



رئاست جمهوری  
معاونت علمی و فناوری

سکوی ملی متن باز هوش مصنوعی:

راهنمای اتصال به مدل‌ها و دریافت پاسخ از طریق API

پیوست: تدارد

۱۴۰۴/۱۱/۰۵

شماره: ۰۶



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

۱۴۰۴ بهمن ماه

تاریخ تهییه: ۱۴۰۴/۰۸/۲۵

تاریخ بهروزرسانی: ۱۴۰۴/۱۱/۰۵



رئاست جمهوری  
سازمان علمی و فناوری

سکوی ملی متن باز هوش مصنوعی:  
راهنمای اتصال به مدل‌ها و دریافت پاسخ از طریق API

پیوست: تدارد

تاریخ: ۱۴۰۴/۱۱/۰۵

شماره: ۰۶



## فهرست مطالب

۳	۱ مقدمه
۳	۱-۱ لیست مدل‌هایی که فعلًا در سکو موجود هستن (بهروزشده در تاریخ ۱۴۰۴/۱۱/۰۵)
۳	۲ پیش‌نیازها
۴	۲-۱ واسط API برای گرفتن لیست مدل‌های در دسترس
۶	۲-۲ واسط API برای ارسال یک درخواست و دریافت پاسخ
۸	۲-۳ پارامترهای رایج در Chat Completions
۹	۲-۴ تست مدل‌های Embedding
۱۰	۲-۵ Reranker
۱۳	۲-۶ تست مدل‌های (ASR) Audio
۱۴	۲-۷ سایر دستورات API موجود
۱۷	۲-۸ پیاده‌سازی RAG و استفاده از طریق API



جمهوری  
سازمان علمی و فناوری

سکوی ملی متن باز هوش مصنوعی:  
راهنمای اتصال به مدل‌ها و دریافت پاسخ از طریق API

پیوست: ندارد

تاریخ: ۱۴۰۴/۱۱/۰۵

شماره: ۰۶



## ۱ مقدمه

این راهنمای 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 فعلًا در سکو موجود هستند (به روز شده در تاریخ ۱۴۰۴/۱۱/۰۵)

تاریخ به روزرسانی: ۱۴۰۴/۱۱/۰۵		
اندازه تقریبی پارامتر	نام مدل	نوع مدل
8B	Qwen3-VL-8B-Instruct	LLM
235B	Qwen3-VL-235B-Instruct	
685B	DeepSeek-V3.1	
32B	Qwen3-32B	
120B	gpt-oss-120b	
72B	Qwen2.5-72B	
32B	Qwen3-32B-128K	
17B	Llama4-Scout-17B-16E	
27B	gemma-3-27b-it	
8B	Dorna2-Llama3.1-8B-Instruct	
235B	Qwen3-235B-A22B-Instruct-2507	SLM (Embedding)
7B	VulnLLM-R-7B	
N/A	Persian_Sentence_EMBEDDING_v3	
N/A	Jina -embeddings-v3	
N/A	jina-embeddings-v4	
560M	intfloat-multilingual-e5-large	
8B	Qwen3-Embedding-8B	SLM (Reranker)
N/A	baai-bge-m3	
150~200M	jina-reranker-v2-base-multilingual	
200~300M	baai-bge-reranker-v2-m3	Vision (OCR)
3B	DeepSeek-OCR	
N/A	rednote-hilab-dots.ocr	Audio (ASR)
1.55B	whisper-large-v3	
1.55B	whisper-large-v3-fa	

## ۲ پیش‌نیازها

### ۱. توکن دسترسی (API Key)

- مثال: sk-xxxxxxxxxxxxxxxxxxxxxx (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>

 <b>رئاست جمهوری</b> <b>سازمان علمی و فناوری</b>	<b>سکوی ملی متن باز هوش مصنوعی:</b> <b>راهنمای اتصال به مدل‌ها و دریافت پاسخ از طریق API</b>	
<b>پیوست: ندارد</b>	<b>تاریخ: ۱۴۰۴/۱۱/۰۵</b>	<b>شماره: ۰۶</b>

## ۲-۱ API برای گرفتن لیست مدل‌های در دسترس

### ۱. ارسال درخواست از طریق Postman

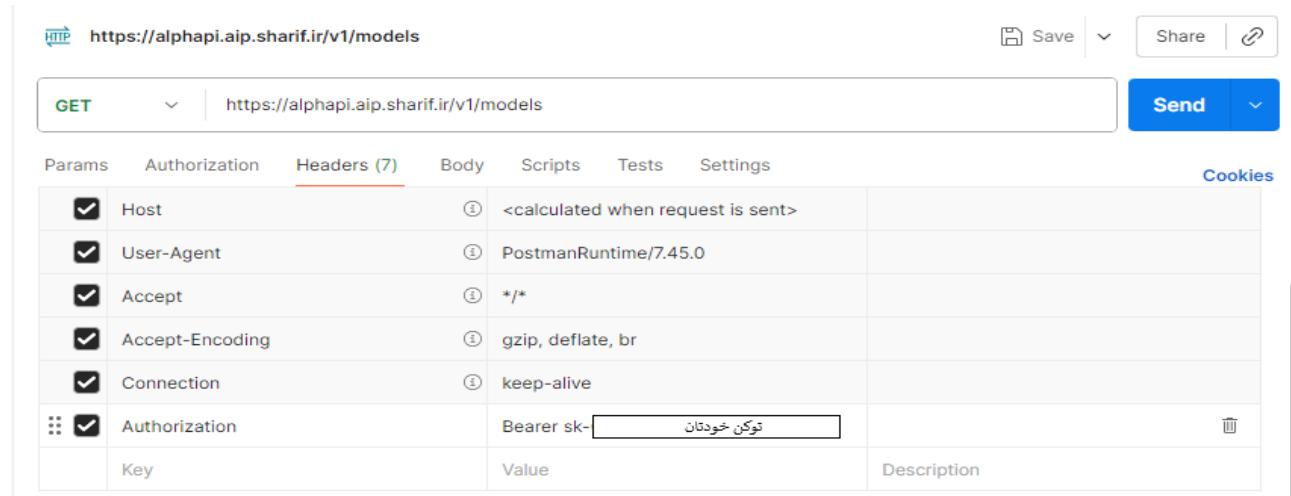
در Postman برای مطمئن شدن از نام دقیق مدل‌ها باید یک ریکوئست GET بسازید.

- روی New کلیک کنید و یک HTTP Request HTTP جدید بسازید.
- متده (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:

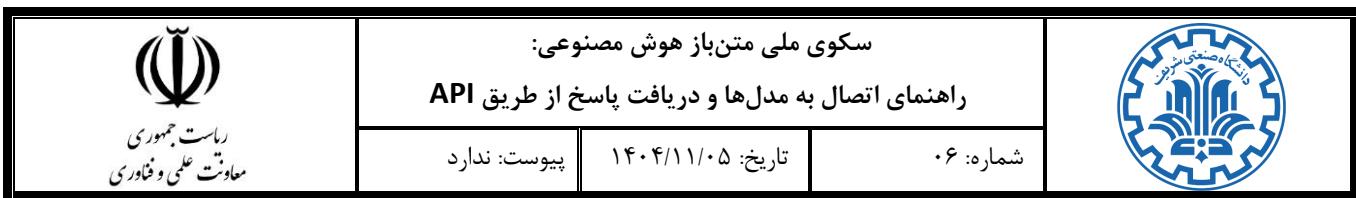
- Authorization → Bearer sk-  
توکن خودتان نیازی به Body نیست.
- روی Send بزنید.



Params		Authorization		Headers (7)		Body	Scripts	Tests	Settings	Cookies
<input checked="" type="checkbox"/>	Host	<input type="checkbox"/>	<calculated when request is sent>							
<input checked="" type="checkbox"/>	User-Agent	<input type="checkbox"/>	PostmanRuntime/7.45.0							
<input checked="" type="checkbox"/>	Accept	<input type="checkbox"/>	/*							
<input checked="" type="checkbox"/>	Accept-Encoding	<input type="checkbox"/>	gzip, deflate, br							
<input checked="" type="checkbox"/>	Connection	<input type="checkbox"/>	keep-alive							
<input checked="" type="checkbox"/>	Authorization	<input type="checkbox"/>	Bearer sk-	токен خودتان						
	Key		Value			Description				

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.



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

Save | Share | [Send](#)

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

Params Authorization Headers (7) Body Scripts Tests Settings Cookies

Body Cookies Headers (4) Test Results

```

1 {
2   "data": [
3     {
4       "id": "DeepSeek-V3.1",
5       "object": "model",
6       "created": 1677610602,
7       "owned_by": "openai"
8     },
9     {
10       "id": "Qwen3-32B",
11       "object": "model",
12       "created": 1677610602,
13       "owned_by": "openai"
14     },
15     {
16       "id": "Qwen2.5-72B",
17       "object": "model",
18       "created": 1677610602,
19       "owned_by": "openai"
20     },
21     {
22       "id": "gemma-3-27b-bit",
23       "object": "model",
24       "created": 1677610602,
25       "owned_by": "openai"
26     }
  ]
}
  
```

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-شما-токن" |
ConvertFrom-Json | ConvertTo-Json -Depth 10
```

## ۳. ارسال درخواست از طریق Python

```
import requests
URL = "https://alphapi.aip.sharif.ir/v1/models"
TOKEN = "sk-شما-токن"
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"))
```

 <b>جمهوری جمهوری جمهوری</b> <b>جمهوری جمهوری جمهوری</b>	<b>سکوی ملی متن باز هوش مصنوعی:</b> <b>راهنمای اتصال به مدل‌ها و دریافت پاسخ از طریق API</b>	
<b>پیوست: ندارد</b>	<b>تاریخ: ۱۴۰۴/۱۱/۰۵</b>	<b>شماره: ۰۶</b>

## 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 تاب Headers بروید و دو مورد زیر را اضافه کنید:

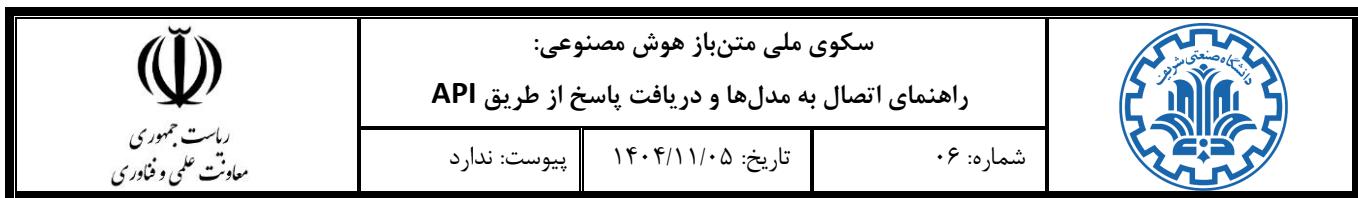
Content-Type: application/json

Authorization: Bearer sk\_توکن خودتان-

- به تاب 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.



HTTP <https://alphapi.aip.sharif.ir/v1/chat/completions> Save Share

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

Params Authorization Headers (9) Body Scripts Tests Settings Cookies

none form-data x-www-form-urlencoded raw binary GraphQL JSON Beautify

```
1 {
2   "model": "DeepSeek-V3.1",
3   "messages": [
4     {
5       "role": "user",
6       "content": "What is the capital of Iran?"
7     }
8   ]
9 }
```

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

Body Cookies Headers (20) Test Results 200 OK 2.85 s 1.71 KB

{ } JSON Preview Visualize

```
1 {
2   "id": "chatmpl-3f9d7f6c-bb43-4cb4-9c28-c33754925d24",
3   "created": 1757158752,
4   "model": "DeepSeek-V3.1",
5   "object": "chat.completion",
6   "choices": [
7     {
8       "finish_reason": "stop",
9       "index": 0,
10      "message": {
11        "content": "Of course! The capital of Iran is **Tehran**.\n\nHere are some key details about it:\n\n* **Population:** Tehran is Iran's largest city, with a metropolitan population of over 15 million people.\n* **Function:** It is the political, economic, and cultural center of the country, housing the seats of the national government.\n* **Location:** The city is situated in the north-central part of Iran, at the foot of the Alborz mountain range.\n\nTehran became the capital of Iran in 1795 under the Qajar dynasty, Agha Mohammad Khan.",
12        "role": "assistant"
13      },
14      "provider_specific_fields": {
15        "stop_reason": null
16      }
17    }
18  ]
19 }
```

تصویر ۴ خروجی دریافتی از مدل بنابر سوال ارسالی



رئاست جمهوری  
سازمان علمی و فناوری

سکوی ملی متن باز هوش مصنوعی:  
راهنمای اتصال به مدل‌ها و دریافت پاسخ از طریق API

پیوست: ندارد

تاریخ: ۱۴۰۴/۱۱/۰۵

شماره: ۰۶



## 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-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-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 (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-شما"
    },
    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.

## ۴-۳ 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.



راست جمهوری  
سازمان اسناد و کتابخانه ملی

سکوی ملی متن باز هوش مصنوعی:  
راهنمای اتصال به مدل‌ها و دریافت پاسخ از طریق API

پیوست: ندارد

تاریخ: ۱۴۰۴/۱۱/۰۵

شماره: ۰۶



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

## ۲-۴ قسّت 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.
- متده را روی POST قرار دهید و آدرس زیر را وارد نمایید:

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

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

Content-Type: application/json

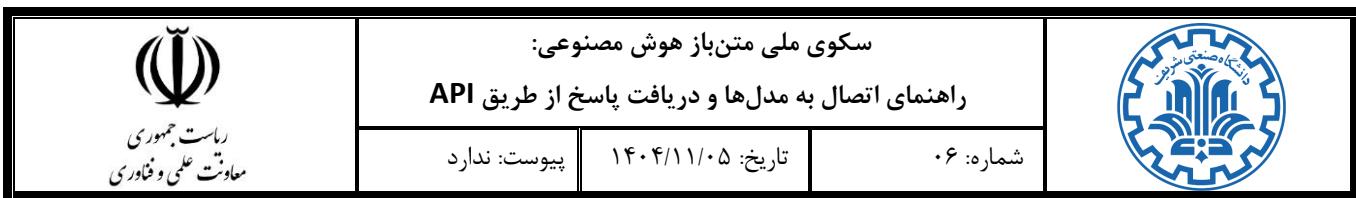
Authorization: Bearer sk\_توکن خودتان-

- به تاب 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.



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

Save | Share | [Copy](#)

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

Send

Params Authorization Headers (10) Body Scripts Settings

none form-data x-www-form-urlencoded raw binary GraphQL JSON [Beautify](#)

```

1 {
2   "model": "baai-bge-m3",
3   "input": ["تهران پایتخت ایران است"]
4 }
5

```

Body Cookies Headers (19) Test Results

{} JSON Preview Visualize

200 OK 659 ms 22.24 KB

Copy | [Raw](#) | [JSON](#) | [Text](#) | [CSV](#)

```

1 {
2   "model": "baai-bge-m3",
3   "data": [
4     {
5       "embedding": [
6         0.005341342184692621,
7         0.06665994971998589,
8         -0.021023757131249428,
9         0.032088521088491516,
10        -0.02368769235909895,
11        0.0010340141598136233,
12        -0.012289442751944665,
13        -0.018541428466537476,
14        0.0017278801584628862,
15        0.02785641565396366,
16        -0.027514878660449445,
17        0.014965847128498554,
18        -0.03241962566971779,
19        -0.003355756482015686,
20        -0.0040431637316942215,
21        0.000841841974761337,

```

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-شما-токن" ^
--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-شما-токن"
}
data = {
    "model": "baai-bge-m3",
    "input": ["تهران پایتخت ایران است"]
}
print(requests.post(url, headers=headers, json=data).json())
```

## ۴-۵ قسم 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.

- متده را روی POST قرار دهید و آدرس زیر را وارد نمایید:

 راست‌جمهوری معاونت علمی و فناوری	<b>سکوی ملی متن باز هوش مصنوعی:</b> <b>راهنمای اتصال به مدل‌ها و دریافت پاسخ از طریق API</b>	
پیوست: تدارد	۱۴۰۴/۱۱/۰۵	شماره: ۰۶

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

خروجی دریافتی مشابه تصویر ۶ خواهد بود.



جمهوری ریاست  
معاونت علمی و فناوری

راهنمای اتصال به مدل‌ها و دریافت پاسخ از طریق API سکوی ملی متن باز هوش مصنوعی:

پیوست: ندارد

تاریخ: ۱۴۰۴/۱۱/۰۵

شماره: ۶



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

Save | Share | 

Send 

**POST** ▾ <https://alphapi.aip.sharif.ir/v1/rerank>

▼

Params • Authorization Headers (10) Body • Scripts Settings

## Cookies

none  form-data  x-www-form-urlencoded  raw  binary  GraphQL  JSON

## Beautify

```
1  {
2     "model": "jina-reranker-v2-base-multilingual",
3     "query": "پیشترین رستوران ایتالیایی در تهران کدام است؟",
4     "documents": [
5         {"ID":01, "text": "رستوران ایتالیایی لاوینا در تهران واقع شده و منو کامل پیشوا و پاستا دارد."},  
6         {"ID":02, "text": "کافه کتاب تهران در خیابان اسلام قرار دارد و امکان آنرا برای افراد میگذند"},  
7         {"ID":03, "text": "رستوران توسکاتی در مرکز تهران پیشوا و پاستا ایتالیایی را میگذند"},  
8         {"ID":04, "text": "Many say the best Italian restaurant in Tehran is a cozy spot with authentic pasta and Neapolitan pizza."},  
9         {"ID":05, "text": "رستوران ایتالیایی پاولو در اطهارن جمیع امکاناتی را دارد."},  
10        {"ID":06, "text": "بیک رستوران پیشوا در تهران با غذاهای پیشوازه ای فناوه ایتالیایی میگذند"},  
11        {"ID":07, "text": "جشنواره فیلم پیرامون سراسر در تهران به همه امکانات پیشوازه ای فناوه ایتالیایی میگذند"},  
12        {"ID":08, "text": "رستوران ایتالیایی لاوینا در تهران با سرویس‌های ممتاز و پیشوازه ای ایتالیایی ارائه می‌نماید"},  
13        {"ID":09, "text": "بیک کاسکاتو ایتالیایی در تهران اسروری و دیزاین پیشوازه ای فناوه ایتالیایی را میگذند"},  
14        {"ID":10, "text": "متن فلواتی دریاره و پیشوازه ای ایتالیایی ارائه می‌نماید و میتوان فریتالی اورپا که پیشوازه ایتالیایی را در تهران معرفت کرد."},  
15    ]  
16 }
```

200 OK • 701 ms • 2.7 KB •  | 000

≡ | ⌂ | 🔎 | 🗃 | 🔗

Body Cookies Headers (11) Test Results | 

```
{ } JSON ▾ ▷ Preview ▷ Visualize | ▾  
1 {  
2   "id": "rerank-7b236b5af45a420fb7df7af20d952578",  
3   "results": [  
4     {  
5       "index": 3,  
6       "relevance_score": 0.8264032602310181,  
7       "document": {  
8         "text": "[ID:D4] Many say the best Italian restaurant in Tehran is a cozy spot with authentic pasta and Neapolitan pizza."  
9       }  
10      },  
11      {  
12        "index": 0,  
13        "relevance_score": 0.8029361367225647,  
14        "document": {  
15          "text": "[ID:01] رستوران ایتالیایی لاوینا در تهران واقع شده و منو کامل پیشوا و پاستا دارد."  
16        }  
17      },  
18      {  
19        "index": 7,  
20        "relevance_score": 0.6365631818771362,  
21      }  
22    ]  
23  }  
24 }
```

## 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-خودتان" ^
--data-raw "{  
    \"model\": \"jina-reranker-v2-base-multilingual\",  
    \"query\": \"پاییخت 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. ادسا، درخواست از طریق کد Python

```
import requests

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



رئاست جمهوری  
سازمان علمی و فناوری

سکوی ملی متن باز هوش مصنوعی:  
راهنمای اتصال به مدل‌ها و دریافت پاسخ از طریق API

پیوست: ندارد

تاریخ: ۱۴۰۴/۱۱/۰۵

شماره: ۰۶



```
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())
```

## ۲-۶ قسٽ 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.
- متده را روی POST قرار دهید و آدرس زیر را وارد نمایید:
- به تب Headers بروید و سه مورد زیر را اضافه کنید:

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

Authorization: Bearer sk\_توکن خودتان-

Accept: application/json

Content-Type: multipart/form-data

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

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

- فایل موردنظر با یکی از فرمتهای mp3 / wav / m4a انتخاب و بارگذاری شود.
- گزینه 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

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



رئاست جمهوری  
سازمان علمی و فناوری

سکوی ملی متن باز هوش مصنوعی:  
راهنمای اتصال به مدل‌ها و دریافت پاسخ از طریق API

پیوست: ندارد

تاریخ: ۱۴۰۴/۱۱/۰۵

شماره: ۰۶



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



خطاهای رایج / تفسیر	Body (نمونه‌ی حداقلی)	Headers	URL	Method	Feature
	<p>"description": "دریافت وضعیت هوا در شهر، بر حسب شهر،</p> <pre>{   "parameters": {     "type": "object",     "properties": {       "city": {         "type": "string"       }     },     "required": ["city"]   } }, "tool_choice": "auto" }</pre>				
۴۰۵/۴۰۴: اگر این مسیر غیرفعال باشد.	<pre>{   "model": "DeepSeek-V3.1",   "prompt": "Translate to French: Hello world",   "max_tokens": 64 }</pre>	Authorization , Content-Type	https://alphapi.aip.sharif.ir/v1/completions	POST	<b>Completions</b> (متن ساده) تمکیل متن غیرچشمی
۴۰۰: اگر prompt/suffix ناقص باشد.	<pre>{   "model": "DeepSeek-V3.1",   "prompt": "function add(a, b) {",   "suffix": "}",   "max_tokens": 64 }</pre>	Authorization , Content-Type	https://alphapi.aip.sharif.ir/v1/completions	POST	<b>FIM Completion (Fill-in-the-Middle)</b> پر کردن وسط متن اکد
—	<pre>{   "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 }</pre>	Authorization , Content-Type	https://alphapi.aip.sharif.ir/v1/chat/completions	POST (دوبار پیاپی)	<b>Context Caching</b> کش سمت‌سرو برای ورودی‌های تکراری
۴۰۱: توکن نامعتبر.	Body بدون	Authorization : (اختیاری) Accept: application/json	https://alphapi.aip.sharif.ir/v1/models	GET	<b>List Models</b> دیدن مدل‌های در دسترس
۴۰۱: مسیر غیرفعال/مجوز ندارید.	Body بدون	Authorization	https://alphapi.aip.sharif.ir/v1/user/balance	GET	<b>Get User Balance</b> دیدن موجودی
400 با پیام‌هایی مثل provider/model not provided: یعنی مدل embedding را روی	<pre>{   "model": "baai-bge-m3",   "input": ["This is a test"] }</pre>	Authorization , Content-Type	https://alphapi.aip.sharif.ir/v1/embeddings	POST	<b>Embeddings</b> بردارسازی متن برای مدل‌های Embedding که را ساپورت می‌کنند.

 راست جمیعی معاونت علمی و فناوری	<b>سکوی ملی متن باز هوش مصنوعی:</b> <b>راهنمای اتصال به مدل‌ها و دریافت پاسخ از طریق API</b>	
پیوست: ندارد	تاریخ: ۱۴۰۴/۱۱/۰۵	شماره: ۰۶

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

## ۲-۸ 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= مثلًا چطور مرخصی بگیرم؟
- 9. TOP\_K = 3
- 10. DOCS = آرایه‌ی Json اسناد

۳. فعال کردن Environment ساخته شده.

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. اسکریپت 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(/[\u062f\u0647]/g, "\u062f").replace(/ى/g, "\u0628").replace(/ك/g, "\u0627").replace(/[\u062a\u062b\u062c\u062d\u062e\u062f\u0628\u0629]/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);
```



رئاست جمهوری  
سازمان اسناد و کتابخانه ملی

سکوی ملی متن باز هوش مصنوعی:

راهنمای اتصال به مدل‌ها و دریافت پاسخ از طریق API

پیوست: ندارد

تاریخ: ۱۴۰۴/۱۱/۰۵

شماره: ۰۶



```
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:"."},
    {role:"user",content:` ${QUESTION}\n\n${CTX}\n\n${ANSWER}` }
  ]
};
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": "ثبت شود. تایید مدیر مستقیم لازم است HR درخواست مخصوصی باید در پورتال"
},
{
  "id": "Procurement",
  "source": "procurement.pdf",
  "text": "اجام می‌شود PR خرید تجهیزات فقط از طریق واحد تدارکات و ثبت فرم"
},
{
  "id": "Onboarding",
  "source": "onboarding.docx",
  "text": "فرایند استخدام شامل مصاحبه فنی و منابع انسانی است"
},
{
  "id": "Leave-Types",
  "source": "hr_policy.md",
  "text": "مرخصی استعلامی 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.