راهنمای اتصال به مدل‌ها از طریق API و استفاده از خدمت LLM Serving

بهمن ماه ۱۴۰۴

تاریخ تهیه: ۱۴۰۴/۰۸/۲۵

تاریخ به‌روزرسانی: ۱۴۰۴/۱۱/۰۵

فهرست مطالب

۱ مقدمه

این راهنما توضیح می‌دهد چگونه می‌توانید با استفاده از API به مدل‌های زبانی (LLM) متصل شوید، درخواست بفرستید و پاسخ دریافت کنید.

۱-۱ لیست مدل‌هایی که فعلاً در سکو موجود هستن (به‌روزشده در تاریخ ۱۴۰۴/۱۱/۰۵)

تاریخ به‌روزرسانی: ۱۴۰۴/۱۱/۰۵		
نوع مدل	نام مدل	اندازه تقریبی پارامتر
LLM	Qwen3-VL-8B-Instruct	8B
	Qwen3-VL-235B-Instruct	235B
	DeepSeek-V3.1	685B
	Qwen3-32B	32B
	gpt-oss-120b	120B
	Qwen2.5-72B	72B
	Qwen3-32B-128K	32B
	Llama4-Scout-17B-16E	17B
	gemma-3-27b-it	27B
	Dorna2-Llama3.1-8B-Instruct	8B
	Qwen3-235B-A22B-Instruct-2507	235B
	VulnLLM-R-7B	7B
SLM (Embedding)	Persian_Sentence_Embedding_v3	N/A
	Jina -embeddings-v3	N/A
	jina-embeddings-v4	N/A
	intfloat-multilingual-e5-large	560M
	Qwen3-Embedding-8B	8B
	baai-bge-m3	N/A
SLM (Reranker)	jina-reranker-v2-base-multilingual	150~200M
	baai-bge-reranker-v2-m3	200~300M
Vision (OCR)	DeepSeek-OCR	3B
	rednote-hilab-dots.ocr	N/A
Audio (ASR)	whisper-large-v3	1.55B
	whisper-large-v3-fa	1.55B

۲ پیش‌نیازها

۱. توکن دسترسی (API Key)

- مثال: sk-xxxxxxxxxxxxxxxxxxxx (توکن شخصی خودتان که برای شما ساخته شده است باید جایگزین xها شود)

۲. ابزار ارسال درخواست HTTP

- Postman
- curl
- Python یا JavaScript

۳. آدرس API

https://alphapi.aip.sharif.ir

۲-۱ واسط API برای گرفتن لیست مدل‌های در دسترس

۱. ارسال درخواست از طریق Postman

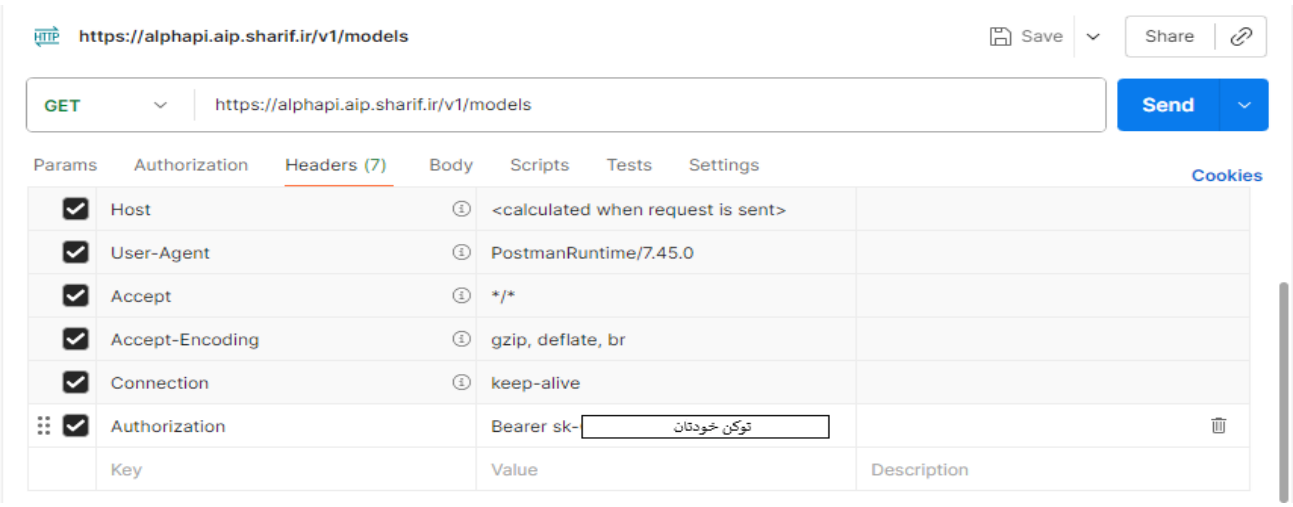
در Postman برای مطمئن شدن از نام دقیق مدل‌ها باید یک ریکوئست GET بسازید.

- روی New کلیک کنید و یک HTTP Request جدید بسازید.
- متد (Method) را روی GET بگذارید.
- آدرس (URL) را مطابق زیر بنویسید:

- https://alphapi.aip.sharif.ir/v1/models

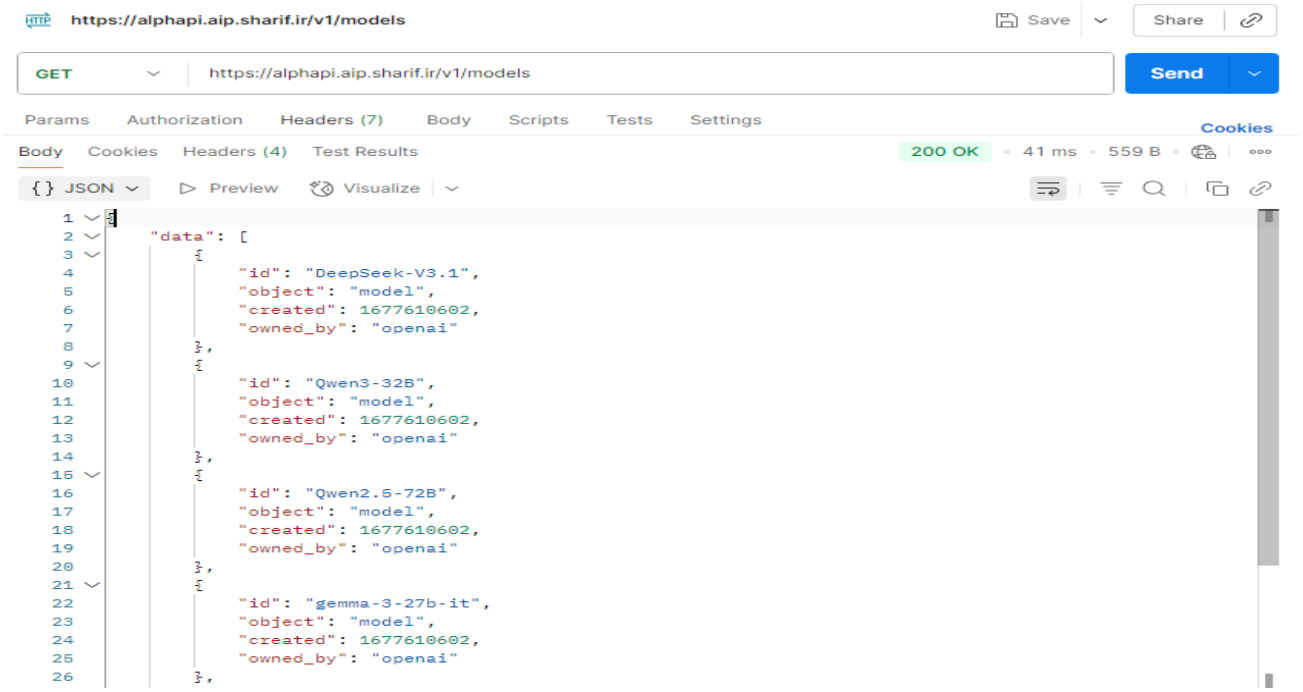
- به تب Headers بروید و توکن خودتان را با عنوان Authorization مشابه تصویر ۱ وارد کنید:
- Authorization → Bearer sk- توکن خودتان

- نیازی به Body نیست.
- روی Send بزنید.



تصویر ۱ تنظیمات احراز هویت در تب Header

خروجی درخواست GET بالا مشابه تصویر ۲ خواهد بود. ID مدل‌هایی که توکن مربوطه به آن‌ها دسترسی دارد قابل رویت است. هر کدام از idها نام دقیق مدلی است که باید در درخواست‌های بعدی استفاده کنید.



تصویر ۲ خروجی درخواست GET برای دیدن نام و مشخصات مدل‌های موجود

۲. ارسال درخواست از طریق Curl

دستور زیر را می‌توان برای دریافت خروجی ارسال کرد:

```
curl.exe -s -X GET "https://alphapi.aip.sharif.ir/v1/models" ^
  -H "Authorization: Bearer sk-توکن-شما" |
  ConvertFrom-Json | ConvertTo-Json -Depth 10
```

۳. ارسال درخواست از طریق Python

```
import requests
URL = "https://alphapi.aip.sharif.ir/v1/models"
TOKEN = "sk"توکن-شما-
headers} =
"  Authorization": f"Bearer {TOKEN},"
"  Accept": "application/json"
{
resp = requests.get(URL, headers=headers, timeout=30)
resp.raise_for_status()

data = resp.json()

# چاپ فقط نام مدل‌ها
for m in data.get("data", []):
    print(m.get("id"))
```

۲-۲ واسط API برای ارسال یک درخواست و دریافت پاسخ

چهار روش برای اتصال به مدل‌ها و دریافت پاسخ در این مرحله نیز وجود دارد که به ترتیب توضیح داده می‌شود.

۱. ارسال درخواست از طریق Postman

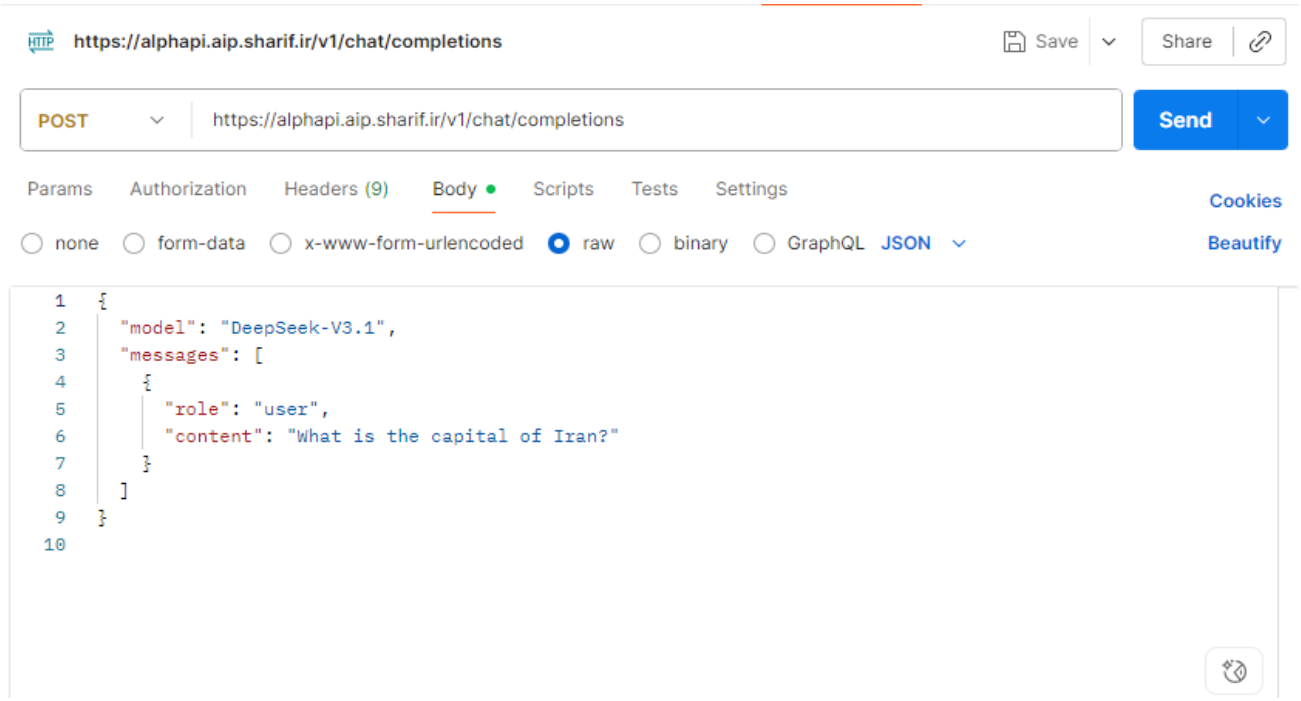
- یک Request جدید بسازید.
- متد URL روی POST باشد و آدرس زیر وارد شود:

https://alphapi.aip.sharif.ir/v1/chat/completions

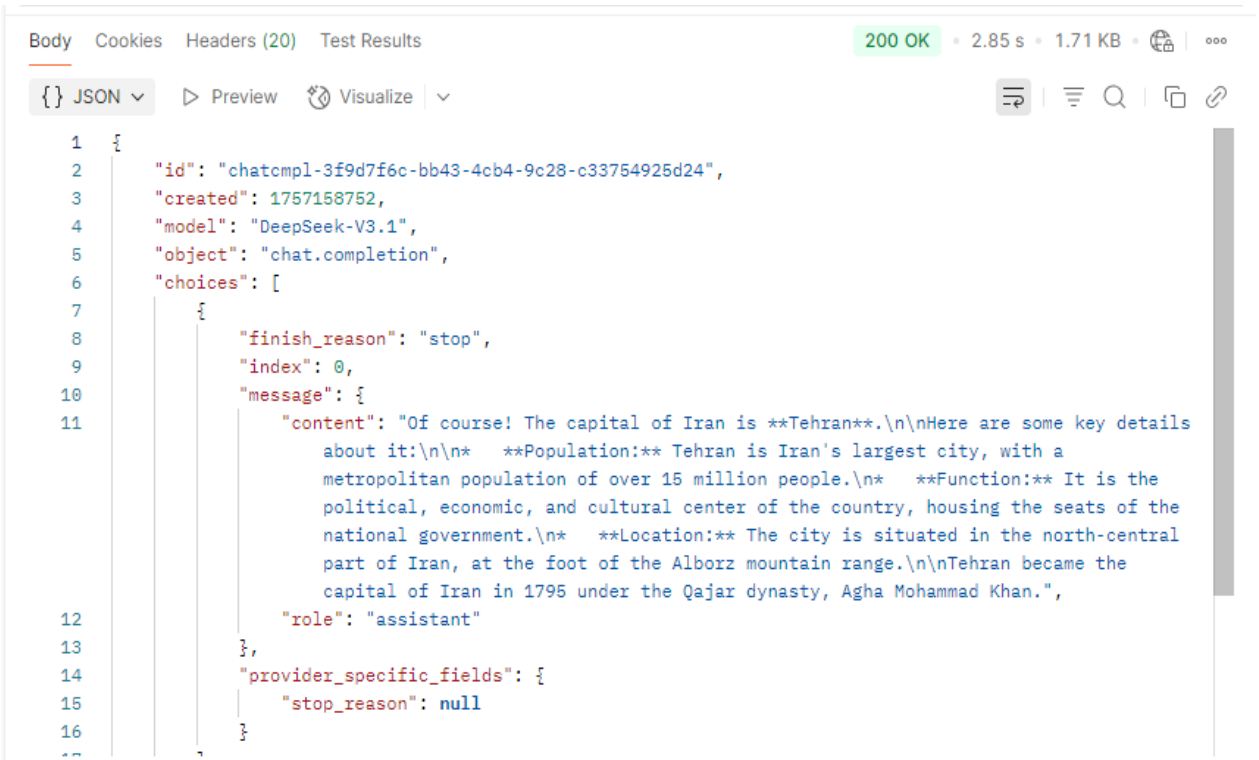
- به تب Headers بروید و دو مورد زیر را اضافه کنید:

Content-Type: application/json
Authorization: Bearer sk- توکن خودتان

- به تب Body بروید، حالت را روی raw و فرمت را روی JSON بگذارید.
- محتوای Body را به صورت امتحانی مشابه تصویر زیر پر کنید.
- روی Send بزنید و پاسخ مدل را در پایین صفحه و در عکس زیر ببینید.



تصویر ۳ نمونه ارسال یک سوال یا Prompt به یکی از مدل‌های سکو



تصویر ۴ خروجی دریافتی از مدل بنا بر سوال ارسالی

۲. **ارسال درخواست از طریق Curl**

```
curl.exe -X POST "https://alphapi.aip.sharif.ir/v1/chat/completions" ^
  -H "Content-Type: application/json" ^
  -H "Authorization: Bearer sk-توکن-شما" ^
  -d "{\"model\":\"DeepSeek-V3.1\",\"messages\":[{\"role\":\"user\",\"content\":\"What is the
capital of Iran?\"}]}"
```

۳. **ارسال درخواست از طریق کد Python**

```
import requests

url = "https://alphapi.aip.sharif.ir/v1/chat/completions"
headers = {
    "Content-Type": "application/json",
    "Authorization": "Bearer sk-توکن-شما"
}
data = {
    "model": "DeepSeek-V3.1",
    "messages": [
        {"role": "user", "content": "What is the capital of Iran?"}
    ]
}
response = requests.post(url, headers=headers, json=data)
print(response.json())
```

۴. **ارسال درخواست از طریق JavaScript (Node.js)**

```
import fetch from "node-fetch";
const url = "https://alphapi.aip.sharif.ir/v1/chat/completions";
const response = await fetch(url, {
  method: "POST",
  headers: {
    "Content-Type": "application/json",
    "Authorization": "Bearer sk-توکن_خودت"
  },
  body: JSON.stringify({
    model: "DeepSeek-V3.1",
    messages: [
      { role: "user", content: "What is the capital of Iran?" }
    ]
  })
});
const result = await response.json();
console.log(result);
```

درخواست بالا را می‌توانید با سوالات دیگر و پرسش از مدل‌های دیگر نیز تکرار کنید.

۲-۳ پارامترهای رایج در Chat Completions

در کنار فیلدهای اصلی در Body مثل model و messages، می‌توان از پارامترهای زیر برای کنترل نحوه تولید پاسخ توسط مدل استفاده کرد.

پارامتر	نوع داده	مقدار پیش‌فرض	توضیح
model	string	—	نام دقیق مدلی که از /v1/models گرفتیم مثل DeepSeek-V3.1
messages	array	—	لیست پیام‌ها در قالب نقش (role: user, assistant, system) و متن (content)
temperature	float (۰ تا ۲)	1.0	کنترل خلاقیت مدل؛ مقادیر بالاتر = خروجی متنوع‌تر، مقادیر پایین‌تر = خروجی متمرکزتر
top_p	float (۰ تا ۱)	1.0	برای محدود کردن temperature ؛ جایگزین روش nucleus sampling تنوع خروجی
max_tokens	integer	بستگی به مدل دارد	حداکثر تعداد توکن خروجی، اگر خیلی کم باشد جواب ناقص می‌آید
n	integer	1	تعداد پاسخ‌هایی که مدل باید همزمان تولید کند
stream	boolean	false	اگر true باشد خروجی به صورت استریم (قطعه‌قطعه) برمی‌گردد
presence_penalty	float (۰ تا ۲)	0	افزایش احتمال تولید موضوعات جدید (جلوگیری از تکرار)
frequency_penalty	float (۰ تا ۲)	0	کاهش احتمال تکرار همان کلمات/عبارات
stop	string یا array	—	تعریف یک یا چند توکن/رشته که با رسیدن به آن تولید متن متوقف شود
user	string	—	شناسه کاربر (برای لاگ یا مدیریت استفاده)، اختیاری است

۲-۴ تست مدل‌های Embedding

۱. ارسال درخواست از طریق Postman

برای دریافت بردار جمله (Sentence Embedding)، مراحل زیر را انجام دهید:

- یک Request جدید در Postman ایجاد کنید.
- متد را روی POST قرار دهید و آدرس زیر را وارد نمایید:

https://alphapi.aip.sharif.ir/v1/embeddings

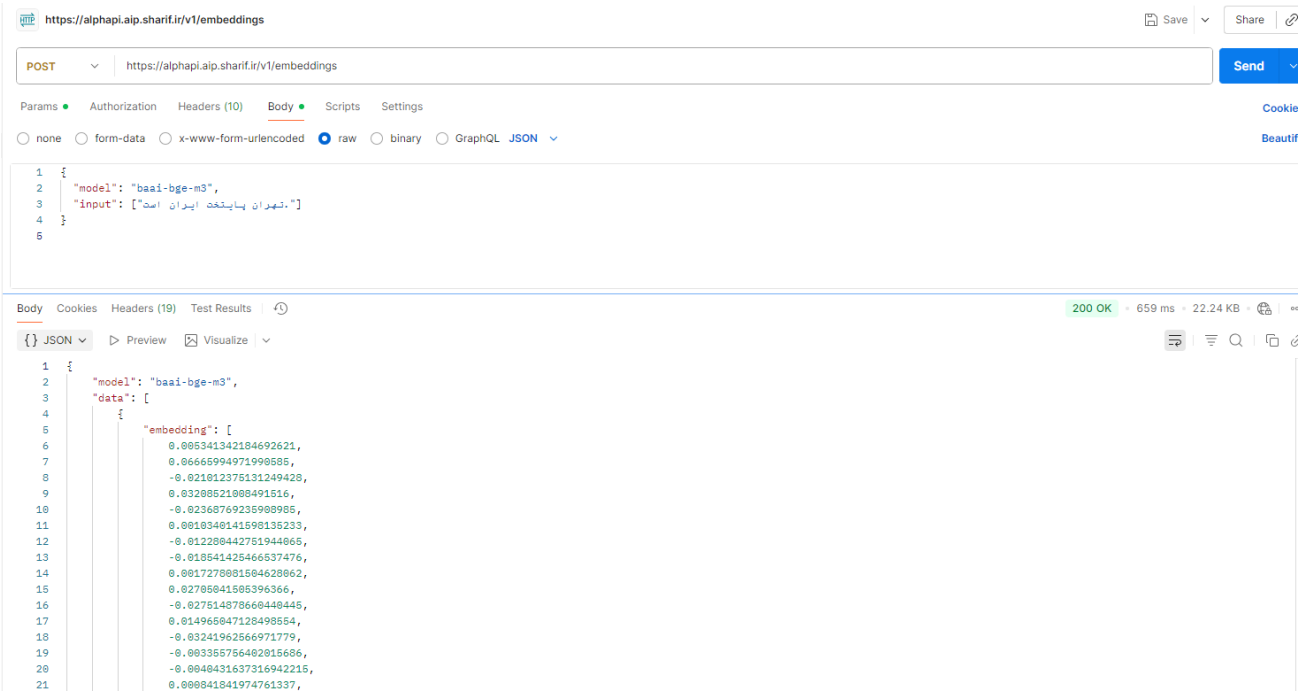
- به تب Headers بروید و دو مورد زیر را اضافه کنید:

Content-Type: application/json
Authorization: Bearer sk- توکن خودتان

- به تب Body بروید، حالت را روی raw و فرمت را روی JSON بگذارید.
- محتوای Body را به صورت امتحانی مشابه زیر پر کنید.

```
{
  "model": "baai-bge-m3",
  "input": ["تهران پایتخت ایران است."]
}
```

خروجی دریافتی مشابه تصویر ۵ خواهد بود.



تصویر ۵ خروجی دریافتی از یک مدل Embedding برای تبدیل جمله به بردار

۲. ارسال درخواست از طریق Curl

```
curl.exe -s -X POST "https://alphapi.aip.sharif.ir/v1/embeddings" ^
  -H "Content-Type: application/json" ^
  -H "Authorization: Bearer sk-توکن-شما" ^
  --data-raw "{\"model\":\"baai-bge-m3\",\"input\":[\"تهران پایتخت ایران است.\"]}"
```

۳. ارسال درخواست از طریق کد Python

```
import requests
url = "https://alphapi.aip.sharif.ir/v1/embeddings"
headers = {
  "Content-Type": "application/json",
  "Authorization": "Bearer sk-توکن-شما"
}
data = {
  "model": "baai-bge-m3",
  "input": ["تهران پایتخت ایران است."]
}
print(requests.post(url, headers=headers, json=data).json())
```

۲-۵ تست مدل‌های Reranker

۱. ارسال درخواست از طریق Postman

برای رتبه‌بندی (Rerank) یک فهرست سند بر اساس یک پرسش، مراحل زیر را انجام دهید:

- یک Request جدید در Postman ایجاد کنید.
- متد را روی POST قرار دهید و آدرس زیر را وارد نمایید:

https://alphapi.aip.sharif.ir/v1/rerank

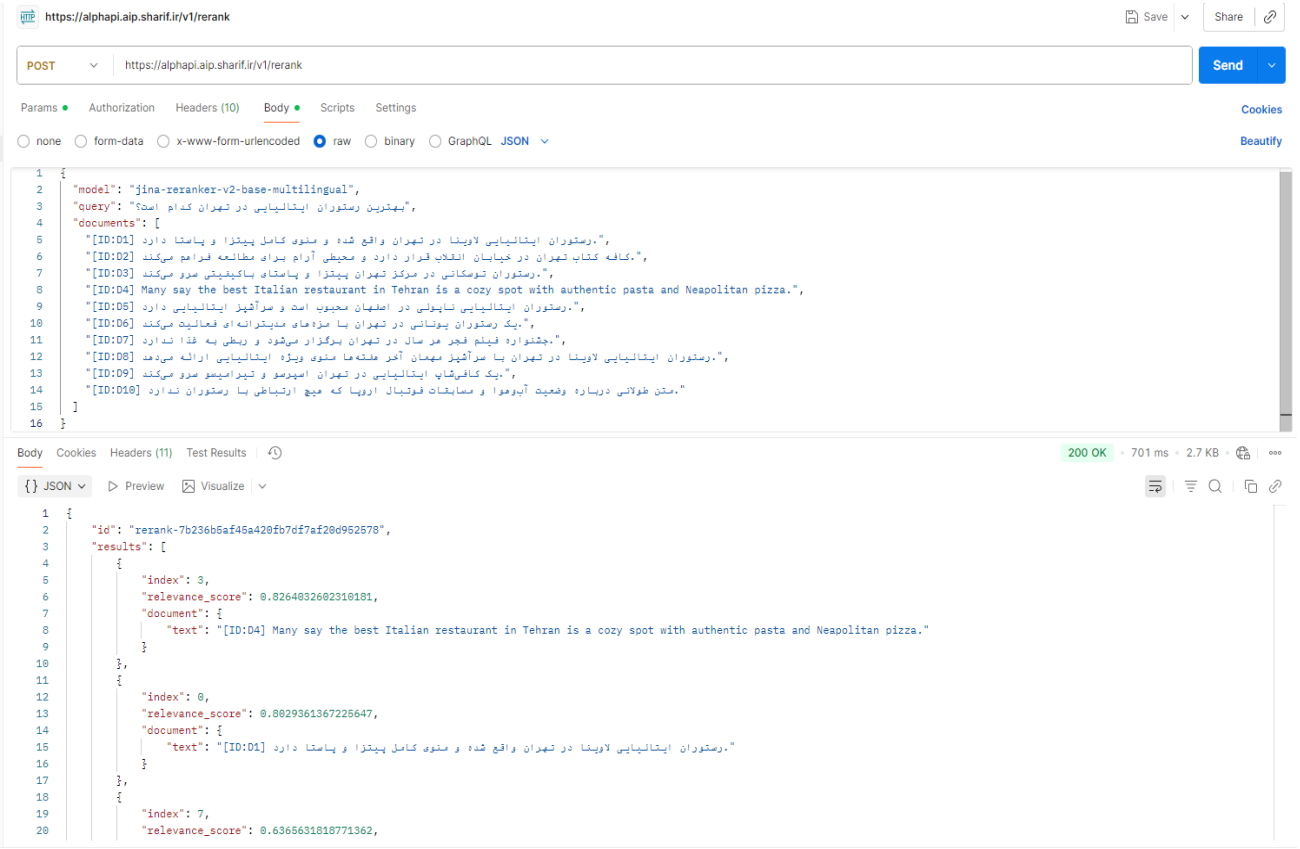
- به تب Headers بروید و دو مورد زیر را اضافه کنید:

Content-Type: application/json
Authorization: Bearer sk- توکن خودتان

- به تب Body بروید، حالت را روی raw و فرمت را روی JSON بگذارید.
- محتوای Body را به صورت امتحانی مشابه زیر پر کنید.

```
{
  "model": "baai-bge-reranker-v2-m3",
  "query": "بهترین رستوران ایتالیایی در تهران کدام است؟",
  "documents": [
    "[ID:D1] رستوران ایتالیایی لاوینا در تهران واقع شده و منوی کامل پیتزا و پاستا دارد.",
    "[ID:D2] کافه کتاب تهران در خیابان انقلاب قرار دارد و محیطی آرام برای مطالعه فراهم می‌کند.",
    "[ID:D3] رستوران توسکانی در مرکز تهران پیتزا و پاستای باکیفیتی سرو می‌کند.",
    "[ID:D4] Many say the best Italian restaurant in Tehran is a cozy spot with authentic pasta and Neapolitan pizza.",
    "[ID:D5] رستوران ایتالیایی ناپولی در اصفهان محبوب است و سرآشپز ایتالیایی دارد.",
    "[ID:D6] یک رستوران یونانی در تهران با مزه‌های مدیترانه‌ای فعالیت می‌کند.",
    "[ID:D7] جشنواره فیلم فجر هر سال در تهران برگزار می‌شود و ربطی به غذا ندارد.",
    "[ID:D8] رستوران ایتالیایی لاوینا در تهران با سرآشپز مهمان آخر هفته‌ها منوی ویژه ایتالیایی ارائه می‌دهد.",
    "[ID:D9] یک کافی‌شاپ ایتالیایی در تهران اسپرسو و تیرامیسو سرو می‌کند.",
    "[ID:D10] متن طولانی درباره وضعیت آب‌وهوا و مسابقات فوتبال اروپا که هیچ ارتباطی با رستوران ندارد."
  ],
  "top_n": 3,
  "return_documents": true
}
```

خروجی دریافتی مشابه تصویر ۶ خواهد بود.



تصویر ۶ خروجی یک مدل Reranker

۲. ارسال درخواست از طریق Curl

```
curl.exe -X POST "https://alphapi.aip.sharif.ir/v1/rerank" ^
  -H "Content-Type: application/json" ^
  -H "Authorization: Bearer sk-توکن-خودتان" ^
  --data-raw "{
      \"model\": \"jina-reranker-v2-base-multilingual\",
      \"query\": \"پایتخت ایران کجاست؟\",
      \"documents\": [
          \"تهران بزرگ‌ترین شهر ایران است و مرکز سیاسی کشور.\",
          \"اصفهان یکی از شهرهای تاریخی ایران است.\",
          \"تبریز مرکز استان آذربایجان شرقی است.\"
      ],
      \"top_n\": 2,
      \"return_documents\": true
  }"
```

۳. ارسال درخواست از طریق کد Python

```
import requests

url = "https://alphapi.aip.sharif.ir/v1/rerank"
headers = {
    "Content-Type": "application/json",
    "Authorization": "Bearer sk-توکن-خودتان"
}
```

```
data = {
    "model": "jina-reranker-v2-base-multilingual",
    "query": "پایتخت ایران کجاست؟",
    "documents": [
        "تهران بزرگ‌ترین شهر ایران است و مرکز سیاسی کشور.",
        "اصفهان یکی از شهرهای تاریخی ایران است.",
        "تبریز مرکز استان آذربایجان شرقی است."
    ],
    "top_n": 2,
    "return_documents": True
}

resp = requests.post(url, headers=headers, json=data)
print(resp.json())
```

۶-۲ تست مدل‌های Audio (ASR)

۱. ارسال درخواست از طریق Postman

برای استفاده از مدل‌های تبدیل صوت به متن همچون Whisper باید مطابق روش زیر عمل کرد. توجه شود که برای استفاده از این دو مدل (Whisper اصلی و Whisper فارسی) باید از آدرس API دیگری استفاده کرد.

- یک Request جدید در Postman ایجاد کنید.
- متد را روی POST قرار دهید و آدرس زیر را وارد نمایید:

https://alphapi.aip.sharif.ir/new/llm/audio/transcriptions

- به تب Headers بروید و سه مورد زیر را اضافه کنید:

Authorization: Bearer sk - توکن خودتان

Accept: application/json

Content-Type: multipart/form-data

- به تب Body بروید و گزینه‌ی from-data را انتخاب کنید. فیلدها را نیز مطابق جدول زیر پر نمایید.

Value	Type	Key
فایل صوتی خودت (audio.mp3)	File	file
whisper-large-v3-fa	Text	model

- فایل موردنظر با یکی از فرمت‌های mp3 / wav / m4a انتخاب و بارگذاری شود.
- گزینه‌ی send زده شود. پاسخ که متن آن محتوای صوتی بارگذاری شده است، در این مرحله قابل رویت است.

۲. ارسال درخواست از طریق Curl

```
curl -X POST "https://alphapi.aip.sharif.ir/new/llm/audio/transcriptions" ^ -H "accept:
application/json" ^ -H "Authorization: Bearer YOUR_KEY" ^ -F
"file=@C:\Users\Test\Downloads\audio.mp3;type=audio/mpeg" ^ -F "model=whisper-large-v3"
```

۳. ارسال درخواست از طریق Python

```
import requests
url = "https://alphapi.aip.sharif.ir/new/llm/audio/transcriptions"
```

```
headers = {
    "Authorization": "Bearer YOUR_KEY",
    "accept": "application/json",
}

files = {
    "file": ("audio.mp3", open(r"C:\Users\Test\Downloads\audio.mp3", "rb"), "audio/mpeg")
}

data = {
    "model": "whisper-large-v3"
}

response = requests.post(
    url,
    headers=headers,
    files=files,
    data=data
)

print("Status code:", response.status_code)
print("Response:", response.text)
```

۲-۷ سایر دستورات API موجود

Feature	Method	URL	Headers	Body (نمونه‌ی حداقلی)	خطاهای رایج / تفسیر
Chat Completion (چت عادی) تولید پاسخ مکالمه‌ای	POST	https://alphapi.aip.sharif.ir/v1/chat/completions	Authorization, Content-Type	{ "model": "DeepSeek-V3.1", "messages": [{ "role": "user", "content": "سلام! حالت چطوره؟" }] }	۴۰۱ : نامعتبر. ۴۲۲/۴۰۰: ساختار Body/ نام مدل اشتباه.
Reasoning پاسخ تحلیلی/مرحله‌ای	POST	https://alphapi.aip.sharif.ir/v1/chat/completions	Authorization, Content-Type	{ "model": "deepseek-r1", "messages": [{ "role": "user", "content": "چرا آسمان آبی است؟" }] }	۴۰۴/۴۰۰: مدل روی سرور وجود ندارد.
Multi-round Conversation حفظ سابقه گفتگو	POST	https://alphapi.aip.sharif.ir/v1/chat/completions	Authorization, Content-Type	{ "model": "DeepSeek-V3.1", "messages": [{ "role": "system", "content": " تو یک دستیار مفید هستی." }, { "role": "user",	۴۰۰: فرمت messages اشتباه است

Feature	Method	URL	Headers	Body (نمونه‌ی حداقلی)	خطاهای رایج / تفسیر
				"content": " من می‌خوام به تهران سفر کنم. چه جاهایی رو پیشنهاد می‌دی؟" }, { "role": "assistant", "content": " میتونی از بام‌لند، دربند و کاخ گلستان دیدن کنی." }, { "role": "user", "content": " برای غذا چی پیشنهاد می‌کنی؟" }] }	
Chat Prefix Completion (Beta) وادار کردن مدل به شروع/سبک خاص	POST	https://alphapi.aip.sharif.ir/v1/chat/completions	Authorization, Content-Type	{ "model": "DeepSeek-V3.1", "messages": [{ "role": "system", "content": " فقط با یک جمله کوتاه جواب بده." }, { "role": "user", "content": " در مورد فواید خواب کافی بگو." }] }	—
JSON Output خروجی JSON معتبر	POST	https://alphapi.aip.sharif.ir/v1/chat/completions	Authorization, Content-Type	{ "model": "DeepSeek-V3.1", "response_format": { "type": "json_object" }, "messages": [{ "role": "user", "content": " سه میوه و قیمت تقریبی آن‌ها را بده." }] }	۴۰۰: اگر مدل/سرور response_format را پشتیبانی نکند.
Function Calling انتخاب/فراخوانی تابع توسط مدل	POST	https://alphapi.aip.sharif.ir/v1/chat/completions	Authorization, Content-Type	{ "model": "DeepSeek-V3.1", "messages": [{ "role": "user", "content": "هوا در تهران چطوره؟" }], "tools": [{ "type": "function", "function": { "name": "get_weather",	۴۰۰: ساختار / اسامی tools اشتباه است.

Feature	Method	URL	Headers	Body (نمونه‌ی حداقلی)	خطاهای رایج / تفسیر

				"description": " دریافت وضعیت هوا برحسب شهر",
				"parameters": {
				"type": "object",
				"properties": {
				"city": {
				"type": "string"
				}
				},
				"required": ["city"]
				}
				}
				}
				],
				"tool_choice": "auto"
				}	

Completions (متن ساده) تکمیل متن غیرچتی	POST	https://alphapi.aip.sharif.ir/v1/completions	Authorization, Content-Type	{
				"model": "DeepSeek-V3.1",
				"prompt": "Translate to French: Hello world",
				"max_tokens": 64
				}
					۴۰۵/۴۰۴: اگر این مسیر غیرفعال باشد.

FIM Completion (Fill-in-the-Middle) پرکردن وسط متن/کد	POST	https://alphapi.aip.sharif.ir/v1/completions	Authorization, Content-Type	{
				"model": "DeepSeek-V3.1",
				"prompt": "function add(a, b) {",
				"suffix": "}",
				"max_tokens": 64
				}
					۴۰۰: اگر prompt/suffix ناقص باشد.

Context Caching کش سمت‌سرور برای ورودی‌های تکراری	POST (دو بار پیاپی)	https://alphapi.aip.sharif.ir/v1/chat/completions	Authorization, Content-Type	{
				"model": "DeepSeek-V3.1",
				"messages": [
				{
				"role": "user",
				"content": "این متن نسبتاً بزرگ است. لطفاً به چند bullet خلاصه کن."
				}
				],
				"temperature": 0.2,
				"context_caching": true
				}
					—

List Models دیدن مدل‌های دردسترس	GET	https://alphapi.aip.sharif.ir/v1/models	Authorization (اختیاری: +Accept: application/json)	بدون Body	۴۰۱: توکن نامعتبر.
Get User Balance دیدن موجودی	GET	https://alphapi.aip.sharif.ir/v1/user/balance	Authorization	بدون Body	۴۰۴/۴۰۱: مسیر غیرفعال/مجوز ندارید.

Embeddings بردارسازی متن برای مدل‌هایی که Embedding را ساپورت می‌کنند.	POST	https://alphapi.aip.sharif.ir/v1/embeddings	Authorization, Content-Type	{
				"model": "baai-bge-m3",
				"input": ["This is a test"]
				}
					400 با پیام‌هایی مثل *provider/model not provided*: یعنی مدل یعنی embedding روی

Feature	Method	URL	Headers	Body (نمونه‌ی حداقلی)	خطاهای رایج / تفسیر
					سرور شما رجیستر نشده.

۸-۲ پیاده‌سازی RAG و استفاده از طریق API

۴. RAG نیمه هوشمند بدون Embedding از طریق Postman

۱. یک Environment جدید ساخته شود.

۲. این متغیرها با مقدارهای مربوطه اضافه شود:

5. base_url = https://alphapi.aip.sharif.ir
6. api_key = sk-YOUR KEY
7. model_id = DeepSeek-V3.1
8. QUESTION= مثلا چطور مرخصی بگیرم؟
9. TOP_K = 3
10. DOCS = آرایه‌ی Json اسناد

۳. فعال کردن Environment ساخته‌شده.

۴. ساخت یک ریکوئست با مشخصات زیر:

- Method: Post
- URL: https://alphapi.aip.sharif.ir/v1/chat/completions
- Headers: Authorization: Bearer {{api_key}}
- Content-Type: application/json
- Body: Json, Raw

۵. اسکریپتی مشابه اسکریپت زیر را در تب Pre-request Script وارد می‌کنیم.

```
const TOP_K = parseInt(pm.environment.get("TOP_K") || "3", 10);
const QUESTION = (pm.environment.get("QUESTION") || "").trim();
const DOCS_RAW = pm.environment.get("DOCS") || "[]";
const MODEL_ID = (pm.environment.get("model_id") || "DeepSeek-V3.1").trim();

function normalizeFa(s=""){return s.replace(/\u200c/g,"
").replace(/[أإآ]/g,"ا").replace(/ی/g,"ی").replace(/ک/g,"ک").replace(/[ـًٌٍَُِّْ]/g,"");}
function tokenize(s=""){return normalizeFa(s).toLowerCase().replace(/[^\w\u0600-\u06FF\s]/g,"
").split(/\s+/).filter(Boolean);}

let docs=[];
try{docs=JSON.parse(DOCS_RAW);}catch(e){docs=[];}

const qTokens=new Set(tokenize(QUESTION));
const scored=docs.map(d=>{
  const text=String(d.text||"");
  const tokens=tokenize(text);
  let score=0;
  for(const t of tokens) if(qTokens.has(t)) score++;
  return {...d,score,snippet:text.slice(0,200).replace(/\s+/g," ").trim()};
}).sort((a,b)=>b.score-a.score).slice(0,TOP_K);
```

```
const CTX=scored.map(d=>`- [${d.id||"doc"} | ${d.source||"src"}] ${d.snippet}`).join("\n");
pm.environment.set("CTX",CTX);

const body={
  model:MODEL_ID,
  temperature:0.2,
  messages:[
    {role:"system",content:"فقط بر اساس متن‌های زیر پاسخ بده. اگر کافی نبود بگو اطلاعات کافی نیست."},
    {role:"user",content:`سؤال: ${QUESTION}\n\nزمینه:\n${CTX}\n\nپاسخ:`}
  ]
};
pm.request.headers.upsert({key:"Content-Type",value:"application/json"});
pm.request.body.update(JSON.stringify(body,null,2));
```

۶. محتوای Body را مشابه زیر پر نمایید:

```
  {
    "id": "HR-Leave",
    "source": "hr_wiki.md",
    "text": "درخواست مرخصی باید در پورتال HR توسط مدیر مستقیم تایید شود. ثبت."
  },
  {
    "id": "Procurement",
    "source": "procurement.pdf",
    "text": "خرید تجهیزات فقط از طریق واحد تدارکات و ثبت فرم PR انجام می‌شود."
  },
  {
    "id": "Onboarding",
    "source": "onboarding.docx",
    "text": "فرایند استخدام شامل مصاحبه فنی و منابع انسانی است."
  },
  {
    "id": "Leave-Types",
    "source": "hr_policy.md",
    "text": "مرخصی استعلاجی با ارائه گواهی پزشک تایید می‌شود. مرخصی ساعتی با تایید مدیر ممکن است."
  }
]
```

۱۱. هوشمند با Embedding از طریق Postman

برای استفاده از RAG به صورت کاملاً هوشمند یا افزودن دیتابیس به آن، نیاز است که از مدل‌های Embedding موجود در سکو استفاده شود تا ایندکس‌های برداری ساخته شوند یا دیتابیس برداری شده (Vector DB) به مدل داده شود و سوال هم توسط یک سرویس محلی برداری شود و نتیجه بر اساس آن حاصل شود.