**文本生成 - 千帆AI应用开发者中心-API参考qianfan-api | 百度智能云文档**

**文本生成**

POST

https://qianfan.baidubce.com/v2/chat/completions

调用本接口，发起一次对话请求。

**支持模型列表**

支持模型列表请查看 [千帆ModelBuilder-模型列表-文本生成](https://cloud.baidu.com/doc/WENXINWORKSHOP/s/Wm9cvy6rl#%E6%96%87%E6%9C%AC%E7%94%9F%E6%88%90) 。

**权限说明**

调用本文API，需使用API Key鉴权方式。使用API Key鉴权调用API流程，具体调用流程，请查看 [认证鉴权](https://cloud.baidu.com/doc/qianfan-api/s/ym9chdsy5) 。

**请求参数**

**Headers 参数**

在 HTTP 请求 Header 中，将字符串Bearer和API Key值拼接，添加到Authorization。

**Body 参数**

**model string 必选**

模型ID，说明：  
（1）预置服务，可选值请参考 [千帆ModelBuilder-模型列表-文本生成](https://cloud.baidu.com/doc/WENXINWORKSHOP/s/Wm9cvy6rl#%E6%96%87%E6%9C%AC%E7%94%9F%E6%88%90) 页，表格中model入参列  
（2）平台训练模型或预置模型，可以通过查看服务详情获取该字段值，详情请查看 [千帆控制台-在线推理](https://console.bce.baidu.com/qianfan/ais/console/onlineService) ：model值为服务详情中对应的API名称，如下图所示：

**messages array 必选**

聊天上下文信息。说明：  
（1）messages成员不能为空，1个成员表示单轮对话，多个成员表示多轮对话，例如：  
· 1个成员示例， "messages": [ {"role": "user","content": "你好"}]  
· 3个成员示例， "messages": [ {"role": "user","content": "你好"},{"role":"assistant","content":"需要什么帮助"},{"role":"user","content":"自我介绍下"}]   
（2）message中的content总长度不能超过对应model的输入字符限制和输入tokens限制

* 子属性：items object {5}
* role string 必选

当前支持以下：  
· user: 表示用户  
· assistant: 表示对话助手  
· system：表示人设，当模型为GLM-Z1-Rumination-32B-0414时，不支持该字段值

* name string 可选
* content anyOf {2} 多选一，只需要符合下列任意一组子节点 必选

对话内容，说明：

（1）不能为空

（2）最后一个message对应的content不能为blank字符，如空格、"\n"、“\r”、“\f”等

* content string
* content array
* items string
* tool\_calls array 可选

函数调用，function call场景下第一轮对话的返回，第二轮对话作为历史信息在message中传入

* tool\_call\_id string 可选

说明：

（1）当role为tool时，该字段必填

（2）模型生成的function call id，对应tool\_calls中的tool\_calls[].id

（3）调用方应该传递真实的、由模型生成id，否则效果有损

**stream boolean 可选**

是否以流式接口的形式返回数据，说明：  
（1）可选值：  
· true ：是，按SSE协议逐块返回内容，以一条 data: [DONE] 消息结束  
· false ：否，默认 false  
（2）beam search模型只能为 false

**stream\_options object {2} 可选**

流式响应的选项，当字段stream为 true 时，该字段生效

* include\_usage boolean 可选

流式响应是否输出usage，说明：  
· true ：是，设置为 true 时，在最后一个chunk会输出一个字段，这个chunk上的usage字段显示整个请求的token统计信息  
· false ：否，流式响应默认不输出usage

* chunk\_include\_usage boolean 可选

流式响应每一包是否都带usage信息，说明：  
· true ：是，设置为 true 时，每一个chunk都带有截止目前chunk为止的usage信息。

**temperature number 可选**

说明：  
（1）较高的数值会使输出更加随机，而较低的数值会使其更加集中和确定  
（2）该参数支持模型及取值范围等，请参考 [千帆ModelBuilder-模型默认参数说明](https://cloud.baidu.com/doc/WENXINWORKSHOP/s/kmamoewth)

**top\_p number 可选**

说明：  
（1）影响输出文本的多样性，取值越大，生成文本的多样性越强  
（2）该参数支持模型及取值范围等，请参考 [千帆ModelBuilder-模型默认参数说明](https://cloud.baidu.com/doc/WENXINWORKSHOP/s/kmamoewth)

**penalty\_score number 可选**

通过对已生成的token增加惩罚，减少重复生成的现象。说明：  
（1）值越大表示惩罚越大  
（2）默认1.0，取值范围：[1.0, 2.0]  
（3）不支持以下模型：  
· DeepSeek-V3  
· DeepSeek-Reasoner  
· DeepSeek-R1-Distill系列  
· QwQ-32B  
· ERNIE X1 Turbo系列  
· ERNIE 4.5系列：ERNIE-4.5-0.3B、ERNIE-4.5-21B-A3B、ERNIE-4.5-VL-28B-A3B

**max\_tokens integer 可选**

指定模型最大输出token数。

**seed integer**

说明：  
（1）取值范围: （0,2147483647），会由模型随机生成，默认值为空  
（2）如果指定，系统将尽最大努力进行确定性采样，以便使用相同seed和参数的重复请求返回相同的结果  
（3）不支持以下模型：  
· ERNIE X1 Turbo系列  
· Qwen2.5  
· Qianfan-Agent-Intent

可选

**stop array 可选**

生成停止标识，当模型生成结果以stop中某个元素结尾时，停止文本生成。说明：  
（1）每个元素长度不超过20字符  
（2）最多4个元素  
（3）不支持以下模型：ERNIE X1 Turbo系列

* items string 可选

**frequency\_penalty number 可选**

说明：  
（1）正值根据新token在文本中的出现频率来对其进行惩罚，从而降低模型逐字重复的可能性  
（2）该参数支持模型及取值范围等，请参考 [千帆ModelBuilder-模型默认参数说明](https://cloud.baidu.com/doc/WENXINWORKSHOP/s/kmamoewth)

**presence\_penalty number 可选**

说明：  
（1）正值根据新token到目前为止是否出现在文本中来对其进行惩罚，从而增加模型谈论新主题的可能性  
（2）该参数支持模型及取值范围等，请参考 [千帆ModelBuilder-模型默认参数说明](https://cloud.baidu.com/doc/WENXINWORKSHOP/s/kmamoewth)

**repetition\_penalty number 可选**

说明：  
（1）控制模型生成文本时连续序列中的重复度，提高repetition\_penalty可以降低模型生成的重复度  
（2）该参数支持模型及取值范围等，请参考 [千帆ModelBuilder-模型默认参数说明](https://cloud.baidu.com/doc/WENXINWORKSHOP/s/kmamoewth)

**tools array**

一个可触发函数的描述列表，支持模型请参考 [开始使用-模型-Function calling-支持模型范围](https://cloud.baidu.com/doc/qianfan-docs/s/xm95lyys5#%E6%94%AF%E6%8C%81%E6%A8%A1%E5%9E%8B%E8%8C%83%E5%9B%B4) 。

可选

* items object {2}
* type string 必选

工具类型，取值function

* function object {3} 必选

函数说明

* name string 必选

函数名

* description string 可选

函数描述

* parameters object {0} 可选

函数请求参数，JSON Schema 格式，参考[JSON Schema描述](https://json-schema.org/understanding-json-schema/)

**tool\_choice anyOf {2} 可选**

多选一，只需要符合下列任意一组子节点

说明：  
（1）支持模型请参考 [开始使用-模型-Function calling-支持模型范围](https://cloud.baidu.com/doc/qianfan-docs/s/xm95lyys5#%E6%94%AF%E6%8C%81%E6%A8%A1%E5%9E%8B%E8%8C%83%E5%9B%B4)  
（2）可选值如下：  
· none ：不希望模型调用任何function，只生成面向用户的文本消息  
· auto ：模型会根据输入内容自动决定是否调用函数以及调用哪些function  
· required ：希望模型总是调用一个或多个function。  
· 通过 {"type": "function", "function": {"name": "my\_function"}} 指定特定 tool，会强制模型调用该 tool。  
（3）当为tool\_choice类型，指在函数调用场景下，提示大模型选择指定的函数，指定的函数名必须在tools中存在

* tool\_choice string

说明：  
· none ：不希望模型调用任何function，只生成面向用户的文本消息  
· auto ：模型会根据输入内容自动决定是否调用函数以及调用哪些function  
· required ：希望模型总是调用一个或多个function

* tool\_choice object {2}
* type string 必选

指定工具类型，固定值function

* function object {1} 必选

指定要使用的函数

* name string 必选

指定要使用的函数名

**parallel\_tool\_calls boolean 可选**

说明：  
（1）支持模型请参考 [开始使用-模型-Function calling-支持模型范围](https://cloud.baidu.com/doc/qianfan-docs/s/xm95lyys5#%E6%94%AF%E6%8C%81%E6%A8%A1%E5%9E%8B%E8%8C%83%E5%9B%B4)  
（2）可选值：  
· true ：表示开启函数并行调用，默认开启  
· false ：表示关闭函数并行调用

**web\_search object {7} 可选**

搜索增强的选项，说明：  
（1）默认不传关闭  
（2）支持模型请参考 [开始使用-模型-联网搜索-支持模型列表](https://cloud.baidu.com/doc/qianfan-docs/s/Wm8r4sw29#%E6%94%AF%E6%8C%81%E6%A8%A1%E5%9E%8B%E5%88%97%E8%A1%A8)

* enable boolean 可选

是否开启实时搜索功能，说明：  
（1）如果关闭实时搜索，角标和溯源信息都不会返回  
（2）可选值：  
· true ：开启  
· false ：关闭，默认 false

* enable\_citation boolean 可选

是否开启上角标返回，说明：  
（1）enable为 true 时生效  
（2）可选值：  
· true ：开启；如果开启，在触发了搜索增强的场景下，响应内容会附上角标，并带上角标对应的搜索溯源信息  
· false ：未开启，默认 false  
（3）如果检索内容包含非公开网页，角标不生效

* enable\_trace boolean 可选

是否返回搜索溯源信息，说明：  
（1）enable 为 true 时生效  
（2）可选值：  
· true ：返回；如果为 true ，在触发了搜索增强的场景下，会返回搜索溯源信息search\_results  
· false ：不返回，默认 false  
（3）如果检索内容为非公开网页，即使触发搜索也不返回溯源信息

* enable\_status boolean 可选

是否返回搜索信号，说明：  
（1）enable 为 true 时生效  
（2） 可选值：  
· true ：返回；如果为 true 触发搜索，会通过 delta\_tag:search\_status 表示这一包是信号包。  
· false ：不返回，默认 false

* search\_mode string 可选

联网搜索模式：  
- auto:大模型基于query判断意图是否需要进行搜索。举例：用户query 1+1等于几 ，默认不过联网搜索  
- required:强制进行联网搜索  
- ernie系列模型不支持该参数

* search\_number integer 可选

检索的文献数量，范围在[1~28]之间

* reference\_number integer 可选

用于给大模型总结的文献数量，范围在[1~28]之间（需≤search\_number，若 reference\_number > search\_number 则将默认赋值 reference\_number=search\_number,如设置：search\_number=6，reference\_number=10，则search\_number=10）

**response\_format object {2} 可选**

指定响应内容的格式，说明：  
（1）对于生成式人工智能大模型，可能会出现效果不满足的情况  
（2）不支持以下模型：ERNIE X1 Turbo系列

* type string 可选

指定响应内容的格式，可选值：  
· json\_object ：以json格式返回，可能出现不满足效果情况  
· text ：以文本格式返回，默认为 text  
· json\_schema ：以json\_scheam规定的格式返回

* json\_schema object {0} 可选

json\_schema格式，请参考 [JSON Schema描述](https://json-schema.org/understanding-json-schema) ；当type为json\_schema时，该参数必填

**metadata map<string,string> 可选**

给每条请求增加自定义标签，可以在日志挖掘环节 [精细化遴选推理日志](https://cloud.baidu.com/doc/WENXINWORKSHOP/s/Gliu6ipor#%E5%85%B3%E4%BA%8Emetadata%E5%8F%82%E6%95%B0) ；说明：  
（1）元素个数最大支持16个  
（2）key和value必须都是string类型

**enable\_thinking boolean 可选**

是否开启思考模式，说明：  
（1）可选值：  
· true ：开启  
· false ：未开启，默认值为 false  
（2）更多内容请参考 [深度思考](https://cloud.baidu.com/doc/qianfan-docs/s/Wm95lyynv)

**thinking\_budget integer 可选**

思维链的最大长度，当模型思考过程生成的Token数超过thinking\_budget时，推理内容会进行截断并立刻开始生成最终回复。说明：  
（1）默认为16384，最小值为100，最大值为各模型支持的思维链长度，详情请参考 [千帆ModelBuilder-模型列表-深度思考](https://cloud.baidu.com/doc/WENXINWORKSHOP/s/Wm9cvy6rl#%E6%B7%B1%E5%BA%A6%E6%80%9D%E8%80%83)  
（2）适用于部分深度思考模型，更多内容请参考 [深度思考](https://cloud.baidu.com/doc/qianfan-docs/s/Wm95lyynv)

**thinking\_strategy string 可选**

思考策略，主要用于减少思维链输出。说明：  
（1）可选值：  
· short\_think ：简短思考  
· chain\_of\_draft ：Chain-of-Draft（草稿链）式思考  
（2）更多内容请参考 [深度思考](https://cloud.baidu.com/doc/qianfan-docs/s/Wm95lyynv)

**reasoning\_effort string 可选**

推理强度，平衡响应速度、输出内容的推理深度及 token 消耗。说明：  
（1）可选值：  
· low  
· medium ：默认值  
· high  
（2）更多内容请参考 [深度思考](https://cloud.baidu.com/doc/qianfan-docs/s/Wm95lyynv)

**user string 可选**

表示最终用户的唯一标识符

请求结构

|  |
| --- |
| Plaintext POST /v2/chat/completions HTTP/1.1Host: qianfan.baidubce.comAuthorization: authorization string{    "model": "deepseek-v3.1-250821",    "messages": [        {            "role": "system",            "content": "You are a helpful assistant."        },        {            "role": "user",            "content": "你好"        }    ]} |

**返回响应**

**Headers 参数**

除公共头域外，还包含以下特殊头域

X-Ratelimit-Limit-Requests integer

一分钟内允许的最大请求次数

X-Ratelimit-Limit-Input-Tokens integer

一分钟内允许的最大输入tokens消耗

X-Ratelimit-Limit-Output-Tokens integer

一分钟内允许的最大输出tokens消耗

X-Ratelimit-Remaining-Requests integer

达到RPM速率限制前，剩余可发送的请求数配额，如果配额用完，将会在0-60s后刷新

X-Ratelimit-Remaining-Input-Tokens integer

达到TPM速率限制前，剩余可消耗的输入tokens数配额，如果配额用完，将会在0-60s后刷新

X-Ratelimit-Remaining-Output-Tokens intege

达到TPM速率限制前，剩余可消耗的输出tokens数配额，如果配额用完，将会在0-60s后刷

**返回参数**

**id string**

本次请求的唯一标识，可用于排查问题

**object string**

回包类型 chat.completion：多轮对话返回

**created integer**

时间戳

**model string**

说明：  
（1）如果是预置服务，返回模型ID  
（2）如果是sft后部署的服务，该字段返回model:modelversionID，model与请求参数相同，是本次请求使用的大模型ID；modelversionID用于溯源

**choices anyOf {2}**

多选一，只需要符合下列任意一组子节点

* choices object {5}

stream=false时，返回该内容，返回类型为choices

* index integer 可选  
  choice列表中的序号
* message object {6} 可选  
  choice列表中的序号
* role string 可选

当前支持以下：

· user: 表示用户

· assistant: 表示对话助手

· system：表示人设

* name string 可选

message名

* content string 可选

对话内容

* tool\_calls array 可选

函数调用，function call场景下第一轮对话的返回，第二轮对话作为历史信息在message中传入

* items object {3}
* idstring

function call的唯一标识，由模型生成

* typestring

固定值function

* functionobject {2}

function call的具体内容

* namestring

函数名称

* argumentsstring

函数参数

* tool\_call\_idstring

说明：

（1）当role=tool时，该字段必填

（2）模型生成的function call id，对应tool\_calls中的tool\_calls[].id

（3）调用方应该传递真实的、由模型生成id，否则效果有损

* reasoning\_contentstring

思维链内容，支持[深度思考](https://cloud.baidu.com/doc/WENXINWORKSHOP/s/Wm9cvy6rl#%E6%B7%B1%E5%BA%A6%E6%80%9D%E8%80%83)模型。

* finish\_reason string 可选  
  输出内容标识，说明：

· stop：模型自然停止或命中提供的停止序列

· length：达到了最大的token数

· content\_filter：输出内容被截断、兜底、替换为\*\*等

· tool\_calls：函数调用

* flag integer 可选  
  安全细分类型，说明：

当stream=false，flag值含义如下：

· 0或不返回：安全

· 1：低危不安全场景，可以继续对话

· 2：禁聊：不允许继续对话，但是可以展示内容

· 3：禁止上屏：不允许继续对话且不能上屏展示

· 4：撤屏

* ban\_round integer 可选  
  当flag 不为 0 时，该字段会告知第几轮对话有敏感信息；如果是当前问题，ban\_round = -1
* choicesobject {6}

stream=true时，返回该内容，返回类型为see\_choices

* index integer 可选

choice列表中的序号

* delta object {3} 可选

响应信息，当stream=true时返回

* role string 可选

仅在流式第一帧返回

* content string 可选

流式响应内容

* tool\_calls array

由模型生成的函数调用，包含函数名称，和调用参数

* items object {3}
* id string 可选

function call的唯一标识，由模型生成

* type string 可选

固定值function

* function object {2} 可选

function call的具体内容

* name string 可选

函数名称

* arguments string 可选

函数参数

* delta\_tag string 可选

响应信息标识，search\_status：触发搜索信号

* finish\_reason string 可选

输出内容标识，说明：

· stop：模型自然停止或命中提供的停止序列

· length：达到了最大的token数

· content\_filter：输出内容被截断、兜底、替换为\*\*等

· tool\_calls：函数调用

* flag integer 可选

安全细分类型，说明：当stream=true时，返回flag表示触发安全

* ban\_round integer

当flag 不为 0 时，该字段会告知第几轮对话有敏感信息；如果是当前问题，ban\_round = -1

响应示例 function call（第一次响应） function call（第二次响应）

JSON

|  |
| --- |
| Plaintext {    "id": "as-qsp8w7ppnv",    "object": "chat.completion",    "created": 1755938117,    "model": "deepseek-v3.1-250821",    "choices": [        {            "index": 0,            "message": {                "role": "assistant",                "content": "你好！很高兴和你交流。请问有什么我可以帮助你的吗？"            },            "finish\_reason": "stop",            "flag": 0        }    ],    "usage": {        "prompt\_tokens": 11,        "completion\_tokens": 15,        "total\_tokens": 26    }} |

**错误码**

如果请求错误，服务器返回的JSON文本包含以下参数。

|  |  |
| --- | --- |
| 名称 | 描述 |
| code | 错误码 |
| message | 错误描述信息，帮助理解和解决发生的错误 |
| type | 错误类型 |