

抖音集团“识典古籍”数字化平台

- ◆ 上线时间：2022年10月
- ◆ 参选类别：年度最佳AI应用工具

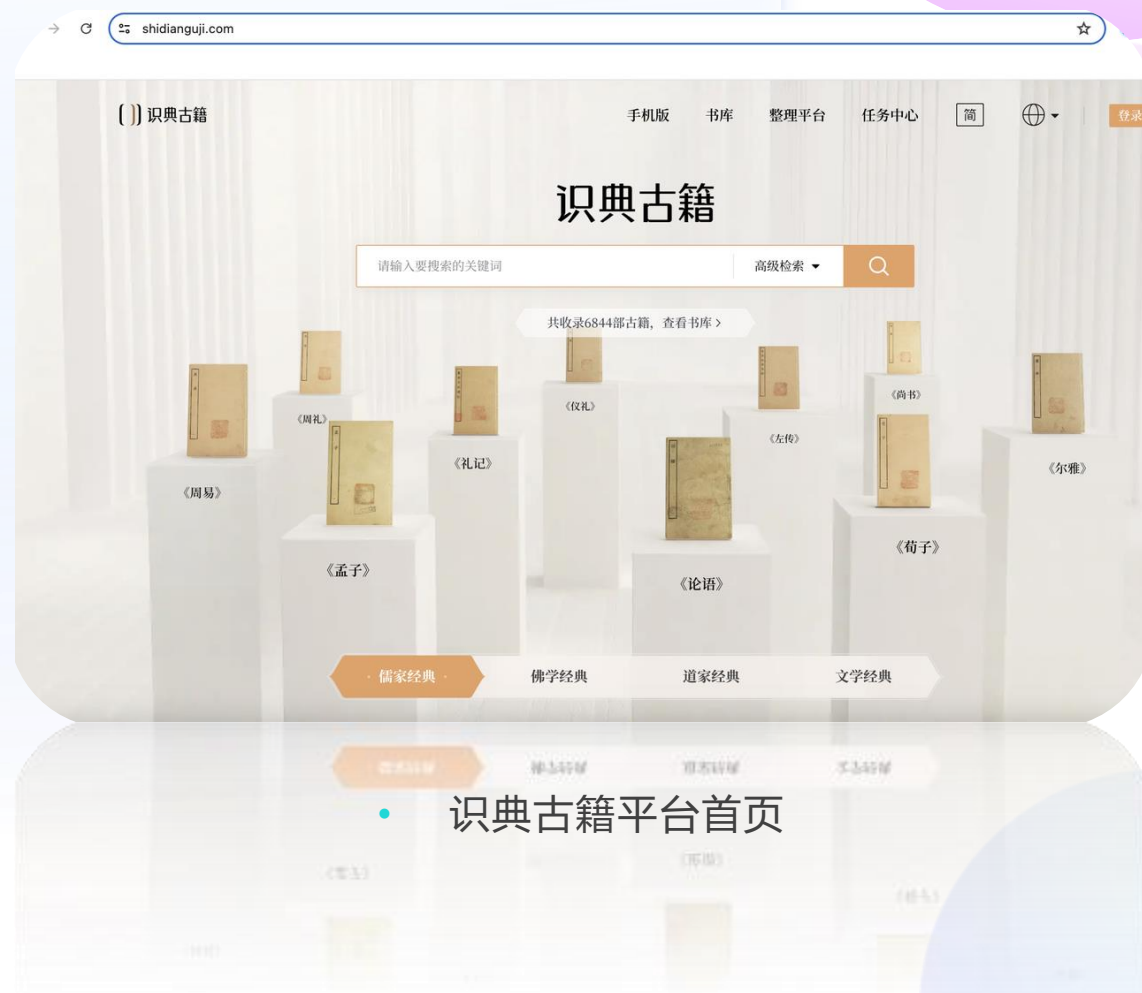
 识典古籍

“识典古籍” 研发背景

- 中国是世界上古籍典藏最丰富的国家，据统计现存古籍20余万种。
- 要把古籍蕴含的思想和智慧传承下去，不仅要**对古籍实体进行保护**，也应该**为古籍的阅读和研究提供便捷条件**。
- 但目前，我国大量古籍**分散收藏在各个图书馆或收录于不同的数据库**，获取难度大，并且难以进行高效检索、关联阅读和深度挖掘与利用。
- 同时，由于古籍的表述方式和现代语言存在较大差异，很多古词语和语法结构对于现代人来说难以理解，**导致古籍研究难度大、大众理解难**等情况，影响了古籍在现代传播。

“识典古籍” 研发背景

- 基于以上情况，抖音集团结合人工智能技术、产品研发能力，与北京大学合作，**打造免费开放、智能的“识典古籍”数字化平台** (www.shidianguji.com)。
- 平台为社会提供公益性服务，让大众可以打破时空限制，在线上获取古籍内容，高效、便捷、无门槛地开展研究、检索和阅读，也可以依托平台，开展古籍数字化整理。



识典古籍平台首页

“识典古籍”的底座支撑技术

- **OCR**：识别出古籍图片影像上的文字
- **自然语言处理(NLP)**：用于句读、分段、实体识别、标题识别、古文分词等
- **大语言模型**：用于白话文翻译、古籍研究智能助手、实体关系提取等
- **WebGL 3D展示技术**：通过3D古籍模型还原古籍原貌，提供真实流畅的古籍阅读体验

“识典古籍” 功能介绍

识典古籍：免费、智能、开放的公益性古籍数字化平台

“识典古籍”平台包括以阅读研究为目的的**阅读平台**，以及以古籍整理为目的的**整理平台**，均面向全社会免费开放。平台从2022年4月启动建设，于2022年10月上线，之后持续不断丰富资源、更新产品能力。

识典古籍阅读平台

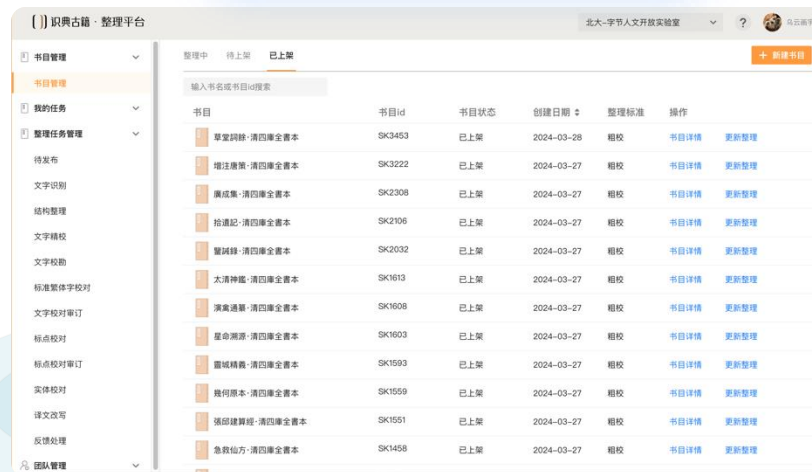


识典古籍Web版
功能丰富的智能化研究工具



手机适配、今日头条“古籍”频道
随时随地查阅古籍

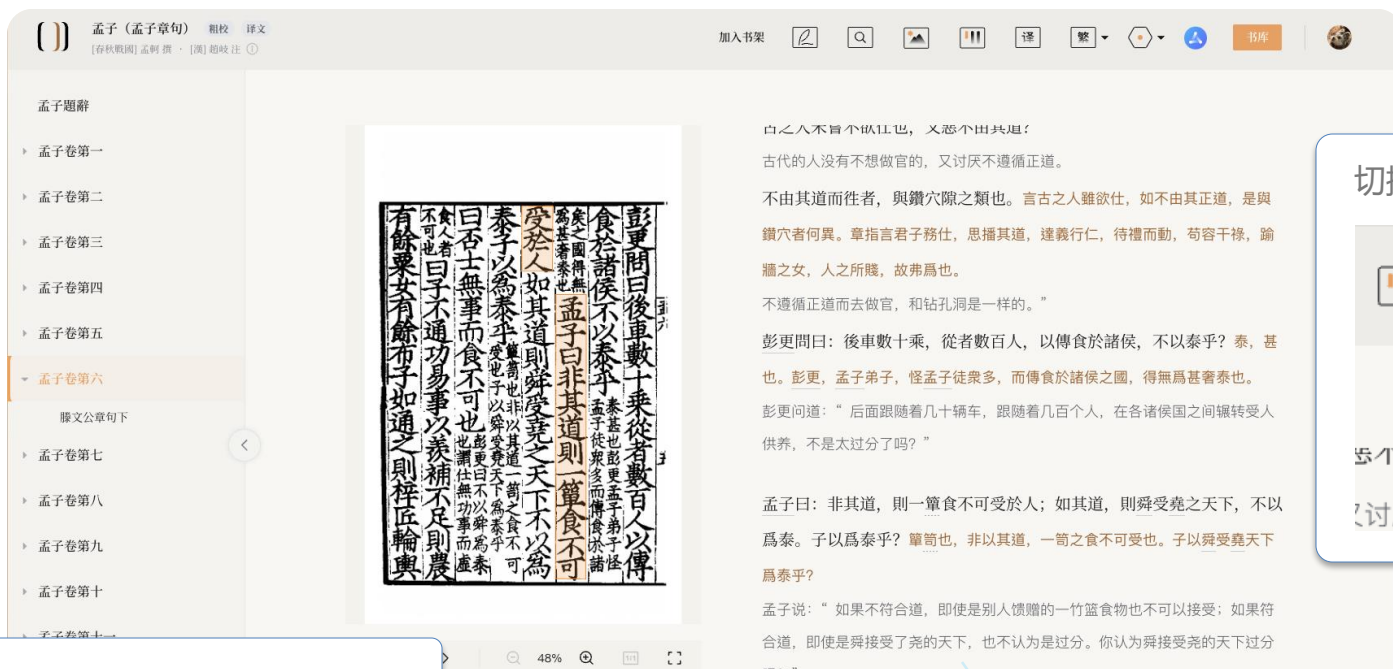
识典古籍整理平台



免费、开放的基于人工智能的
协作化古籍整理平台

“识典古籍” 阅读平台功能

图文对照、繁简转换、文白对照



忠于底本，文字可定位原图

孟子曰：非其道，則一簞食不可受於人；如其道，則舜



问AI



标记



写想法



复制



查看引用



定位原图



问题反馈

常见古籍提供文白对照

切换注疏；繁简体字或底本原字



译

繁

简

机器简体

繁

标准繁体字

原

底本原字

五个出共担！

讨厌不遵循正道。

“识典古籍” 阅读平台功能

字典释义、实体百科

也。彭更，孟子弟子，怪孟子徒众多，而传食于诸侯之国，得无为甚奢泰也。

彭更问道：“后面跟着几十辆车，跟着几百个人，在各诸侯国之间辗转受人供养，不是太过分了吗？”

孟子曰：非其道，则一簞食不可受于人；如其道，则舜受尧之天下，不以

也。子以舜受尧天下

者，不可也。

彭更说：“不。士人不做事而吃饭，是不

曰：子不通功易事，以养补不足，则

梓匠轮舆皆得食于子。孟子言凡人当通功易事，乃可各以奉其用。梓匠，木工

也。轮人、舆人，作车者交易则得食于子之所有矣。周礼攻木之工七，梓匠、

轮、舆，是其四者，养余也。于此有人焉，入则孝，出则悌，守先王之道，

以待后之学者，而不得食于子。子何尊梓匠轮舆而轻为仁义者哉？入则事

搜本书 搜全库 查字典

簞

簞

原文繁体：簞 简体：簞

簞《汉语大词典》

dān

1.古代用来盛饭食的盛器。以竹或苇编成，圆形，有盖。2.竹或苇编的小箱。3.瓢类盛器。参见“簞醪投川”。4.见“簞竹”。

内容来源：《汉语大词典》

查看更多>

阅读中遇到疑难字词，选中就会弹出《汉语大词典》《康熙字典》和《说文解字注》的相关词典数据，帮助人们理解词义。

公明仪

战国时代音乐家

公明仪是战国时代的音乐家，出生于鲁国南武城。他自幼有音乐天赋，但家境贫寒，无法买乐器，只好跳踯踏舞。后来，一位宫廷乐师将毕生钟爱之琴馈赠于他。公明仪因此欣喜若狂，常常见人就笑，被人认为是疯了。公明仪的生活年代和身份有多种不同的记载，有的说他曾问礼于孔子，有的认为他是孟子的弟子。他也是成语“对牛弹琴”的典故由来。

我要反馈>

查看头条百科

针对重点古籍做了命名实体识别，古籍中的人名、地名、时间、职官、书名都以下划线的方式标记出来。点击实体，便会对应到识典百科中，能够让人们进一步了解这些不熟悉的专名，达到增强阅读的效果。

“识典古籍” 阅读平台功能

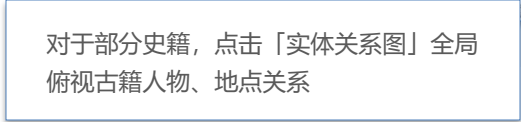
AI问答



可对整篇古籍原文做总结，
或者对任何与古籍相关的
内容进行提问

当遇到一些不太理解的语句，可
以唤起AI问答助手进行解答

还可以询问与该篇古籍相关的其他文本。AI助手
会生成相关内容，点击生成的链接，将直接跳转
对应古籍页面



实体来自原文中实体标记;
使用豆包大语言模型自动提
取文中实体关系

提供跳转到原文段落的链接：
来自古籍原文、回到古籍原文

“识典古籍” 整理平台功能

文字识别、自动标点、自动翻译、多人协作等

识典古籍整理平台具有**可多人协作、覆盖古籍整理全流程、智能化等特点**。平台整合了古籍文字识别(OCR)、自动标点、自动命名实体识别、自动翻译等人工智能工具，以及配套的协作校对工具等，其中OCR准确率已经达到了96%至97%。

识典古籍整理平台的推出，方便了有古籍整理需求的机构、学者和个人借助人工智能技术和多人分工协作的方式**提高古籍整理效率**，这将带动更多古籍被数字化，促进古籍知识普惠。

识典古籍·整理平台

北大-字节人文开放实验室

?

马云雷

书目管理

我的任务

整理任务管理

待发布

文字识别

结构整理

文字精校

文字校勘

标准繁体字校对

文字校对审订

标点校对

标点校对审订

实体校对

译文改写

反馈处理

团队管理

整理中

待上架

已上架

+ 新建书目

批量申请上架

输入书名或书目id搜索

☐ 全选本页

当前选中 0 项

书目	书目id	待确认内容卷	待发起整理卷	整理中	创建日期	操作
☐ 周易·民国四部丛刊景长沙景氏藏古堂藏明翻宋岳氏刊本	SBC0004	0	12	0	2023-03-03	书目详情 发起整理
☐ 诗经·民国四部丛刊景长沙景氏藏古堂藏明翻宋刊巾箱本	SBC0003	0	0	1	2023-03-03	书目详情
☐ 周易·民国四部丛刊景上海涵芬楼藏宋刊	SBC0001	0	0	1	2023-03-03	书目详情
☐ 洞真太上八素真经精曜三景妙诀·明正统道藏本	DZ1320	0	1	0	2023-02-21	书目详情 发起整理
☐ 太上清静元洞真玉字妙经·明正统道藏本	DZ0986	0	0	1	2023-02-21	书目详情
☐ 正一宫教清静道经·明正统道藏本	DZ0799	0	0	1	2023-02-21	书目详情
☐ 太上玄黄净明洞仙度人经法·明正统道藏本	DZ0563	0	0	1	2023-02-13	书目详情
☐ 金阙帝君三元真一经·明正统道藏本	DZ0253	0	0	1	2023-02-13	书目详情
☐ 大洞经吉祥神咒法·明正统道藏本	DZ1461	0	0	1	2023-02-13	书目详情
					2023-02-13	书目详情
					2022-12-29	书目详情 发起整理
					2022-12-29	书目详情
					2022-12-29	书目详情 发起整理
					2022-12-18	书目详情

整理平台运用OCR、自动标点、自动分段、自动校勘、命名实体识别、自动翻译等人工智能技术。

全部技术和功能免费开放，最终生成一个可开放共享的古籍文本，并发布在阅读平台中。

包含文字识别、文字精校、文字校勘等文字录入检校环节；上传古籍后，可根据实际工作需要选择不同流程；可添加多个用户协作。

选择整理流程

书目: 民国四部丛刊景长沙景氏藏古堂藏明翻宋岳氏刊本《周易》

☒ xml上传12卷 ☐ OCR0卷

粗校流程

☐ 文字识别 ☐ 自动标点 ☐ 自动分段 ☒ 结构整理

精校流程

☐ 文字精校 ☐ 文字校勘 ☐ 标准繁体字校对 ☐ 文字校对审订

☐ 标点校对 ☐ 标点校对审订

附加处理

☐ 自动实体标注 ☐ 实体整理 ☐ 译文改写

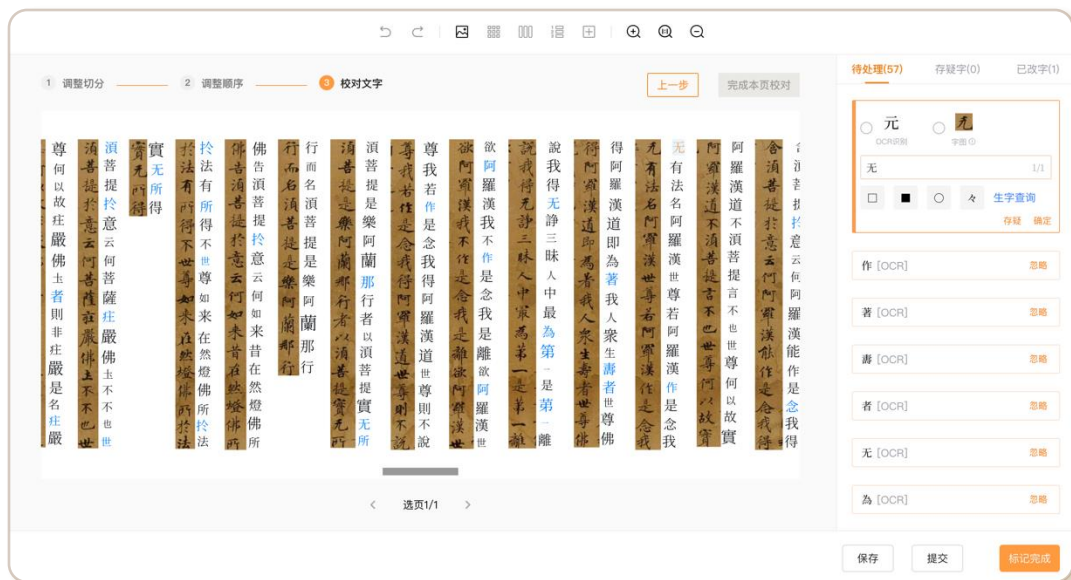
☐ 责任人自动领取后续环节任务

取消

发起整理

“识典古籍” 整理平台功能

整理功能示例



上传图片后，OCR技术可以将古籍图像的文本进行识别。
对于OCR不确定的字，会有颜色标记，以行对齐的方式，
便于定位这些有问题的文字、参照底本修改。



自动实体识别可以对人名、地名、书名、时间和职官进行标记。
这些实体可以人工再进行修改，找出其中的错误或缺漏。
识别到的实体信息，也可以与百科词条相对应。

“识典古籍”运营数据及所获荣誉

已免费公开**6800余种**古籍（截止2024年8月）

累计有超过**8706万人次**通过网页版、今日头条古籍频道等访问识典古籍平台

每月超过**240万用户**通过“识典古籍”各渠道查阅古籍

识典古籍

- 入选凤凰网《行动者联盟2022“年度十大公益创意”》
- 入选中国互联网发展基金会、中国乡村发展基金会主办的《e起向善·数字公益”“网络正能量”2022年度项目》
- 入选中国计算机学会《2023年度技术公益案例集》
- 入选中国网络社会组织联合会《2024互联网平台企业履行社会责任评估报告》优秀实践案例

“识典古籍” 古籍数字化案例——《永乐大典》

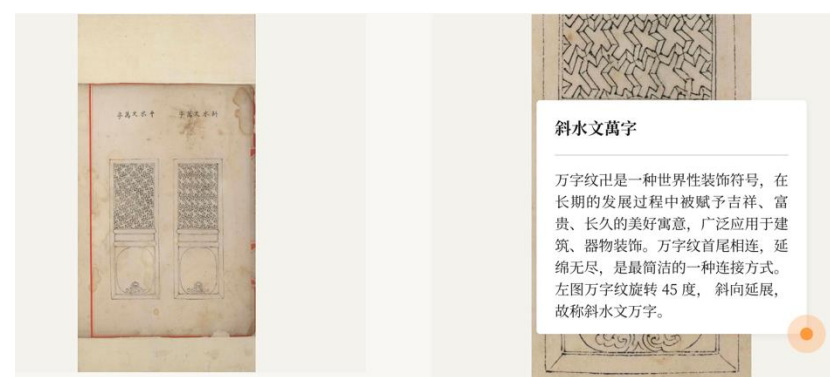
- 《永乐大典》是明成祖(朱棣)永乐年间编纂的一部大型百科全书，共22877卷，11095册，保存了中国14世纪以前的文学、艺术、史地、哲学和应用科学等方面的丰富资料，是中华典籍瑰宝。如今《永乐大典》副本仅有400余册存世，分布在8个国家和地区，其中国家图书馆共收藏《永乐大典》224 册。
- 2023年2月，《永乐大典》中40册在“识典古籍”平台上线，让大典有了数字居所，大众也可以通过“识典古籍”平台免费阅读这部珍贵典籍。



《永乐大典》上线“识典古籍”平台



用户可以进行繁简转换、高清底本对照、查看引文等，高效顺畅地读懂大典内容



对重点内容进行知识标引示范，降低了内容理解门槛

“识典古籍” 古籍数字化案例——《永乐大典》

- 《永乐大典》的文献价值和史料价值主要表现在其引述之丰，全书引用了我国明初以前七八千种图书。因此，不少失传的古书都是从大典中进行辑录而重现于世。明朝至今，已有上千种古籍（含方志）从《永乐大典》中辑出。
- 同时，为了实现沉浸式体验，“识典古籍”平台还通过**Web 3D、光影还原等交互技术**对《永乐大典》原貌进行还原，让用户可以更加身临其境地“触摸”大典。



搜全文	
找到 34 条搜索结果：水（隸、林）经（經、彙、經、垚、垚、垚、經）	
水經五	[明] 解缙等 辑《永乐大典》永樂大典卷之一萬一千一百三十一【八脂】“水”
水經六	[明] 解缙等 辑《永乐大典》永樂大典卷之一萬一千一百三十二【八脂】“水”
水經三	[明] 解缙等 辑《永乐大典》永樂大典卷之一萬一千一百二十九【八脂】“水”

在《永乐大典》数据库“缀玉”板块，用户只要点击或检索大典引用的书籍名称，即可将对应书目的内容找出来

以大家耳熟能详的《水经注》为例，只要检索“水经注”，即可找到此次公开的40册《永乐大典》中引用的《水经注》全部内容



《永乐大典》3D展示

媒体、专家报道和评价

- **新华每日电讯**报道，字节跳动联手北大研发古籍数字化平台，利用智能技术加速中华古籍资源的数字化建设，向全社会提供公益化服务。科技创新正在让万千古籍重返“青春期”。
- **光明网**报道，以“北京大学-字节跳动数字人文开放实验室”为研发基地构建的古籍数字化平台，让古籍能够以文本的形态加以检索、关联阅读和深度挖掘与利用。这样的“工具箱”打开，已是在纸面上开启了一场“凿空之旅”。
- **三联生活周刊**报道，六百多年前，中国两千名文化精英耗费五年心力，编写了世界上最大的百科全书。百年的动荡中，这部大书流离失所。现在，存世的40册《永乐大典》终于结束了多舛的命运，更令人欣慰的是，每个人都可以登录网站，以图文对照的方式，一览大典风采，还可以体验字韵索引，查阅内容，宝典有了数字居所。
- **丽江市东巴文化研究院院长李德静**，从事东巴古籍文献抢救及翻译整理等工作已有数十年时间，她表示：“东巴文史诗《创世纪》知识库在“中华古籍资源库”“识典古籍”平台上线非常有实践意义，其将古老的文化进行解读并转化为数字产品，呈现方式也符合现代人的思维方式，尤其符合年轻人的审美和阅读习惯，能够更好地传播与传承我们古老的民族文化。”

*点击链接可查看具体内容