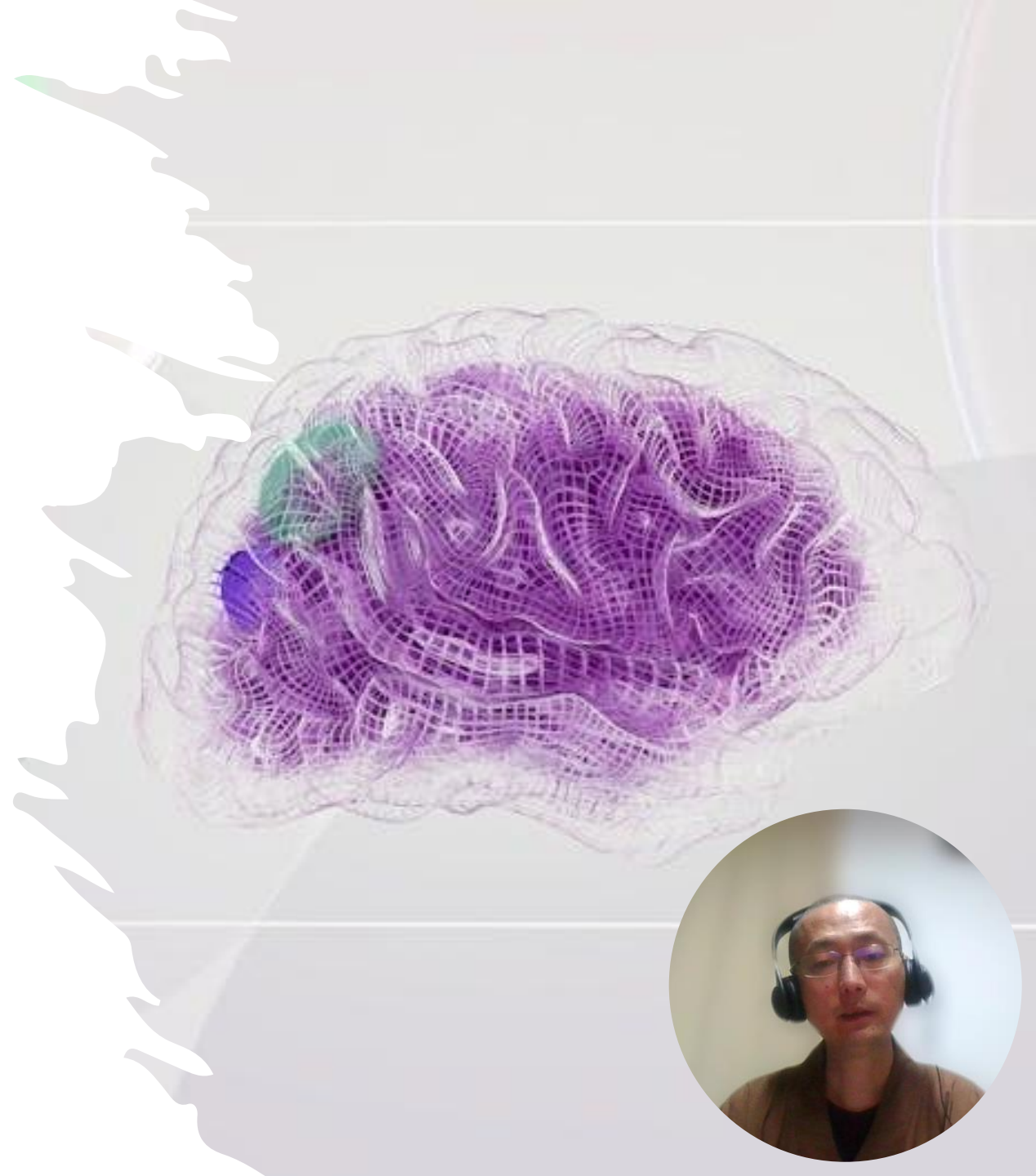


後數位化時代的 OCR算法與應用創新

釋賢超 北京市海淀區龍泉寺藏經辦公室

<https://www.gj.cool/author>

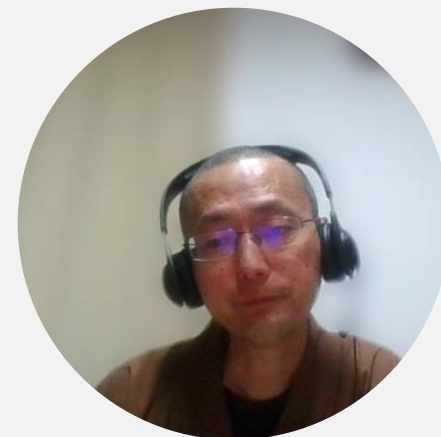


内容大纲

OCR：发展前景和融合框架

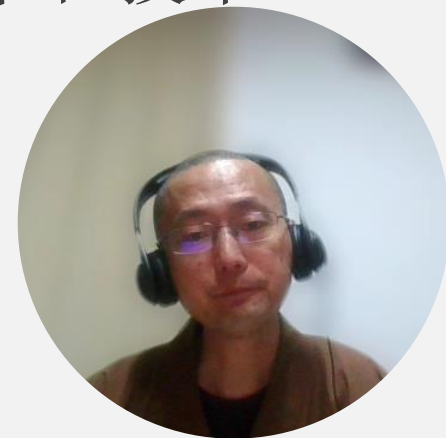
古籍酷：应用平台

籍智达：事业拓展



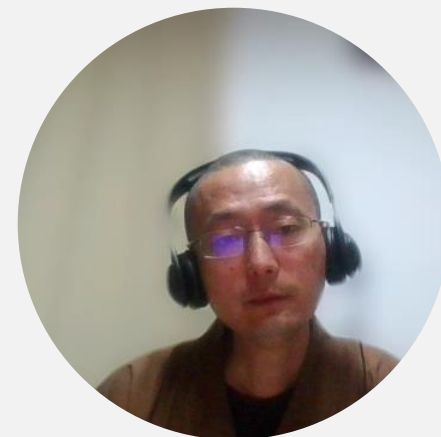
OCR与大模型

- 多模态大模型可以实现**OCR**的大部分能力
- **OCR**也能够辅助大模型获得更好表现
- **OCR**相对大模型的优势——性价比；劣势——通用性
- 中文古籍的特殊性：数据偏少，使用面窄，难以吸引资本和技术

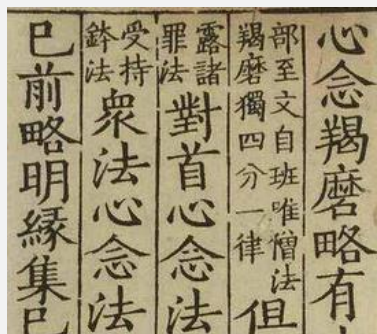
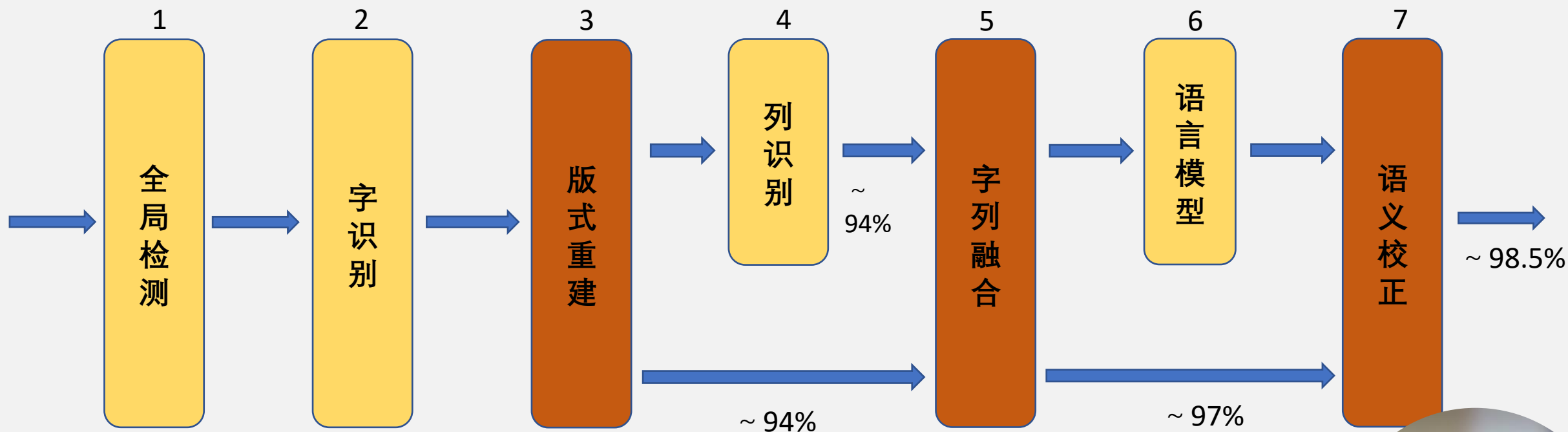


中文古籍OCR框架：多模型融合

- 版式特征：双行夹注
- 开发时间：2021年5月-12月
- 知识准备：2020年字符串比对
- 算法创新：针对古籍版式的字符级版式重建



Multi-Model Ensemble Framework



心念羯磨略有十四【義分三別一但心念法二對首】【心念法
【部至文自班唯僧法】【羯磨獨四分一律】但心念法有三
【露諸】【罪法】對首心念法有七【安居法說淨法受藥法受
【受持】【鉢法】衆法心念法有四【說戒法自恣法受僧得】
已前略明緣集已後辨緣成壞



古籍酷：应用平台

古籍酷 AI服务

中文古籍OCR、自动标点、文言文翻译

The Coolest AI Service for Chinese Ancient Books

OCR, Punctuation, Translation

古籍OCR 𠄎
OCR for Ancient Books

自动标点 𠄎
Auto Punctuation

文言文翻译 𠄎
Classical Chinese Translation

快速体验 𠄎
Fast Experience

快速体验 Fast Experience

准备就绪，请上传/拍照



問梅詩社詩鈔卷一
後附東籬會詩
癸未
探梅結社紀事
九興詩

我過北郭里每懷沈尚書詩社問遺老
僉日久荒蕪中吳盛文敷大雅輪誰扶
黃子豪于詩夙以山水娛邀我出西郊
尋詩訪林逋歸田有彭子欣然命籃輿
春風來吹送一路三人俱行行詣古寺
冷香襲襟裾巡簷佇清影啜茗菰蔬
再游小雲栖

上传 拍照

标点

翻译

清理

快速体验 Fast Experience

庵記云

暇日與文文起

朱德升過此

今亦三人來後先若合符黃子大歡喜



自动标点

問梅詩社詩鈔卷一後附。東籬會詩。癸未探梅結社紀事。九興詩。我過北郭里，每懷沈尚書。詩社問遺老，僉日久荒蕪。中吳盛文敷，大雅輪誰扶。黃子豪于詩，夙以山水娛。邀我出西郊，尋詩訪林逋。歸田有彭子，欣然命籃輿。春風來吹送，一路三人俱。行行詣古寺，冷香襲襟裾。巡簷佇清影，啜茗菰蔬。再游小雲栖周忠介公書額。不辭幽徑紆，仰瞻忠介書。凜然生眉鬚，相傳過此地。公嘗偕文朱公作是庵記云：暇日與文文起、朱德升過此，今亦三人來。後先若合符，黃子大歡喜。



上传 拍照

标点

翻译

清理

快速体验 Fast Experience

文言文翻译

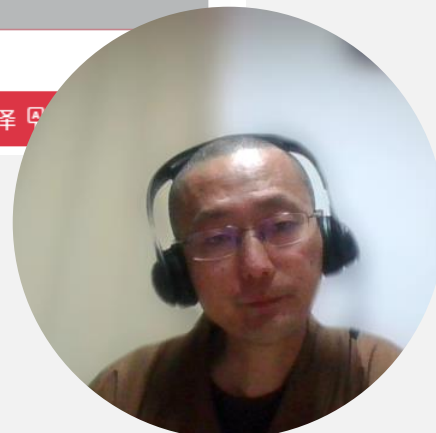
请他读《梅诗》和《诗经》，然后才将之附在后面。东门诗是这样。二十六日，探索梅子结社纪事。九兴诗。我路过北郭外城，常常思念着沉于尚书。诗社问遣老，都说长期荒芜。中吴盛有文才，大雅有谁扶持？黄子豪富于诗，早以山水为娱乐。邀我去西郊，寻访林逋。归田前有彭子，高兴心地坐在车上。春风吹来送走了我，一路上三人都来了。他走到古寺，脱下了清香，穿上了衣襟。我只有等待清晨的影子，吃茶叶也吃蔬菜。第二次游历小云，是周朝忠诚耿介的公文。不辭幽隱曲折，仰瞻忠节书籍。严生凜然眉毛长满胡须，一路前来经过这里。不辭幽隱曲折，仰瞻忠介书。谢安曾经和文朱公一起作这样的记载：闲暇的时候与文起、朱德升一起经过这里，如今也有三个人来。不辭幽隱曲徑，仰瞻忠介书，皇后若合符，黄子大喜。



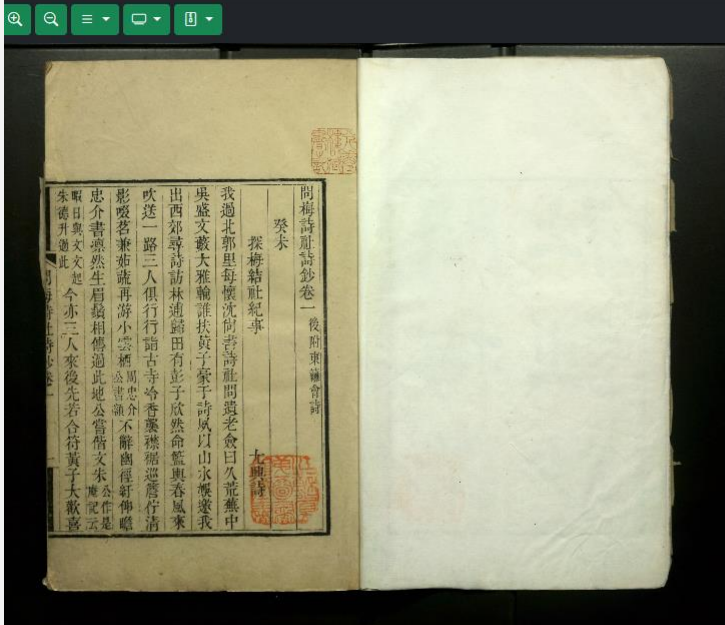
上传 拍照

标点

翻译



1、OCR_Pro(自动识别) API



A001.jpg

1 / 1

URL <70MB

账户 1331 / 10000

▼ 字段

字段	内容	说明
Authorization *	gjcool eyJ0eXAiOiJKV1QiLCJhbGciOiJIUzI1	在Header中设置。
img *	选择文件 A001.jpg	MIME: image/jpeg, image/png, image/tiff, image/webp, image/heic, image/heif, image/jp2.
layout	不设置	0: 竖版, 1: 横版。默认值: 0。
area	[]	默认值: '[]'. 尝试通过鼠标绘制。
compact	不设置	合法取值: 1, 2, 4, 6. 数值大于1时, 可以改善密集文字的识别效果, 但会消耗等值的使用次数。默认值: 1。

1. *: 必设字段。

2. 非必设字段的“不设置”与使用默认值的效果相同。

3. <access_token>的过期时间为24小时。同一账户下的所有<access_token>共享使用次数, 每天的使用次数有上限。

▼ 结果

Json ☒ Text

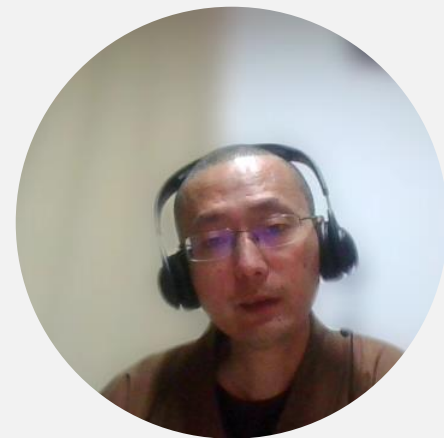
```
{  "FileName": "A001",  "ContentType": "image/jpeg",  "CharNumber": 183,  "LineNumber": 19,  "Width": 5769,  "Height": 4658,  "Size": 3601695,  "Layout": 0}
```

▼ 代码

☒ cURL ☐ JS ☐ Python

```
curl -X POST https://api.jzd.cool:9013/ocr_pro -H "Authorization: gjcool eyJ0eXAiOiJKV1QiLCJhbGciOiJIUzI1NiJ9.eyJmcmVzaCI6ZmFsc2UsImhhdCI6MTcxMzE0NDk1MSwianRlpljoiZjcwOGM2YmEtZDY5Yy00YTk0LWw1ZjAtYyYyNWw1M2U3MmZkliwidHlwZSI6ImFjY2VzcyIsbnN1YiIiLCJPSm9NcHBtVIZCcEEva1I1ZUNBRWc9PVxuliwibmJmljoxNzEzMlQ0OTUxLCJleHAiOiE3MTMyMzEzNTF9.TS07Xj4u5cwtfAqY-3D34zSC1tcCoKQNJ4T5jRwAM1TIC3ZFQa38KfyCrrPHRN5Tfu65zQ9DLYlBHK8zno69gMawKBBt_oMKD0ZLcmxAJkL5mfZdc5bZ9zwlstW6Fe_mxdRhgwVJ2xl0qc1Dr85ZgjAXhJwByZRpTYodPHzA1k" -F "img=@<本地文件路径>"
```

1. 关于cURL。<本地文件路径>如果包含中文字符, 需要先进行urlencode编码。
2. 关于Javascript。API服务端支持CORS(跨域资源共享), 即允许用户通过自己域名下的网页发起调用请求。



详尽的文档说明

Body:			
字段	是否必填	格式	说明
apiid	是	form-data	使用授权服务的唯一用户标识。可从 账户页面 获取。
password	是	form-data	密码
encrypt	否	form-data	0: password字段使用明文格式; 1: password字段使用加密格式。 加密步骤: (1)密码明文用UTF-8编码; (2)利用公钥加密; (3)加密后的密文再用base64编码。 若不作特意设定, 则取默认值0。使用密码加密功能, 需申请公钥。
is_long	否	form-data	是否生成长效access_token。 0: 否。过期时间为24小时。 1: 是。过期时间为90天。 若不作特意设定, 则取默认值0。
正常返回:			
application/json			
字段	含义	格式	
access_token	用于授权服务的鉴权。	JSON Web Token ↗	
refresh_token	用于刷新access token, 过期时间为30天。	JSON Web Token ↗	
msg	错误类型	apiid或password校验错误, 返回值为 "Bad request" 非正式授权用户或用户授权过期, 返回值为 "User blocked" encrypt设置错误, 返回值为 "encrypt wrong" 密文解密失败, 返回值为 "Decrypt failed" 其他密码相关错误, 返回值为 "password wrong"	



1. 认证类API

- 1.1 授权
- 1.2 刷新access token

2. 图片类API

- 2.1 自动识别
- 2.2 超分辨率
- 2.3 双层PDF

3. 文本类API

- 3.1 自动标点
- 3.2 文白翻译

2.1 自动识别

URL: https://api.jzd.cool:8000/ocr_pro

Method: POST

Headers:

字段	内容	说明
Authorization	gjcool access_token	"gjcool"与access_token之间用一个空格连接

Body:

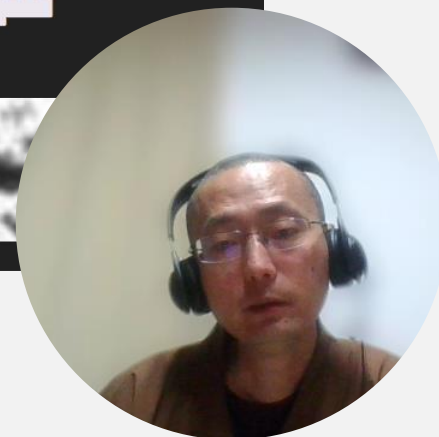
字段	是否必填	格式	说明
img	是	form-data	图片, file, MIME: image/jpeg, image/png, image/tiff, image/webp, image/heic, image/heif, image/jp2, 体积不得超过70MB。
layout	否	form-data	版式, 整数, 取值范围: 0, 1。 0: 竖版, 文字方向从上至下, 行排列方向从右至左, 适用于绝大多数古籍图像; 1: 横版, 文字方向从左至右, 行排列方向从上至下, 适用于现代出版物。 若不作特意设定, 则取默认值0。
area	否	form-data	区域, 字符串, 格式: "[a1,b1,c1,d1],[a2,b2,c2,d2]", a1, b1, c1, d1分别代表单个区域的左边界、上边界、右边界、下边界的像素值, 都是整数, 用","间隔, 外部用"[]"包围, 各个区域用","间隔, 最外层用"[]"包围。 若不作特意设定或者设置为"[]", 则默认针对全图进行识别, 适合于大多数简单版式的情况。 针对复杂版式, 可以通过设置此字段, 针对图片中的若干区域依次识别。
compact	否	form-data	密集系数, 整数, 取值范围: 1, 2, 4, 6。 数值大于1时, 可以改善密集文字的识别效果, 但会消耗等值的使用次数。 若不作特意设定, 则取默认值1。

正常返回:		
application/json		
字段	含义	格式
FileName	图片的文件名 (不含扩展名)	字符串。
ContentType	图片的MIME	字符串。
CharNumber	总字数	整数。
LineNumber	总行数	整数。
Width	图片宽度的像素值	整数。
Height	图片高度的像素值	整数。
Size	图片的存储空, 单位byte	整数。
Layout	识别版式	整数。
Area	识别区域	字符串, 含义同
Compact	识别密集系数	整数。
chars	单字识别结果	列表, 数据项为长度为1的字符串, 数据项个数与总字数一致。
coors	单字坐标	列表, 数据项为[x1,y1,x2,y2], 分别代表字框的左边界、上边界、右边界、下边界的像素值, 都是整数, 数据项个数与总字数一致。
char_probs	单字识别结果的置信度	列表, 数据项为四位小数, 数据项个数与总字数一致。
coor_probs	单字坐标的置信度	列表, 数据项为四位小数, 数据项个数与总字数一致。
char_ids	单字的序号	列表, 数据项为整数, 数据项个数与总字数一致。

异常返回:		
application/json		
字段	含义	格式
msg	错误类型	1. 设置错误 "image size too large": img体积超过限定值。 "image format wrong": img格式不支持。 "img wrong": img错误。 "img read failed": img读取错误。 "heic/heif failed": img heic/heif读取错误。 "layout wrong": layout错误。 "area wrong": area错误。 "image height or width wrong": img的高度或宽度错误。 "compact wrong": compact错误。 2. 执行错误 "OCR failed": OCR失败。 3. 认证错误 "user blocked": 未获得正式授权。 "token wrong": access_token错误。 "reach usage limit": 使用量达到上限或者授权过期。 "request too frequent": 请求过于频繁。 "reach traffic limit wait *** hours": 达到流量上限, 等待***小时后重试。 以上为常见错误类型, 其他错误类型恕不赘述。



无限缩放，毫厘可见



辅助编辑： 多多益善

復

拼音

异体字

形近字

语义替换

向前补字

向后补字

易错字

1. 覆

2. 復

3. 復

4. 復

5. 復

6. 復

7. 復

8. 復

9. 復

10. 復

11. 復

12. 復

13. 復

14. 復

復

拼音

异体字

形近字

语义替换

向前补字

向后补字

易错字

1. 復

2. 復

3. 復

4. 復

5. 履

6. 履

7. 獲

8. 觀

9. 役

10. 預

復

拼音

异体字

形近字

语义替换

向前补字

向后补字

易错字

1. 佛

2. 復

3. 復

4. 復

5. 心

6. 经

7. 藏

8. 經

9. 王

10. 藏

復

拼音

异体字

形近字

语义替换

向前补字

向后补字

易错字

1. 溫

2. 履

3. 履

4. 頌

5. 作

康熙字典

CTEXT

Unihan

GlyphWiki

CHISE

HNG

MDBG

WWWJDIC

HanDeDict

CHDICT

復: 4(1)

金剛頂一切如來真實攝大現證大教王經卷第二

七

羅

尊寶生如來昇茶羅左邊月輪而住說此偈

施南

奇哉我無比稱為寶供養

於三界王勝教勅受供養

盧遮那復入一切如來歌詠供養三昧耶所

生名金剛三摩地一切如來族大天女從自

心出

踰囉

二

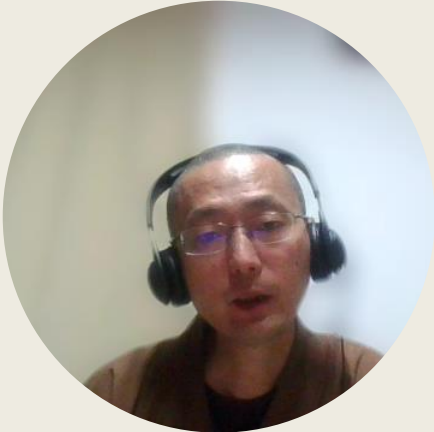
合

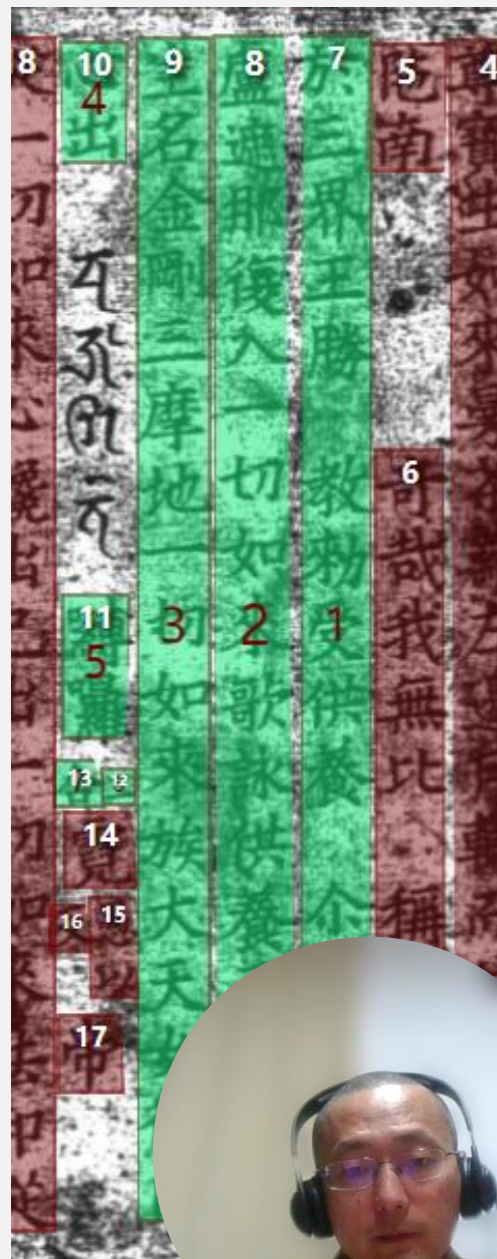
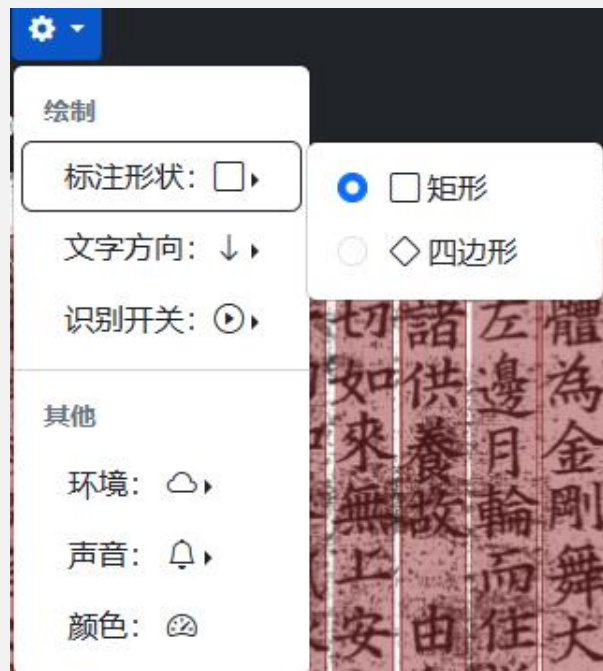
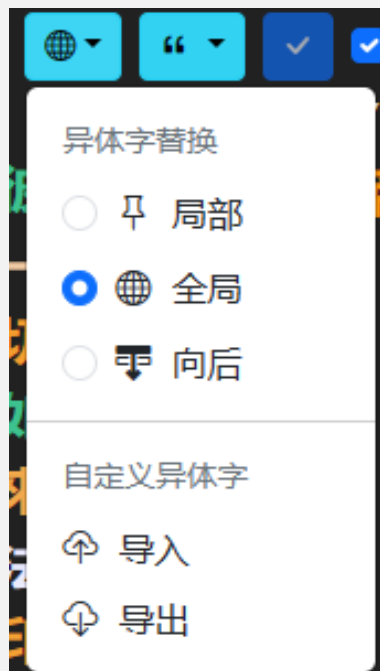
覽

愚以

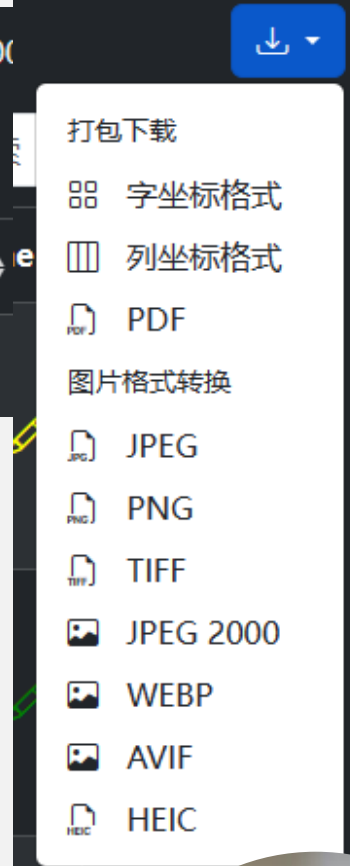
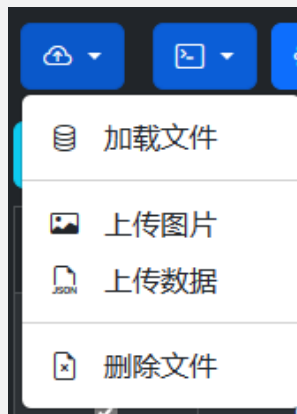
反

行數: 48; 字數: 479

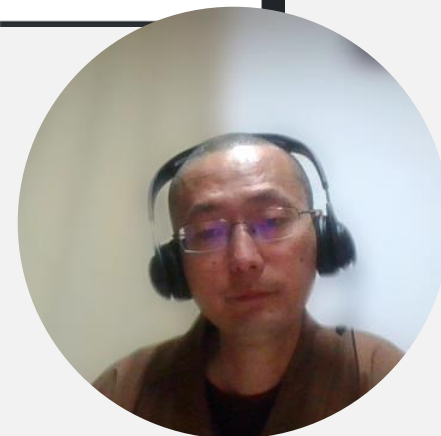




批量处理

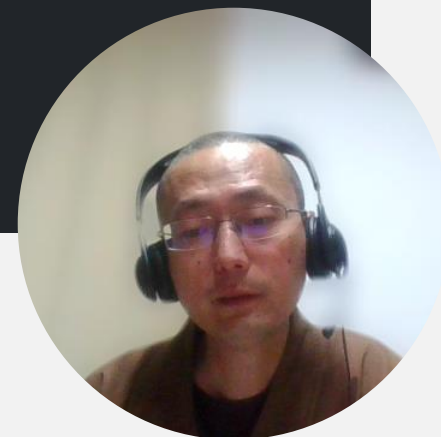


ocr: 377/100000. pdf: 106/100000.											
搜索											
	文件名	体积	上传日期	图片预览	上传进度	文本	字数	OCR	Line	PDF	
☒	0430_0.jpg	1.84 MB	2024-04-05 15:05:59		已上传	金剛頂一切如來真實攝大乘現證大教王經卷第二 七 羅 尊寶生如來鼻茶羅左邊月輪而住說此嘸 陀南 奇哉我無比稱為寶供養 於三界王勝教勅受供養爾時世尊毗 盧遮那復入...	479				
☐	0440_1.jpg	1.91 MB	2024-04-05 15:06:10		已上传	金剛頂一切如來真實攝大乘現證大教王經卷第三 十四 羅 係薩普 二 合 吒鐵大得 健吒嚧嚧令震動 我今說法印成就令清淨於舌觀金剛 能作諸事業次說羯磨印應結金剛拳...	415				
☐	0446_0.jpg	1.72 MB	2024-04-05 15:06:15		已上传	金剛頂經瑜伽修習毗盧遮那三摩地法七 羅 即證薩婆若定中遍禮佛願加持堅固 一切諸佛聞金剛界言已盡入金剛中便說金剛心嚧 一 薩婆他他藥多 引 二 鼻三 菩提三涅槃茶...	413				
☐	0451_0.jpg	1.80 MB	2024-04-05 15:06:20		已上传	金剛恐怖集會方廣軌儀觀自在菩薩三出最勝心明王經 五 羅 種願所謂持明仙無相三昧所現摩訶底 惹等若得持明仙願為明仙中輪惹身相美 白髮紺青色便成二八童子之形瞬目...	448				
金剛恐怖集會方廣軌儀觀自在菩薩三出											



隐私政策

- 1、本网站**不搜集、不存储用户数据**（包括但不限于：上传图片、OCR结果以及用户编辑内容，上传文本与自动标点结果），也不对用户数据进行“**合理利用**”（例如改进模型和相关算法等），仅对用户最近一次的上传图片和OCR结果进行临时性存储。
- 2、本网站不存储用户密码明文，并对用户密码进行了高强度加密，理论上可以保证除用户本人之外的任何人无法在合理时间内还原密码明文。
- 3、本网站不对用户使用OCR、自动标点所产生的数据主张权利或授予权利，不对用户合理使用数据的行为附加任何条件或限制，但同时也不承担与之相关的任何直接或间接的法律责任和道德义务。
- 4、出于维护网站正常运行的需要，除了用户注册时提供的信息之外，本着“**最低程度**”、“**不触及用户内容**”的原则，用户以下的使用行为和数据将被酌情记录：
 - (1) 提交ocr或自动标点单次请求的时间、IP地址；
 - (2) 上传图片的体积；
 - (3) 自动标点原文的字数；
 - (4) **持续长时间API请求所提交的图片或文本。**



更新历史

版本：RC 5.24

一般更新：

- 1、API演示。优化jp2文件加载，增加下载进度提示。

— 2024.3.29

版本：RC 5.23

一般更新：

- 1、首页快速体验。增加上传进度提示。
- 2、API演示。增加演示账号的当前用量的实时反馈。

— 2024.3.28

版本：RC 5.22

重要更新：

- 1、双层PDF API参数调整。compression取值范围调整为0-5的整数，数值越大则图像压缩程度越大，生成的PDF文件体积越小。
- 2、标注平台。优化图片上传按钮，新增文件信息展示。批量处理界面的图片上传、数据上传分为两个独立按钮，增加当前使用量的实时反馈。

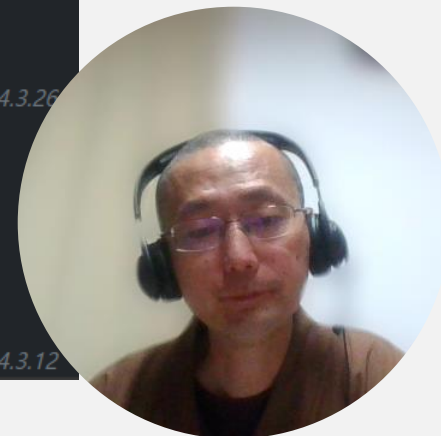
— 2024.3.26

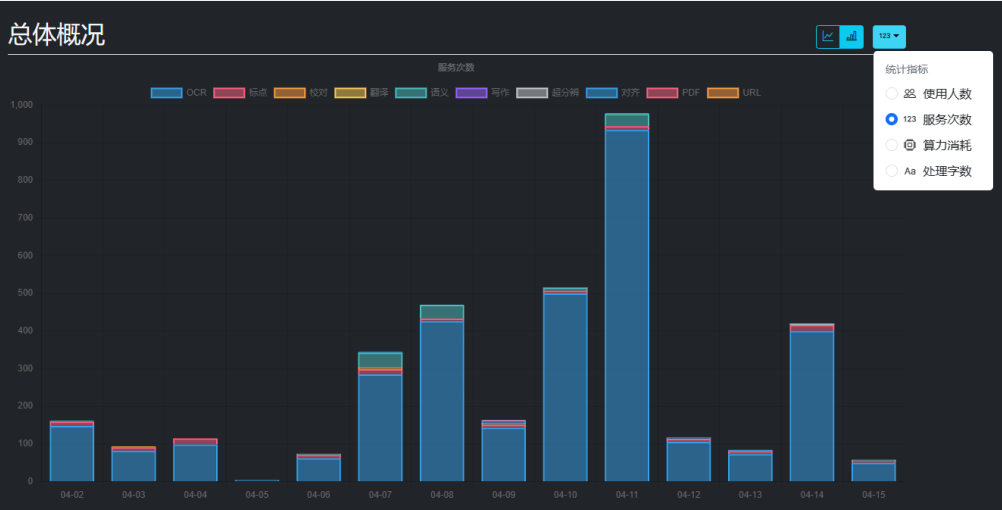
版本：RC 5.21

一般更新：

- 1、导航栏调整。

— 2024.3.12

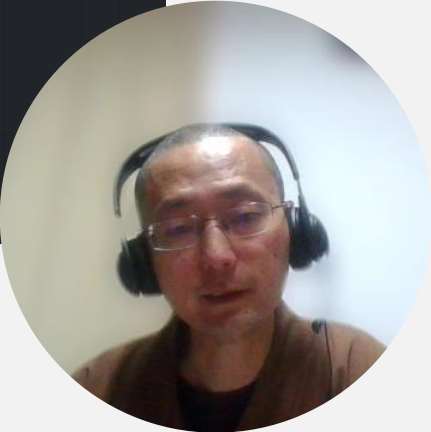
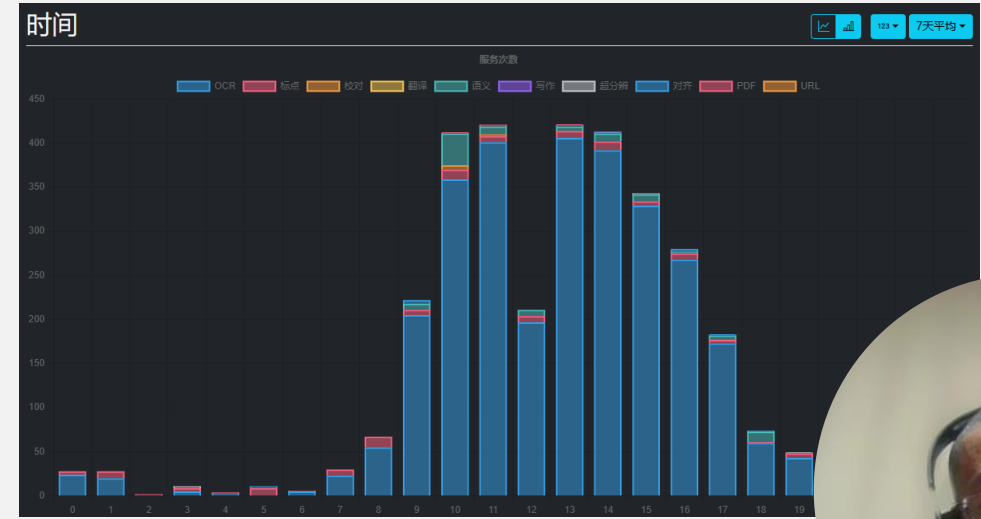
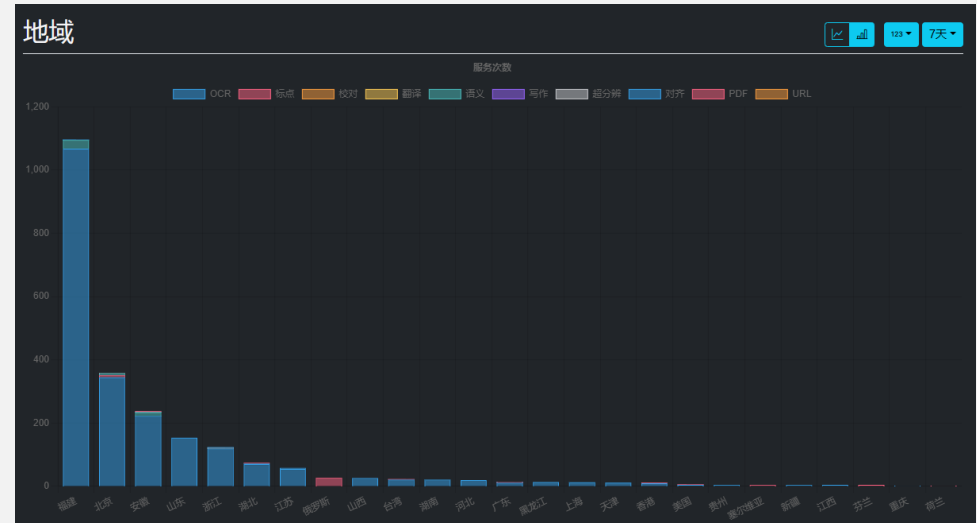




服务类型

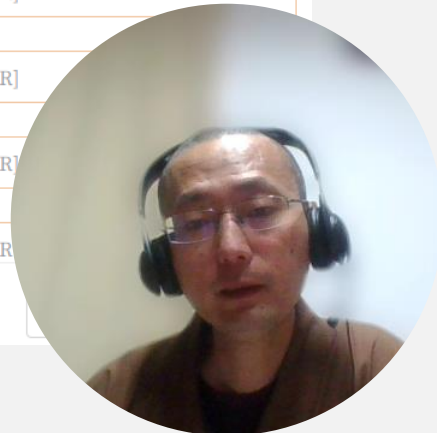
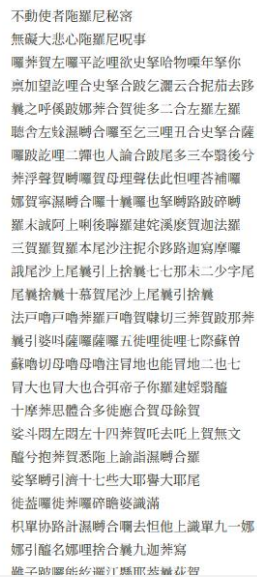
服务类型: OCR

日期	使用人数	服务次数	算力消耗	处理字数
2024-04-15	11	50	88	12325
2024-04-14	35	400	568	93069
2024-04-13	20	73	136	28733
2024-04-12	18	106	250	46711
2024-04-11	54	934	1389	81346
2024-04-10	26	500	794	33158
2024-04-09	31	143	374	44263
2024-04-08	36	426	562	74659
2024-04-07	47	285	584	93242
2024-04-06	15	62	106	24020
2024-04-05	2	4	7	2159
2024-04-04	20	99	177	38991
2024-04-03	12	82	209	32214
2024-04-02	24	148	252	42209



文字识别校对帮助文档

1. 调整切分：保证每个应被识别的字，都有字框；并且字框归属于正确的列框。
2. 调整顺序：文字间的阅读顺序正确。
3. 校对文字：每个字都识别为正确的写法。



中研院文字識別與校對平台

文字校正模式

校正選擇：○區域 ●文字

顯示：原圖 文字圖 提示顏色(2) 單行顯示 深色背景 顯示符號

檢視方式：文字區在左 字型：新經明體 快捷鍵說明

檢視大小：文字區 圖片區 固定檢視比例

笑	笑	不	不	援	援	非	非	難	難	十	十	不	不	尚	尚	過	過
即	即	為	為	人	人	僂	僂	而	而	萬	萬	如	如	寶	寶	日	日
背	背	之	之	之	之	倖	倖	責	責	金	金	揭	揭	聞	聞	塞	塞
後	後	三	三	功	功	水	水	經	經	不	不	數	數	其	其	外	外
有	有	思	思	以	以	泉	泉	略	略	費	費	十	十	欲	欲	繼	繼
他	他	哉	哉	生	生	之	之	之	之	則	則	萬	萬	以	以	夷	夷
如	如	今	今	意	意	二	二	所	所	門	門	金	金	空	空	本	本
公	公	前	前	氣	氣	捷	捷	易	易	下	下	略	略	口	口	非	非
所	所	後	後	而	而	也	也	也	也	又	又	之	之	犁	犁	孝	孝
聞	聞	虞	虞	即	即	權	權	今	今	以	以	綴	綴	歸	歸	子	子
堂	堂	書	書	反	反	酋	酋	之	之	求	求	一	一	樽	樽	順	順
堂	堂	尚	尚	其	其	之	之	論	論	廣	廣	步	步	酋	酋	孫	孫
天	天	在	在	事	事	均	均	者	者	辱	辱	可	可	相	相	水	水
朗	朗	稱	稱	以	以	革	革	▽	▽	國	國	以	以	與	與	清	清

西園聞見錄

卷之八十三

見其易得非僥倖水泉之二捷也權酋之肯革面受約束取
誰為之夫援人之功以生意氣而即反其事以敗之古云
者至矣可不為之三思哉今前後虞書尚在稱獻稱納者
為中國之笑即背後有他如公所聞堂堂天朝之體斷無
責大羊異類者也彼以詐來我亦以詐應之防之如防虎
則邊境安矣當王安石賈似道之開邊其言未必不以貼
之笑者一時痛哭乃在萬世以公純忠憂國猛氣吞胡固
此然好謀而成先師炯戒美成在久莊子至言不肖弒義
孫千萬之一否即今火真雖逸而西邊信當為之事甚多

西園聞見錄

- 1 輸入缺字構字式
- 2 輸入校正備註
- 3 插入文字
- 4 插入符號
- 5 區間標記符號開始
- 6 區間標記符號結束
- 7 標記為錯別字
- 8 刪除文字

插入單個文字在前
插入單個文字在後
編輯整行文字

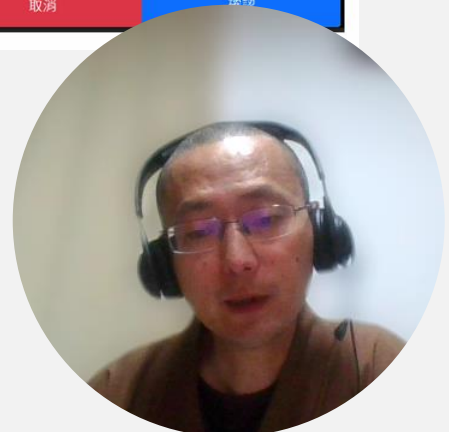
編輯整行文字

編輯整行文字

檢視大小：[滑桿]

西園聞見錄

取消 確定



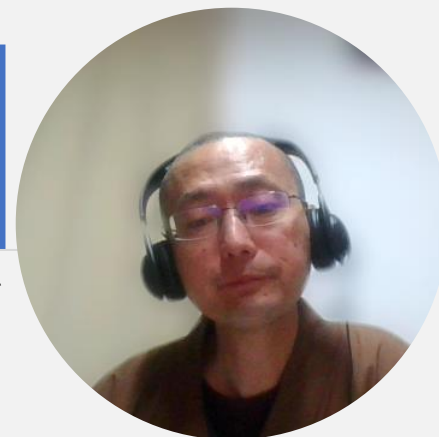
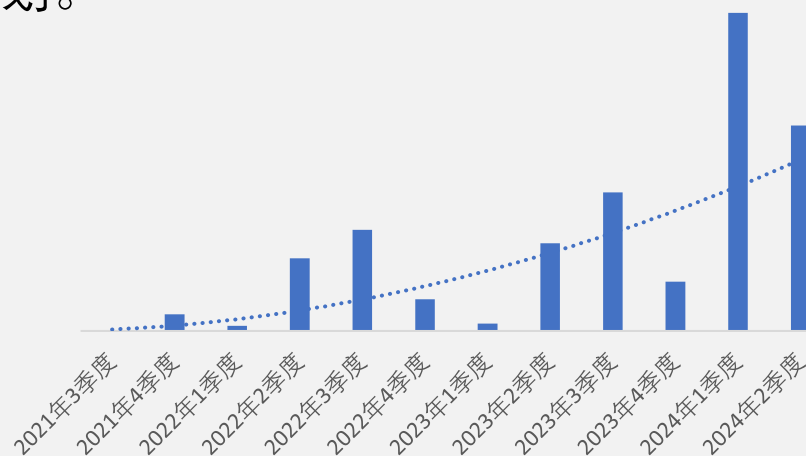
籍智达：事业拓展

2021年，注册成立。

2022年，API网络服务。校对平台网络免费体验。

2023年，服务器租用。发布“数字万舟”公益计划。

2024年，服务器出售。校对平台本地化部署。



合作单位

复旦大学中国语言文学系

浙江大学数字人文研究中心

山东大学数字人文实验室

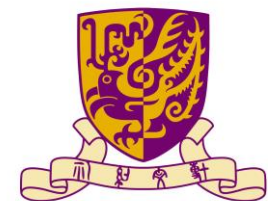
武汉大学数字人文研究中心

对外经济贸易大学中国语言文学学院

江苏师范大学中华家文化研究院

商丘师范学院汉梁文化研究中心

香港中文大学图书馆



香港中文大學
The Chinese University of Hong Kong



“数字万舟”计划简介

中文古籍OCR工业版自2022年发布以来广受各界好评，得到了一批忠实用户的鼎力支持。他们的积极反馈和殷切期望，为我们注入了源源不断的前进动力。正所谓“士不可不弘毅，任重而道远。”一方面是尘封的古籍在手中日渐老去，另一方面是数字列车从身边飞驰而过。愿望与现实间的巨大落差，无时无刻不在触动每一位中华学人心中的软处。

有感于此，古籍酷正式对外发布“数字万舟”计划，面向社会征集优秀中文古籍数字化项目，为传承中华古籍文化贡献绵薄之力。符合申请条件的项目，将获赠中文古籍OCR工业版免费使用额度。

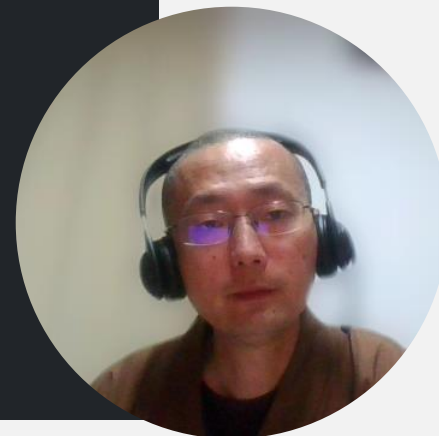
申请者

申请者应为以下三者之一：

- 1、学术或教育单位；
- 2、社会非营利机构；
- 3、个人。

申请条件

- 1、健康。符合申请者所在国家或地区以及服务对象所在国家或地区的法律和社会伦理，具有正面社会价值。
- 2、公益。不含任何盈利性成分。
- 3、透明。在网络主页及时公布项目的进展情况和成果信息，准确描述古籍酷的技术贡献，在描述中应包含“古籍酷”和“gj.cool”的明确字样。
- 4、授权古籍酷及其关联单位对该项目的公开内容进行转载或展示。



数字万舟的受赠者

• 单位：

- 复旦大学中国语言文学系
- 对外经济贸易大学中国语言文学学院
- 江苏师范大学中华家文化研究院
- 商丘师范学院汉梁文化研究中心
- 铜仁市图书馆

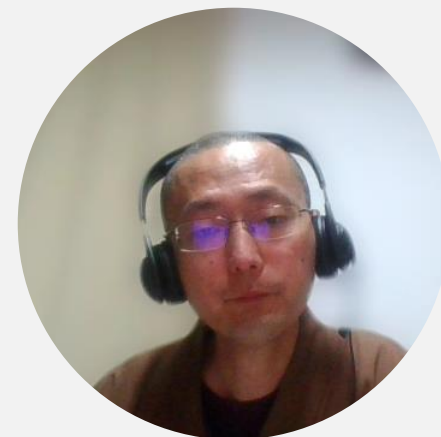
个人：

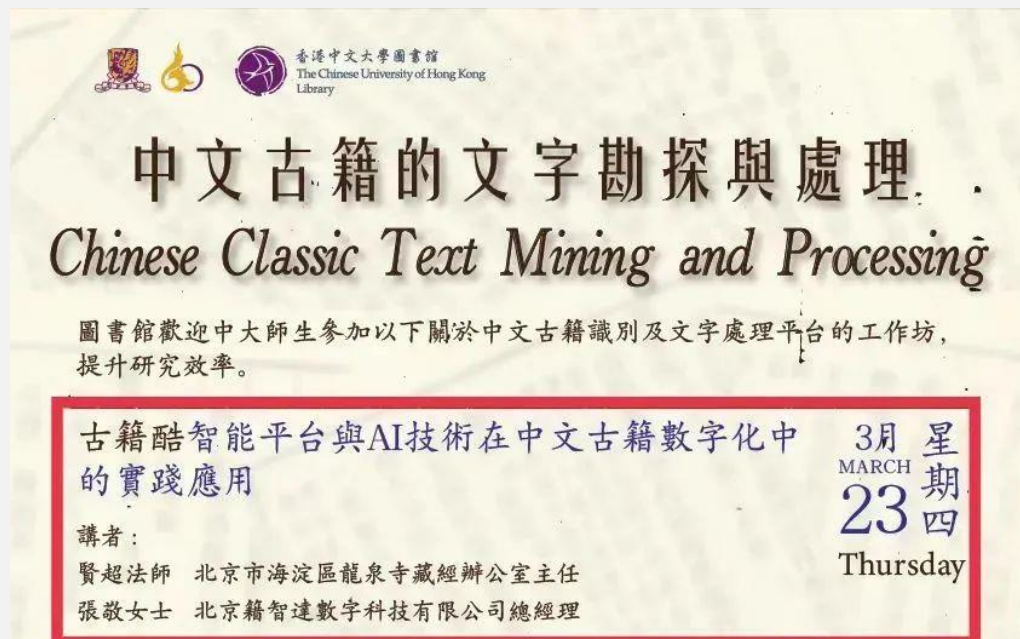
- 哈佛大学、埃默里大学、俄亥俄州立大学、匹兹堡大学、马来西亚理科大学，……
- 北京大学、复旦大学、浙江大学、北京师范大学、中央美术学院、华东师范大学、华中师范大学、华中科技大学、西北大学、湖北大学、广东外语外贸大学、河南财经政法大学、南京理工大学、厦门大学、香港中文大学、澳门理工大学，……



项目类型

- 说文，集韵，训诂小学
- 左传，清史稿，清宫档案
- 经籍志，省郡县府镇志，山水志，运河文献
- 文集，小说
- 类书
- 中医方剂
- 家谱，族谱，家训
- 科举文献
- 民间文献
- 石经、石刻
- 政区史、海洋史、农村史
- 专名提取、历史计量学
- 语料库
- 维基文库、中国哲学电子书计划

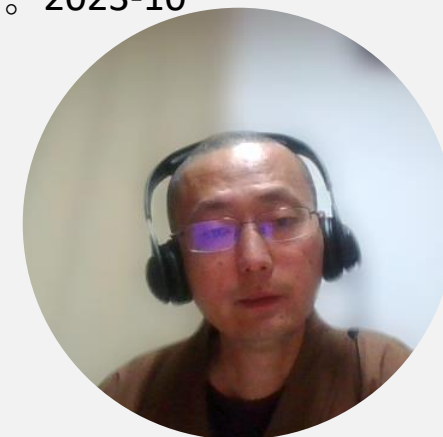




香港中文大学图书馆中文古籍的文字勘探与处理工作坊。2023-3



杭州：第二届东亚古籍数字人文研讨会。2023-10



籍智达与嘉兴精严讲寺签署战略合作协议。本人受聘嘉兴市嘉兴藏文化研究会顾问。2023-11





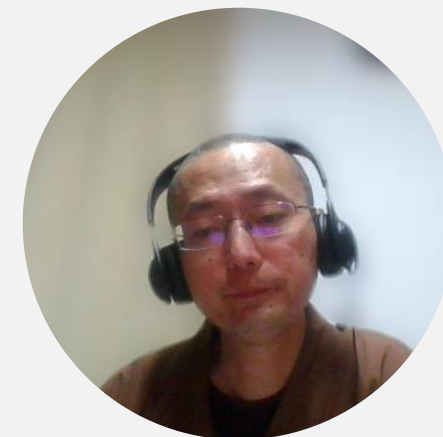
黄河科技学院：中华优秀传统文化与黄河文化学术论坛。2023-11



北京：第二十届全国科技翻译研讨会。2023-11



曲阜：全国家文化研究机构联席会议二次会议暨颜氏家训家风与中华民族家文化研讨会。2023-11



展望

