单个文件解析

创建解析任务

接口说明

适用于通过 API 创建解析任务的场景，用户须先申请 Token。 注意：

- 单个文件大小不能超过 200MB,文件页数不超出 600 页
- 每个账号每天享有 2000 页最高优先级解析额度，超过 2000 页的部分优先级降低
- 因网络限制，github、aws 等国外 URL 会请求超时
- 该接口不支持文件直接上传
- header头中需要包含 Authorization 字段，格式为 Bearer + 空格 + Token

Python 请求示例

```python
import requests

token = "官网申请的api token"
url = "https://mineru.net/api/v4/extract/task"
header = {
    "Content-Type": "application/json",
    "Authorization": f"Bearer {token}"
}
data = {
    "url": "https://cdn-mineru.openxlab.org.cn/demo/example.pdf",
    "model_version": "vlm"
}
```

```
res = requests.post(url,headers=header,json=data)
print(res.status_code)
print(res.json())
print(res.json()["data"])
```

CURL 请求示例

```
curl --location --request POST 'https://mineru.net/api/v4/extract/task'
\
--header 'Authorization: Bearer ***' \
--header 'Content-Type: application/json' \
--header 'Accept: */*' \
--data-raw '{
    "url": "https://cdn-mineru.openxlab.org.cn/demo/example.pdf",
    "model_version": "vlm"
}'
```

请求体参数说明

参数	类型	是否必选	示例	
url	string	是	https://static.openxlab.org.cn/opendatalab/pdf/demo.pdf	文件 URL，支持.pdf、.doc、.docx、格式
is_ocr	bool	否	false	是否启动 ocr 功能，默

参数	类型	是否必选	示例	
enable_formula	bool	否	true	是否开启公式识别，默
enable_table	bool	否	true	是否开启表格识别，默
language	string	否	ch	指定文档语言，默认 c https://www.paddleo OCRv5/PP-OCRv5_n pipeline模型有效
data_id	string	否	abc**	解析对象对应的数据 ID （_）、短划线（-）、英 符，可以用于唯一标识
callback	string	否	http://127.0.0.1/callback	解析结果回调通知您的 议的地址。该字段为空 callback 接口必须支持 Type:application/json content。解析接口按照 content，调用您的 ca checksum：字符串格式 字符串，通过 SHA256 查询。为防篡改，您可 成字符串，与 checksu content：JSON 字符串 象。关于 content 结果

参数	类型	是否必选	示例	
				例，对应任务查询结果 说明:您的服务端 callb 结果后，如果返回的 H 功，其他的 HTTP 状态 mineru 将最多重复推送 推送 5 次后仍未接收成 callback 接口的状态。
seed	string	否	abc**	随机字符串，该值用于 数字、下划线（_）组成 用于在接收到内容安全 服务发起。 说明：当使用 callback
extra_formats	[string]	否	["docx","html"]	markdown、json为默认 docx、html、latex三种
page_ranges	string	否	1–600	指定页码范围，格式为 表示选取第2页、第4页 [2,4,5,6])；"2––2"：表 中"–2"表示倒数第二页
model_version	string	否	vlm	mineru模型版本，两个

请求体示例

```
{
  "url": "https://static.openxlab.org.cn/opendatalab/pdf/demo.pdf",
  "model_version": "vlm",
  "data_id": "abcd"
}
```

响应参数说明

参数	类型	示例	说明
code	int	0	接口状态码，成功：0
msg	string	ok	接口处理信息，成功："ok"
trace_id	string	c876cd60b202f2396de1f9e39a1b0172	请求 ID
data.task_id	string	a90e6ab6-44f3-4554-b459-b62fe4c6b436	提取任务 id，可用于查询任务结果

响应示例

```
{
  "code": 0,
  "data": {
    "task_id": "a90e6ab6-44f3-4554-b4***"
  },
  "msg": "ok",
```

```
  "trace_id": "c876cd60b202f2396de1f9e39a1b0172"
}
```

获取任务结果

接口说明

通过 task_id 查询提取任务目前的进度，任务处理完成后，接口会响应对应的提取详情。

Python 请求示例

```python
import requests

token = "官网申请的api token"
url = f"https://mineru.net/api/v4/extract/task/{task_id}"
header = {
    "Content-Type": "application/json",
    "Authorization": f"Bearer {token}"
}

res = requests.get(url, headers=header)
print(res.status_code)
print(res.json())
print(res.json()["data"])
```

CURL 请求示例

```
curl --location --request GET
'https://mineru.net/api/v4/extract/task/{task_id}' \
```

```
--header 'Authorization: Bearer *****' \
--header 'Accept: */*'
```

响应参数说明

参数	类型	示例
code	int	0
msg	string	ok
trace_id	string	c876cd60b202f2396de1f9e39a1b0172
data.task_id	string	abc**
data.data_id	string	abc**
data.state	string	done
data.full_zip_url	string	https://cdn-mineru.openxlab.org.cn/pdf/018e53ad-d4f1-475d-b380-36bf24db9914.zip
data.err_msg	string	文件格式不支持，请上传符合要求的文件类型

参数	类型	示例
data.extract_progress.extracted_pages	int	1
data.extract_progress.start_time	string	2025-01-20 11:43:20
data.extract_progress.total_pages	int	2

响应示例


```
{
  "code": 0,
  "data": {
    "task_id": "47726b6e-46ca-4bb9-******",
    "state": "running",
    "err_msg": "",
    "extract_progress": {
      "extracted_pages": 1,
      "total_pages": 2,
      "start_time": "2025-01-20 11:43:20"
    }
  },
  "msg": "ok",
  "trace_id": "c876cd60b202f2396de1f9e39a1b0172"
}
```

```
{
  "code": 0,
  "data": {
    "task_id": "47726b6e-46ca-4bb9-******",
    "state": "done",
```

```
    "full_zip_url": "https://cdn-mineru.openxlab.org.cn/pdf/018e53ad-
d4f1-475d-b380-36bf24db9914.zip",
    "err_msg": ""
  },
  "msg": "ok",
  "trace_id": "c876cd60b202f2396de1f9e39a1b0172"
}
```


批量文件解析

文件批量上传解析

接口说明

适用于本地文件上传解析的场景，可通过此接口批量申请文件上传链接，上传文件后，系统会自动提交解析任务 注意：

- 申请的文件上传链接有效期为 24 小时，请在有效期内完成文件上传
- 上传文件时，无须设置 Content-Type 请求头
- 文件上传完成后，无须调用提交解析任务接口。系统会自动扫描已上传完成文件自动提交解析任务
- 单次申请链接不能超过 200 个
- header头中需要包含 Authorization 字段，格式为 Bearer + 空格 + Token

Python 请求示例

```python
import requests

token = "官网申请的api token"
```

```python
url = "https://mineru.net/api/v4/file-urls/batch"
header = {
    "Content-Type": "application/json",
    "Authorization": f"Bearer {token}"
}
data = {
    "files": [
        {"name":"demo.pdf", "data_id": "abcd"}
    ],
    "model_version":"vlm"
}
file_path = ["demo.pdf"]
try:
    response = requests.post(url,headers=header,json=data)
    if response.status_code == 200:
        result = response.json()
        print('response success. result:{}'.format(result))
        if result["code"] == 0:
            batch_id = result["data"]["batch_id"]
            urls = result["data"]["file_urls"]
            print('batch_id:{},urls:{}'.format(batch_id, urls))
            for i in range(0, len(urls)):
                with open(file_path[i], 'rb') as f:
                    res_upload = requests.put(urls[i], data=f)
                    if res_upload.status_code == 200:
                        print(f"{urls[i]} upload success")
                    else:
                        print(f"{urls[i]} upload failed")
        else:
            print('apply upload url failed,reason:
{}'.format(result.msg))
    else:
        print('response not success. status:{} ,result:
{}'.format(response.status_code, response))
except Exception as err:
    print(err)
```

CURL 请求示例

```
curl --location --request POST 'https://mineru.net/api/v4/file-
urls/batch' \
--header 'Authorization: Bearer ***' \
--header 'Content-Type: application/json' \
--header 'Accept: */*' \
--data-raw '{
    "files": [
        {"name":"demo.pdf", "data_id": "abcd"}
    ],
    "model_version": "vlm"
}'
```

CURL 文件上传示例

```
curl -X PUT -T /path/to/your/file.pdf 'https://****'
```

请求体参数说明

参数	类型	是否必选	示例	
enable_formula	bool	否	true	是否开启公式识别，默认 tru
enable_table	bool	否	true	是否开启表格识别，默认 tru

参数	类型	是否必选	示例	
language	string	否	ch	指定文档语言，默认 ch，其 https://www.paddleocr.ai/ OCRv5/PP-OCRv5_multi_l pipeline模型有效
file.name	string	是	demo.pdf	文件名，支 持.pdf、.doc、.docx、.ppt、 格式
file.is_ocr	bool	否	true	是否启动 ocr 功能，默认 fa
file.data_id	string	否	abc**	解析对象对应的数据 ID。由 （_）、短划线（-）、英文句 符，可以用于唯一标识您的
file.page_ranges	string	否	1-600	指定页码范围，格式为逗号分 表示选取第2页、第4页至第 [2,4,5,6]）；"2--2"：表示从 中"-2"表示倒数第二页）。
callback	string	否	http://127.0.0.1/callback	解析结果回调通知您的 URL 议的地址。该字段为空时， callback 接口必须支持 POS Type:application/json 传输 content。解析接口按照以下

参数	类型	是否必选	示例	
				content，调用您的 callback checksum：字符串格式，由 字符串，通过 SHA256 算法 查询。为防篡改，您可以在获 成字符串，与 checksum 做 content：JSON 字符串格式 象。关于 content 结果的示 例，对应任务查询结果的 da 说明:您的服务端 callback 接 结果后，如果返回的 HTTP 功，其他的 HTTP 状态码均 mineru 将最多重复推送 5 次 推送 5 次后仍未接收成功， callback 接口的状态。
seed	string	否	abc**	随机字符串，该值用于回调请 数字、下划线（_）组成，不 用于在接收到内容安全的回调 服务发起。 说明:当使用 callback 时，该
extra_formats	[string]	否	["docx","html"]	markdown、json为默认导出 docx、html、latex三种格式
model_version	string	否	vlm	mineru模型版本，两个选项:

请求体示例

```
{
  "files": [{ "name": "demo.pdf", "data_id": "abcd" }],
  "model_version": "vlm"
}
```

响应参数说明

参数	类型	示例	说明
code	int	0	接口状态码，成功： 0
msg	string	ok	接口处理信息，成功："ok"
trace_id	string	c876cd60b202f2396de1f9e39a1b0172	请求 ID
data.batch_id	string	2bb2f0ec-a336-4a0a-b61a-****	批量提取任务id，可用于批量查询解析结果
data.files	[string]	["https://mineru.oss-cn-shanghai.aliyuncs.com/api-upload/***"]	文件上传链接

响应示例

```
{
  "code": 0,
  "data": {
    "batch_id": "2bb2f0ec-a336-4a0a-b61a-241afaf9cc87",
    "file_urls": [
        "https://***"
    ]
  }
  "msg": "ok",
  "trace_id": "c876cd60b202f2396de1f9e39a1b0172"
}
```


url 批量上传解析

接口说明

适用于通过 API 批量创建提取任务的场景 注意：

- 单次申请链接不能超过 200 个
- 文件大小不能超过 200MB,文件页数不超出 600 页
- 因网络限制，github、aws 等国外 URL 会请求超时
- header头中需要包含 Authorization 字段，格式为 Bearer + 空格 + Token

Python 请求示例

```python
import requests

token = "官网申请的api token"
url = "https://mineru.net/api/v4/extract/task/batch"
header = {
    "Content-Type": "application/json",
```

```
    "Authorization": f"Bearer {token}"
}
data = {
    "files": [
        {"url":"https://cdn-mineru.openxlab.org.cn/demo/example.pdf",
"data_id": "abcd"}
    ],
    "model_version": "vlm"
}
try:
    response = requests.post(url,headers=header,json=data)
    if response.status_code == 200:
        result = response.json()
        print('response success. result:{}'.format(result))
        if result["code"] == 0:
            batch_id = result["data"]["batch_id"]
            print('batch_id:{}'.format(batch_id))
        else:
            print('submit task failed,reason:{}'.format(result.msg))
    else:
        print('response not success. status:{} ,result:
{}'.format(response.status_code, response))
except Exception as err:
    print(err)
```

CURL 请求示例

```
curl --location --request POST
'https://mineru.net/api/v4/extract/task/batch' \
--header 'Authorization: Bearer ***' \
--header 'Content-Type: application/json' \
--header 'Accept: */*' \
--data-raw '{
    "files": [
        {"url":"https://cdn-mineru.openxlab.org.cn/demo/example.pdf",
```

```
"data_id": "abcd"}
    ],
    "model_version": "vlm"
}'
```

请求体参数说明

参数	类型	是否必选	示例	
enable_formula	bool	否	true	是否开启公式识别，默认 tru
enable_table	bool	否	true	是否开启表格识别，默认 tru
language	string	否	ch	指定文档语言，默认 ch，其 https://www.paddleocr.ai/ OCRv5/PP-OCRv5_multi_l pipeline模型有效
file.url	string	是	demo.pdf	文件链接，支 持.pdf、.doc、.docx、.ppt、 格式
file.is_ocr	bool	否	true	是否启动 ocr 功能，默认 fa

参数	类型	是否必选	示例	
file.data_id	string	否	abc**	解析对象对应的数据 ID。由（_）、短划线（–）、英文句 符，可以用于唯一标识您的
file.page_ranges	string	否	1–600	指定页码范围，格式为逗号 表示选取第2页、第4页至第 [2,4,5,6])；"2––2"：表示从 中"–2"表示倒数第二页）。
callback	string	否	http://127.0.0.1/callback	解析结果回调通知您的 URL 议的地址。该字段为空时， callback 接口必须支持 POS Type:application/json 传输 content。解析接口按照以下 content，调用您的 callbac checksum：字符串格式，由 字符串，通过 SHA256 算法 查询。为防篡改，您可以在 成字符串，与 checksum 做 content：JSON 字符串格式 象。关于 content 结果的示 例，对应任务查询结果的 da 说明:您的服务端 callback 接 结果后，如果返回的 HTTP 功，其他的 HTTP 状态码均

参数	类型	是否必选	示例	
				mineru 将最多重复推送 5 次 推送 5 次后仍未接收成功， callback 接口的状态。
seed	string	否	abc**	随机字符串，该值用于回调 数字、下划线（_）组成，不 用于在接收到内容安全的回 服务发起。 说明：当使用 callback 时，
extra_formats	[string]	否	["docx","html"]	markdown、json为默认导出 docx、html、latex三种格式
model_version	string	否	vlm	mineru模型版本，两个选项:

请求体示例

```
{
    "files": [
        {"url":"https://cdn-mineru.openxlab.org.cn/demo/example.pdf",
"data_id": "abcd"}
    ],
    "model_version": "vlm"
}
```

响应参数说明

参数	类型	示例	说明
code	int	0	接口状态码，成功：0
msg	string	ok	接口处理信息，成功："ok"
trace_id	string	c876cd60b202f2396de1f9e39a1b0172	请求 ID
data.batch_id	string	2bb2f0ec-a336-4a0a-b61a-****	批量提取任务 id，可用于批量查询解析结果

响应示例

```
{
  "code": 0,
  "data": {
    "batch_id": "2bb2f0ec-a336-4a0a-b61a-241afaf9cc87"
  },
  "msg": "ok",
  "trace_id": "c876cd60b202f2396de1f9e39a1b0172"
}
```

批量获取任务结果

接口说明

通过 batch_id 批量查询提取任务的进度。

Python 请求示例

```python
import requests

token = "官网申请的api token"
url = f"https://mineru.net/api/v4/extract-results/batch/{batch_id}"
header = {
    "Content-Type": "application/json",
    "Authorization": f"Bearer {token}"
}

res = requests.get(url, headers=header)
print(res.status_code)
print(res.json())
print(res.json()["data"])
```

CURL 请求示例

```
curl --location --request GET 'https://mineru.net/api/v4/extract-
results/batch/{batch_id}' \
--header 'Authorization: Bearer *****' \
--header 'Accept: */*'
```

响应参数说明

参数	类型	示例
code	int	0

参数	类型	示例
msg	string	ok
trace_id	string	c876cd60b202f2396de1f9
data.batch_id	string	2bb2f0ec-a336-4a0a-b6 241afaf9cc87
data.extract_result.file_name	string	demo.pdf
data.extract_result.state	string	done
data.extract_result.full_zip_url	string	https://cdn- mineru.openxlab.org.cn/p d4f1-475d-b380-36bf24
data.extract_result.err_msg	string	文件格式不支持，请上传符 类型
data.extract_result.data_id	string	abc**
data.extract_result.extract_progress.extracted_pages	int	1
data.extract_result.extract_progress.start_time	string	2025-01-20 11:43:20

参数	类型	示例
data.extract_result.extract_progress.total_pages	int	2

响应示例


```
{
  "code": 0,
  "data": {
    "batch_id": "2bb2f0ec-a336-4a0a-b61a-241afaf9cc87",
    "extract_result": [
      {
        "file_name": "example.pdf",
        "state": "done",
        "err_msg": "",
        "full_zip_url": "https://cdn-
mineru.openxlab.org.cn/pdf/018e53ad-d4f1-475d-b380-36bf24db9914.zip"
      },
      {
        "file_name":"demo.pdf",
        "state": "running",
        "err_msg": "",
        "extract_progress": {
          "extracted_pages": 1,
          "total_pages": 2,
          "start_time": "2025-01-20 11:43:20"
        }
      }
    ]
  },
  "msg": "ok",
  "trace_id": "c876cd60b202f2396de1f9e39a1b0172"
}
```

常见错误码

错误码	说明	解决建议
A0202	Token 错误	检查 Token 是否正确，请检查是否有Bearer前缀 或者更换新 Token
A0211	Token 过期	更换新 Token
-500	传参错误	请确保参数类型及Content-Type正确
-10001	服务异常	请稍后再试
-10002	请求参数错误	检查请求参数格式
-60001	生成上传 URL 失败，请稍后再试	请稍后再试
-60002	获取匹配的文件格式失败	检测文件类型失败，请求的文件名及链接中带有正确的后缀名，且文件为 pdf,doc,docx,ppt,pptx,png,jp(e)g 中的一种
-60003	文件读取失败	请检查文件是否损坏并重新上传
-60004	空文件	请上传有效文件
-60005	文件大小超出限制	检查文件大小，最大支持 200MB

错误码	说明	解决建议
-60006	文件页数超过限制	请拆分文件后重试
-60007	模型服务暂时不可用	请稍后重试或联系技术支持
-60008	文件读取超时	检查 URL 可访问
-60009	任务提交队列已满	请稍后再试
-60010	解析失败	请稍后再试
-60011	获取有效文件失败	请确保文件已上传
-60012	找不到任务	请确保task_id有效且未删除
-60013	没有权限访问该任务	只能访问自己提交的任务
-60014	删除运行中的任务	运行中的任务暂不支持删除
-60015	文件转换失败	可以手动转为pdf再上传
-60016	文件转换失败	文件转换为指定格式失败，可以尝试其他格式导出或重试