  -H "Authorization: Bearer sk-<ваш-ключ>" \
  -H "Content-Type: application/json" \
  -d '{
    "model": "gpt-oss-120b",
    "messages": [
      {"role": "user", "content": "Привет! Расскажи о себе."}
    ]
  }'
```

Параметры запроса

Параметр	Тип	Описание
<code>model</code>	string	Название модели (обязательный)
<code>messages</code>	array	Массив сообщений (обязательный)
<code>temperature</code>	float	Температура генерации (0.0–2.0, по умолчанию 1.0)
<code>max_tokens</code>	int	Максимальное количество токенов в ответе
<code>top_p</code>	float	Nucleus sampling (0.0–1.0)
<code>stream</code>	bool	Потоковая генерация (по умолчанию false)
<code>stop</code>	string/array	Стоп-последовательности

Пример ответа

```
{
  "id": "chatcmpl-abc123",
  "object": "chat.completion",
```

```

"created": 1710000000,
"model": "gpt-oss-120b",
"choices": [
  {
    "index": 0,
    "message": {
      "role": "assistant",
      "content": "Привет! Я языковая модель..."
    },
    "finish_reason": "stop"
  }
],
"usage": {
  "prompt_tokens": 12,
  "completion_tokens": 25,
  "total_tokens": 37
}
}

```

Streaming

```

curl https://api.aillm.ru/v1/chat/completions \
  -H "Authorization: Bearer sk-<ваш-ключ>" \
  -H "Content-Type: application/json" \
  -d '{
    "model": "gpt-oss-120b",
    "messages": [
      {"role": "user", "content": "Напиши короткое стихотворение"}
    ],
    "stream": true
  }'

```

Ответ приходит в формате Server-Sent Events (SSE):

```

data: {"id":"chatcmpl-abc123","choices":[{"delta":{"content":"Привет"},"inde

data: {"id":"chatcmpl-abc123","choices":[{"delta":{"content":" мир"},"index"

data: [DONE]

```



Использование с SDK

Python (OpenAI SDK)

```
from openai import OpenAI

client = OpenAI(
    base_url="https://api.aillm.ru/v1",
    api_key="sk-<ваш-ключ>",
)

response = client.chat.completions.create(
    model="gpt-oss-120b",
    messages=[
        {"role": "user", "content": "Привет!"}
    ],
)

print(response.choices[0].message.content)
```

Python (streaming)

```
stream = client.chat.completions.create(
    model="gpt-oss-120b",
    messages=[
        {"role": "user", "content": "Расскажи историю"}
    ],
    stream=True,
)

for chunk in stream:
    if chunk.choices[0].delta.content:
        print(chunk.choices[0].delta.content, end="")
```

JavaScript/TypeScript

```
import OpenAI from "openai";

const client = new OpenAI({
```

```
    baseUrl: "https://api.aillm.ru/v1",
    apiKey: "sk-<ваш-ключ>",
  });

const response = await client.chat.completions.create({
  model: "gpt-oss-120b",
  messages: [
    { role: "user", content: "Привет!" },
  ],
});

console.log(response.choices[0].message.content);
```

Проверка состояния

Health check

```
curl https://api.aillm.ru/health/liveliness
# "I'm alive!"
```

Коды ошибок

Код	Описание
401	Невалидный или отсутствующий API-ключ
429	Превышен лимит запросов или бюджета
500	Внутренняя ошибка сервера
503	Модель недоступна (загружается или не развёрнута)