



中国科学院计算机网络信息中心  
Computer Network Information Center,  
Chinese Academy of Sciences



# 中国科技云MaaS平台

大模型API开放平台、AI工作台

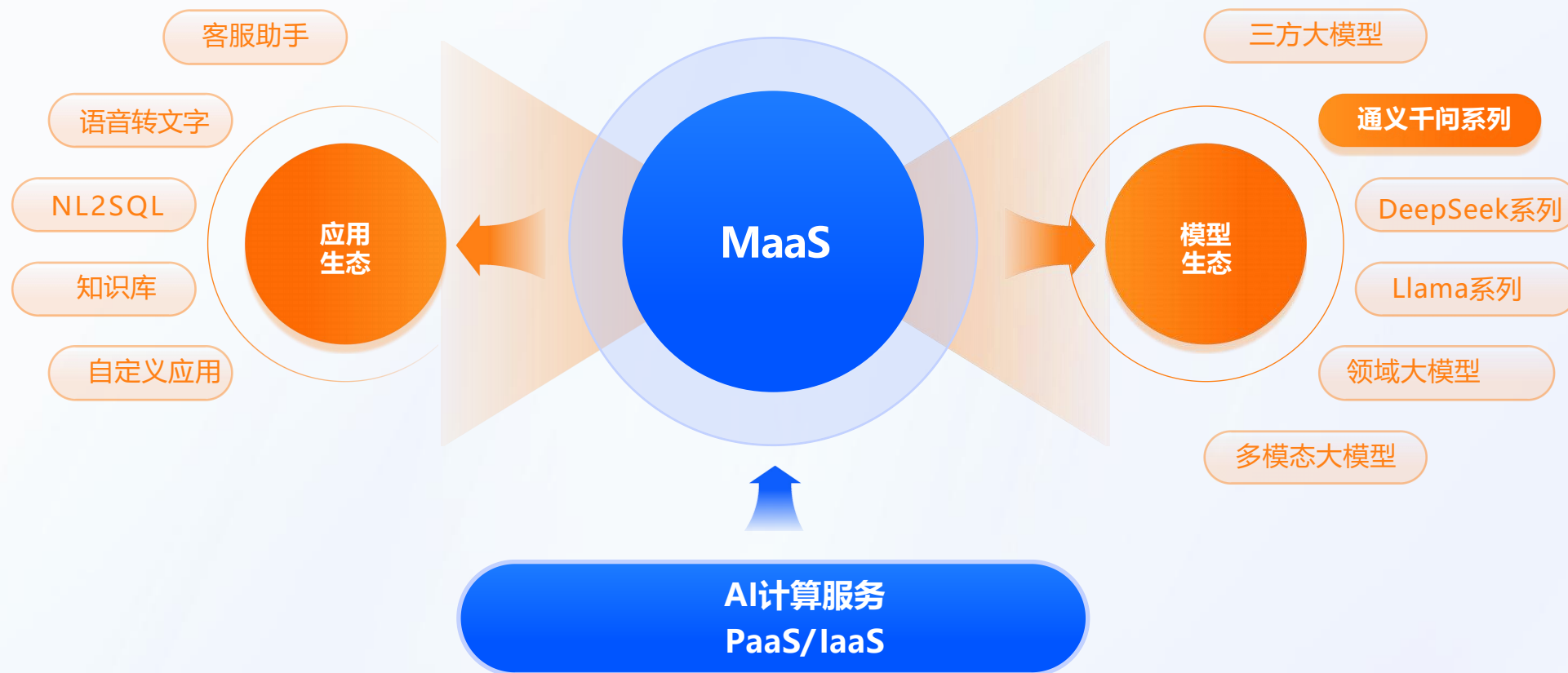
中国科学院计算机网络信息中心

裴昶华

2025年6月10日

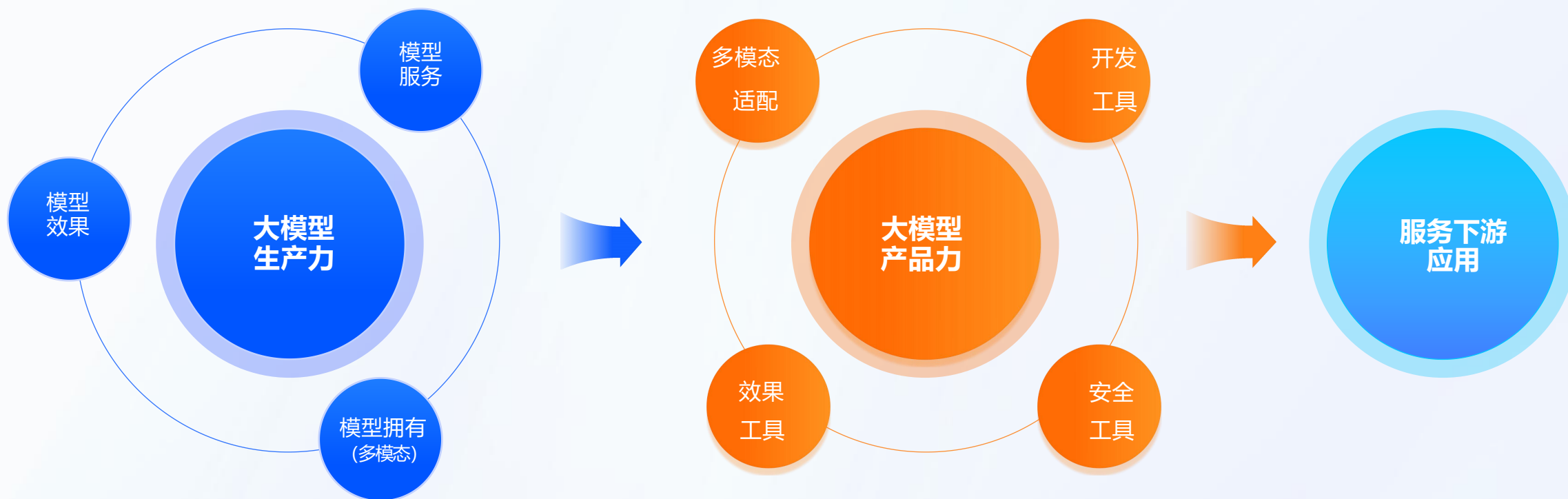
# MaaS平台定位

基于大模型，面向单位和个人开发者用户，打造一站式大模型服务和大模型应用构建平台



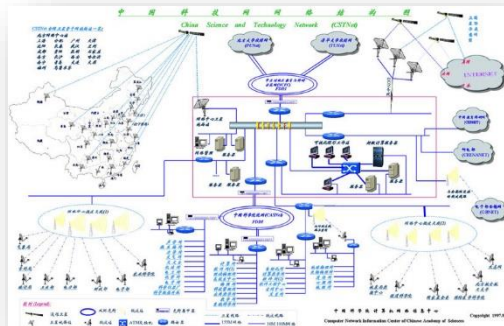
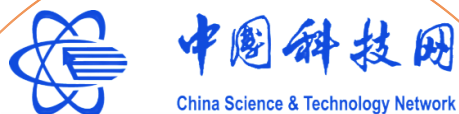
# 大模型服务及应用已逐步进入“深水区”

用户对效果需求带来对大模型平台能力，从要求**生产力**到**产品力**的转变



# 始终坚守服务科研创新的初心，不断前行

30年前实现中国全功能接入国际互联网，这标志着中国互联网时代的开端。

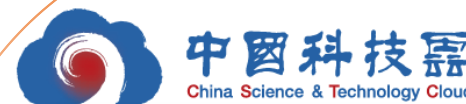


1994

从“网”到“云”演进



逐步实现“云网融合”  
正在迈向“云智网融合”



2018

下一个30年，奋楫扬帆，勇立时代潮头

# 中国科技云本地部署DeepSeek，提供三类服务

构建**API开放平台**助力**300+**项目推进，研发**对话服务**，上线**学术助手**、**团队知识库**





# 基于本地部署大模型，中国科技云提供一体化服务

针对不同科研应用场景，定制本地化部署和应用方案

01

API 接口定制开发

02

模型微调

03

RAG 知识库

04

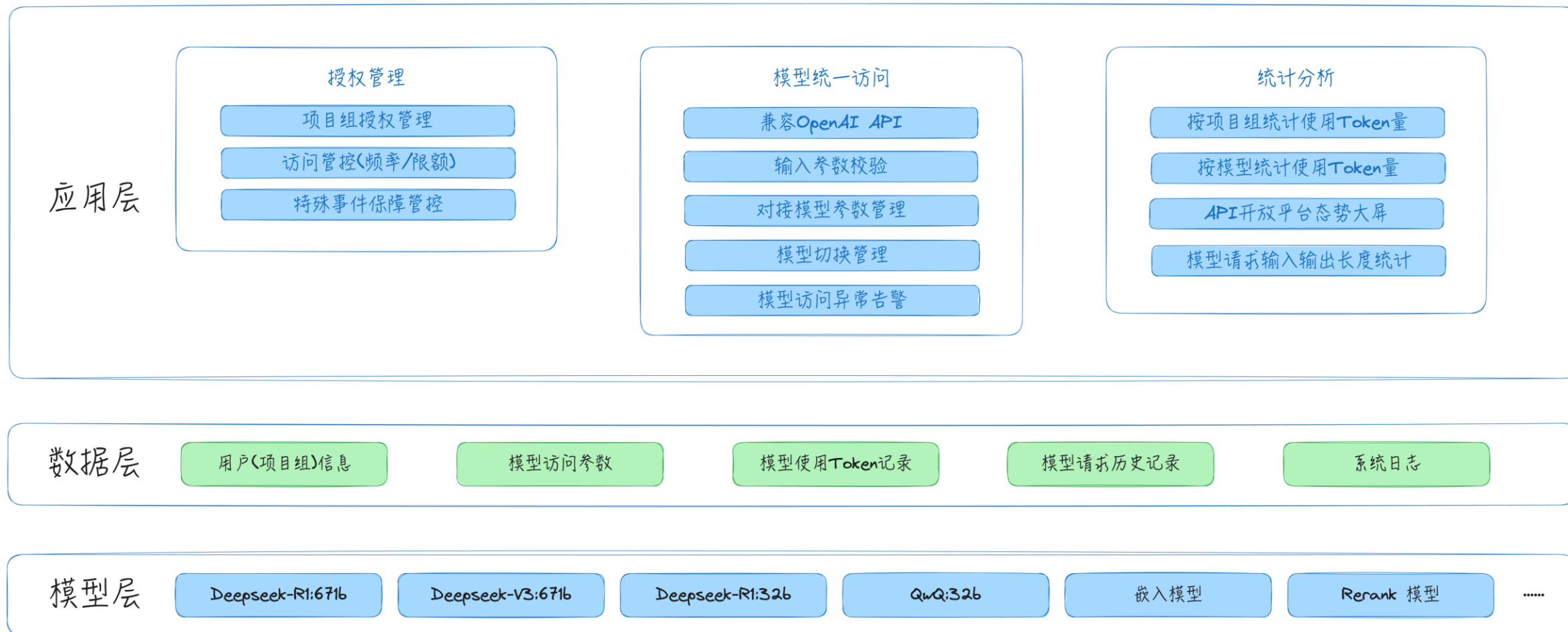
Agent开发

05

深度适配

# 中国科技云大模型 API 开放平台

## API 开放平台



# 大模型 API 服务



完全兼容 OpenAI API 格式，方便各类应用的直接接入 (如 ragflow, dify, openwebui等)



提供不同版本的模型选择 (deepseek-v3:671b, deepseek-r1:671b, deepseek-r1:32b, 嵌入模型, QwQ-32b) , 后续会增加 其他模型



使用API key进行认证，方便自研应用的接入，模型请求接口均返回请求消耗的token数，方便应用层对用户行为进行计量

已支持 300+ 课题组使用

API.md

2025-02-16

## 中国科技云 Deepseek 和相关模型服务 API 使用文档

API 文档是为开发者提供的接口调用说明,开发者可以通过 API 文档了解模型使用相关接口的调用方式,参数说明,返回结果等信息.

本文档中的全部 API 均为 OpenAI-API-Compatible, 除文档中描述的直接请求外, 均支持通过 openai sdk 调用. 只需用我们告知的 Token 替换 OPENAI\_API\_TOKEN 的值, 修改 BASE\_URL 即可.

### 基本说明

**BASE\_URL:** <https://uni-api.cstcloud.cn/v1>

**Token:** 联系 liuyude@cstnet.cn 获得

### Models

获取模型列表

请求方式: GET

请求地址: /models

请求参数: 无

```
curl -X GET "https://uni-api.cstcloud.cn/v1/models" -H "Authorization: Bearer {Token}"
```

返回结果: 模型列表

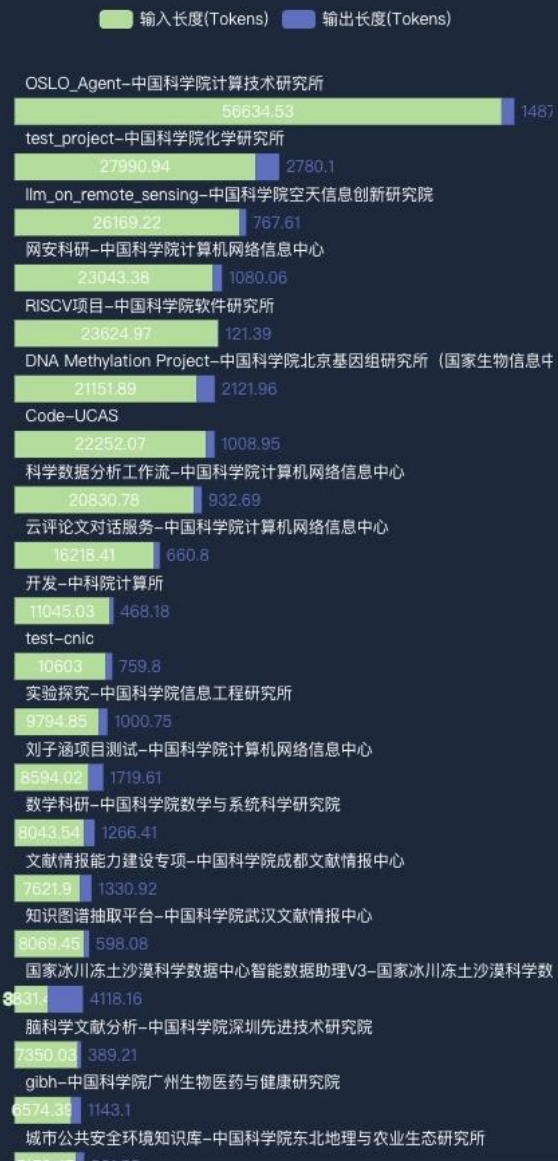
```
{
  "object": "list",
  "data": [
    {
      "id": "model-id-0",
      "object": "model",
      "created": 1686935002,
      "owned_by": "CSTCloud"
    },
    {
      "id": "model-id-1",
      "object": "model",
      "created": 1686935002,
```



# 支持模型列表

- DeepSeek-R1:671b – 0528版
- DeepSeek-V3:671b - 0324版
- DeepSeek-R1:32b
- 千问推理模型 QwQ:32b
- 千问 Qwen3:235b
- 星火推理模型 Spark-X1:70b
- 嵌入模型 gte-Qwen2-7B(嵌入维度 3584, 上下文长度 32K)
- 嵌入模型 bge-large-zh:latest
- Rerank模型 bge-reranker-v2-m3
- Web Search API / AI Search API
- 计划近期上线:
  - 千问多模态 qwen2.5-vl:72b
  - OCR 模型

API请求平均输入和平均输出Token数

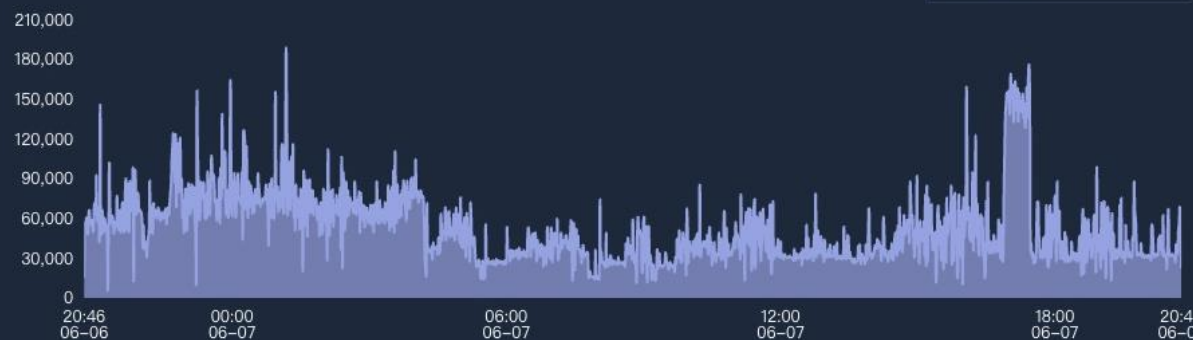


中国科技云 Deepseek API 使用态势

10,605,544,584

2025-02-16 以来通过 API 请求 Deepseek 模型消耗 Token 总量

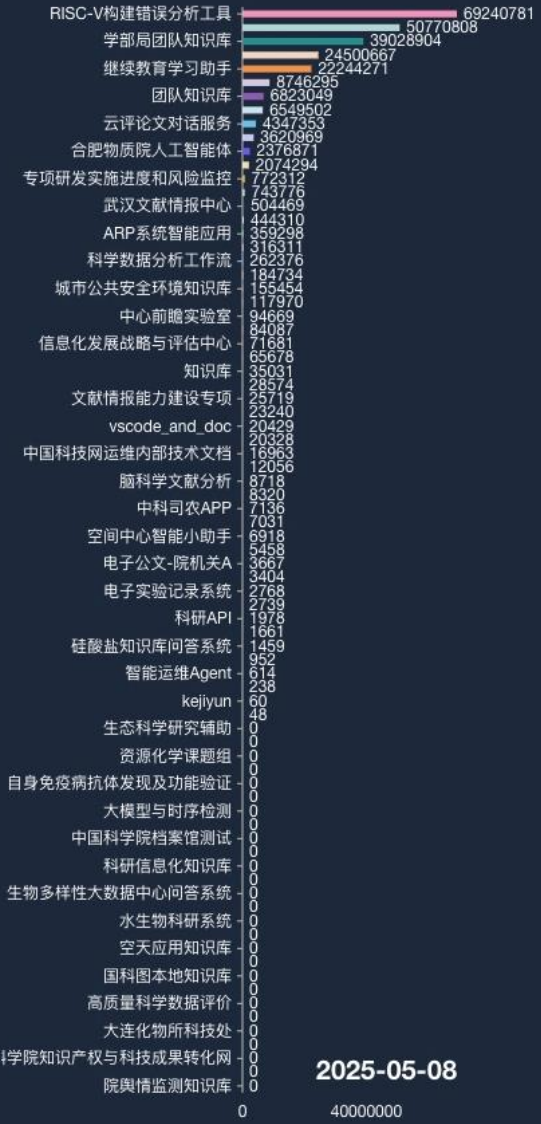
24小时API访问 - 按模型统计(Token数)



24小时API访问 - 按课题组统计(Token数)



每日 Token 使用量排名



2025-05-08

# 如何使用

- 浏览器访问 <https://uni-api.cstcloud.cn>

大模型 API 开放平台

我的 API Key

API 使用手册

API Key 管理

API Key 申请审核

新增模型授权审核

API Key 管理

Token 限额管理

用户授权管理

模型管理

模型对接信息管理

临时访问控制策略管理

使用统计

按模型统计 Token 用量

按 API Key 统计 Token 用量

态势大屏

退出系统

中国科大云大模型 API 开放平台

创建 & 管理 大模型 API Key

陈灿

输入申请信息

确认提交

名称(用于简短标识本 API Key 用途) \*

名称

联系人姓名 \*

联系人姓名

联系人 Email \*

联系人 Email

所在单位 \*

所在单位

联系电话 \*

联系电话

用途说明(描述大模型 API 将用于做什么研究) \*

用途说明...

选择需要授权的模型 \*

☐ deepseek-r1:32b-16k

☐ bge-large-zh:latest

☐ deepseek-v3:671b

☐ deepseek-r1:671b-64k

☐ bge-reranker-v2-m3

☐ qwq:32b

☐ deepseek-r1:32b

☐ spark-x1:70b

☐ gte-qwen2:7b

☐ qwen3:235b

☐ deepseek-r1:671b

Copyright © 2025 CNIC. All rights reserved. · v1.0.0 · 京ICP备09112257号-130 · 京公网安备11010802040559号

# 获取API Key & 如何使用 & 追加授权模型申请

中国科技云大模型API接口使用手册

说明

BASE\_URL

Models

Chats

Embeddings

Rerank

技术支持

模型发布记录

Q 搜索文档 36 K

Home

⚙

首页 > Chats

Chats

约 1539 字 大约 5 分钟 2025-04-25

此页内容

和大模型对话

特别提示

和大模型对话

请求方式: POST

请求地址: /chat/completions

请求参数:

- model: 必选, 模型名, 对应接口 /v1/models 中的 id, 例如 "model": "deepseek-r1:32b-16k"
- messages: 必选, 推理请求消息结构, list类型,0KB<messages内容包含的字符数<4MB,支持中英文。tokenizer之后的token数量小于或等于 maxInputTokenLen, maxSeqLen-1, max\_position\_embeddings 和 1MB 之间的最小值。其中,max\_position\_embeddings从权重文件 config.json 中获取,其他相关参数从配置文件中获取
  - role: 必选, 消息角色, 可选值为 "system" 或 "user" 或 "assistant"

提示

deepseek-r1 模型只支持 "user" 类型的消息

- content: 必选, 消息内容, 字符串类型

- stream: 可选, 是否为流式推理, 默认值为 false, [true: 流式推理, false: 文本推理]
- presence\_penalty: 可选, 存在惩罚介于-2.0和2.0之间,它影响模型如何根据到目前为止是否出现在文本中来惩罚新token。正值将通过惩罚已经使用的词,增加模型谈论新主题的可能性, 负值将通过惩罚已经使用的词,减少模型谈论新主题的可能性。默认值为 0.0, 可选范围[-2.0, 2.0]
- frequency\_penalty: 可选, 频率惩罚介于-2.0和2.0之间,它影响模型如何根据token的频率来惩罚新token。正值将通过惩罚常用词,增加模型谈论新主题的可能性, 负值将通过惩罚常用词,减少模型谈论新主题的可能性。默认值为 0.0, 可选范围[-2.0, 2.0]
- repetition\_penalty: 可选, 重复惩罚是一种技术,用于减少在文本生成过程中出现重复片段的概率。它对之前已经生成的文本进行惩罚,使得模型更倾向于选择新的, 不重复的内容。默认值为 1.0, 可选范围(0.0, 2.0]
- temperature: 可选, 温度参数介于0.0和1.0之间,它控制模型生成的文本的创造性。较低的温度会导致更可预测的文本,而较高的温度会导致更随机的文本。默认值为 1.0, 可选范围[0.0, 2.0], 0.0表示完全确定性,1.0表示完全随机性, deepseek-r1 模型官方建议设置为 0.6
- top\_p: 可选, 控制模型生成过程中考虑的词汇范围,使用累计概率选择候选词,直到累计概率超过给定的阈值。该参数也可以控制生成结果的多样性,它基于累积概率选择候选词,直到累计概率超过给定的阈值为止。默

# 如何使用 — 获取已授权的模型接口

获取模型列表

请求方式: GET

请求地址: /models

请求参数: 无

```
1 curl -X GET "https://uni-api.cstcloud.cn/v1/models" -H "Authorization: Bearer {Token}"
```

返回结果: 模型列表

```
1 {
2   "object": "list",
3   "data": [
4     {
5       "id": "model-id-0",
6       "object": "model",
7       "created": 1686935002,
8       "owned_by": "CSTCloud"
9     },
10    {
11      "id": "model-id-1",
12      "object": "model",
13      "created": 1686935002,
14      "owned_by": "CSTCloud",
15    },
16    {
17      "id": "model-id-2",
18      "object": "model",
19      "created": 1686935002,
20      "owned_by": "CSTCloud"
21    },
22  ]
23 }
```

- 快速研发对话服务(更新两个大版本, 几十个小版本), 基于中国科技云认证联盟进行身份认证, 提供多轮对话、对话列表保存、联网搜索、文档上传等功能。 <https://deepseek.ai4cas.cn>

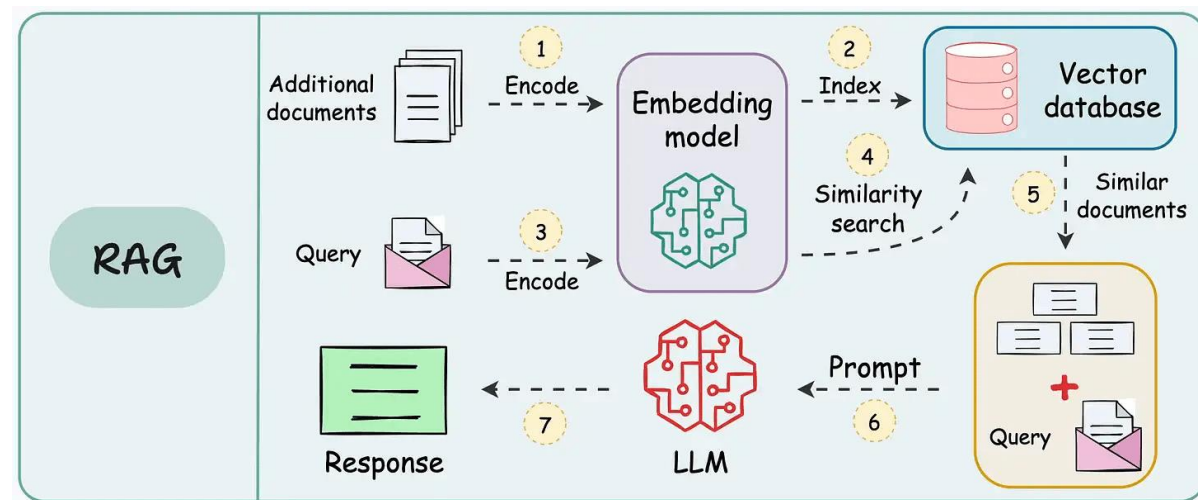
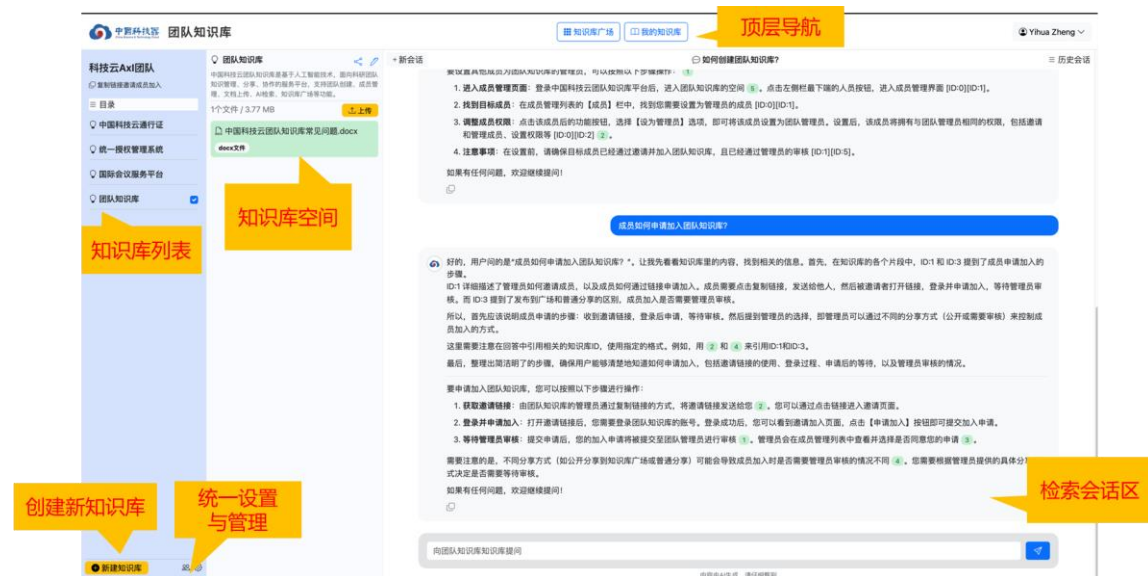




# 自主研发团队知识库 <https://kb.ai4cas.cn>

## ● AI时代的团队知识库

- 面向团队，支持团队创建、邀请成员，协同管理团队文档
- 团队内支持多知识库创建
- 团队内支持多知识库联合对话
- 支持多种方式的知识库分享
  - 知识库广场公开分享
  - 私密分享，授权加入
- DeepSeek智能支持
- 简洁友好，文件上传即刻具有智能能力



- 开启新会话
- 助手列表
- 历史对话
- 你是一名科研类的英...
- 时序分析有哪些相关...
- 作为一名中文学术论...
- 帮我搜索有没有time ...

提示语设定

工作流程跟踪

学术助手: <https://academic-new.ai4cas.cn>

你好，我是云小助

作为你的学术小助手，我既能思、能写、能总结，也能陪你聊个天，解个闷。

**智能对话**

您可同云小助直接对话，问这问那都能行

**提示语设定**

通过设定提示语，引导大模型完成学术任务

**智能体工作流**

链式自动化完成复杂的学术搜索、深度学习等任务

- 学术润色
- 中英互译
- 学术搜索(云评版)
- 论文总结
- 内容改写
- 要点精炼

- 深度学术(云评版)
- 撰写文章标题

请输入对话内容

☐ 深度思考

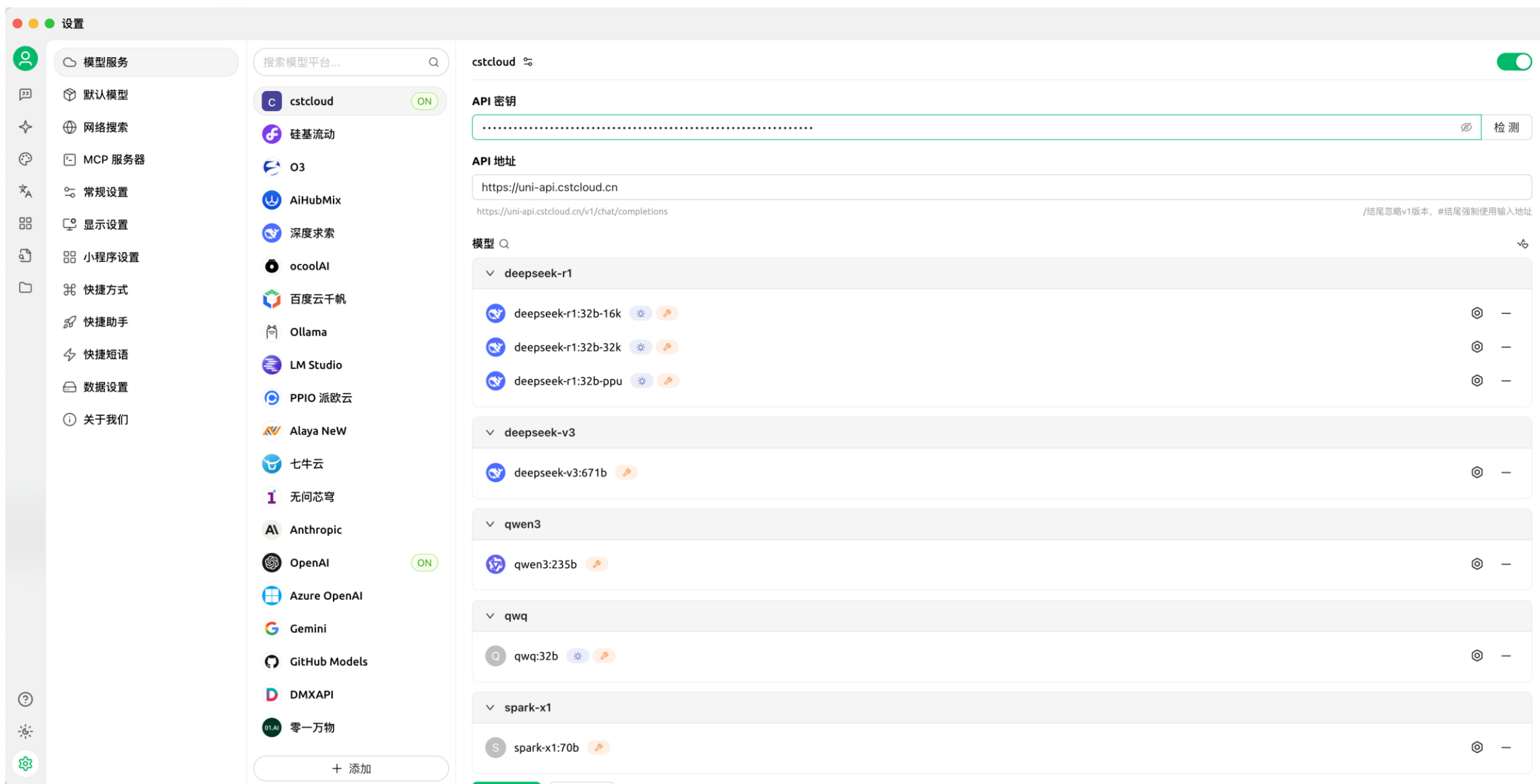
# 如何使用 – 使用 langchainjs

```
test_langchain.js > client > mcpServers > "timezone-toolkit" > restart
26 const model = new ChatOpenAI({
27   temperature: 0,
28   configuration: {
29     baseUrl: "https://uni-api.cstcloud.cn/v1",
30   },
31 });
32 };
33
34 const agent = createReactAgent({
35   llm: model,
36   tools,
37 });
38
39 try {
40   const mathResponse = await agent.streamEvents({
41     messages: [{ role: "user", content: "现在北京什么时间" }],
42   }, {
43     version: 'v2',
44   },);
45   for await (const chunk of mathResponse) {
46     console.log(chunk.data);
47   }
48 } catch (error) {
49   console.error("Error during agent execution:", error);
50   // Tools throw ToolException for tool-specific errors
51   if (error.name === "ToolException") {
52     console.error("Tool execution failed:", error.message);
53   }
54 }
55
56 await client.close();
```

```
问题 输出 调试控制台 终端 端口 GITLENS
+ node 窗 ... ^ x

"name": "mcp_timezone-toolkit_get_current_time",
"additional_kwargs": {},
"response_metadata": {},
"tool_call_id": "chatcmpl-tool-6ecf7a0f8d5e4718a701f2aa2ed64285",
"artifact": []
},
AIMessageChunk {
  "id": "chatcmpl-c9e22f993c5749d3a609cce864f5d996",
  "content": "好的，我现在需要处理用户的问题“现在北京什么时间”。首先，用户询问的是北京当前的时间，所以我需要确定北京所在的时区。北京属于中国标准时间，对应的IANA时区名称是“Asia/Shanghai”。\n\n接下来，我要查看可用的工具函数。发现有一个叫‘mcp_timezone-toolkit_get_current_time’的函数，这个函数可以获取指定时区的当前时间，参数需要时区名称。因此，我应该调用这个函数，并将时区参数设为“Asia/Shanghai”。\n\n用户没有指定输出格式，所以使用默认的“medium”格式。然后，我需要构造一个工具调用的JSON对象，包含函数名称和参数。确认参数中的时区正确无误，并且满足函数所需的必填字段。最后，确保返回的工具调用格式正确，放在指定的XML标签中。这样系统就能正确执行函数并返回北京当前时间的信息了。</think>\n\n当前北京时间是：**5月14日中午11点01分**（时区：Asia/Shanghai, UTC+8）。",
  "additional_kwargs": {},
  "response_metadata": {
    "usage": {
      "prompt_tokens": 1513,
      "total_tokens": 1742,
      "completion_tokens": 229
    }
  },
  "tool_calls": [],
  "tool_call_chunks": [],
```

# 如何使用 – 对接工具



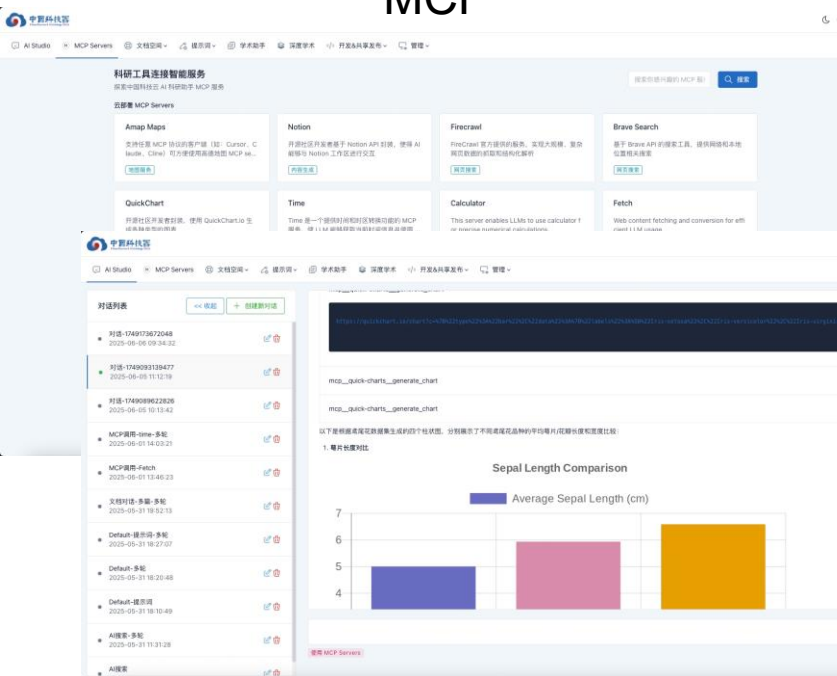
# 中国科技云 AI 工作台

- MCP Servers 共享平台 / MCP Hub / MCP 在线使用平台
- 大模型支持的提示词生成/优化/共享平台
- 论文深度解读

MCP

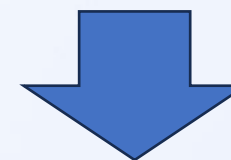
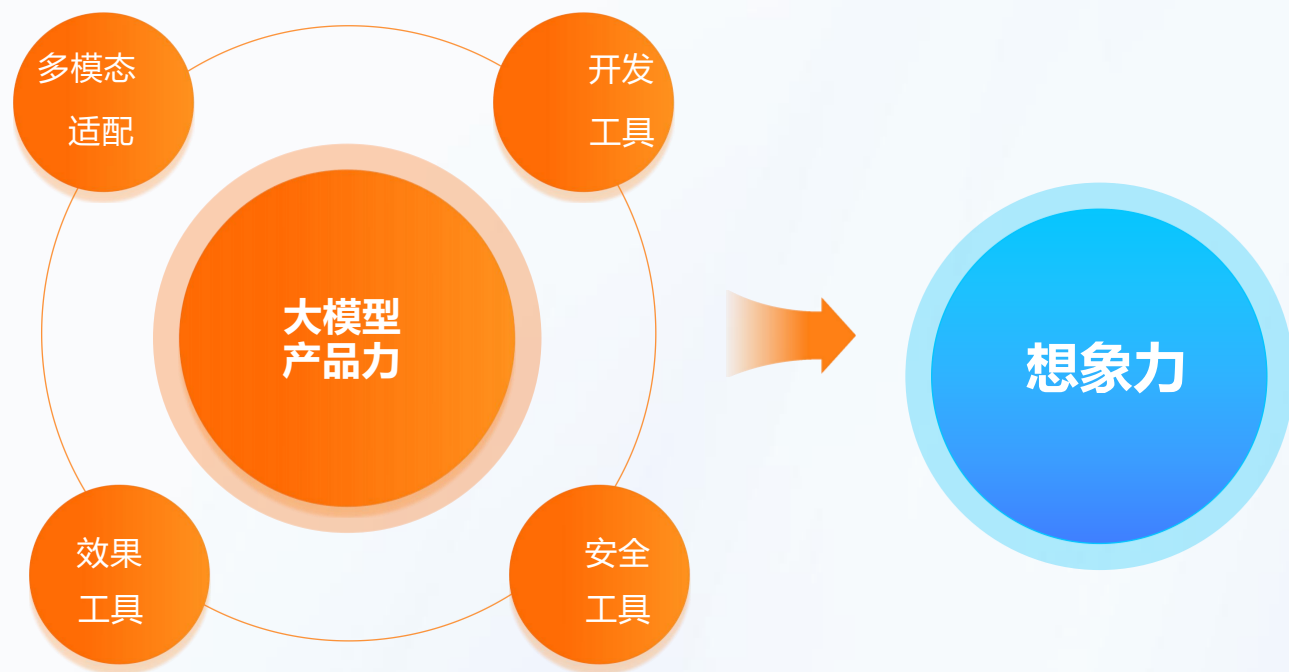
提示词

论文解读





# 大模型服务及应用已进入“乐高化”



除了提供MCP server，更多的是通过MCP server组合，售卖“想象力”



# 欢迎大家试用！

# 谢谢！