**mineru测试文档**

* 删除页眉、页脚、脚注、页码等元素，确保语义连贯
* 输出符合人类阅读顺序的文本，适用于单栏、多栏及复杂排版
* 保留原文档的结构，包括标题、段落、列表等
* 提取图像、图片描述、表格、表格标题及脚注
* 自动识别并转换文档中的公式为LaTeX格式
* 自动识别并转换文档中的表格为HTML格式
* 自动检测扫描版PDF和乱码PDF，并启用OCR功能
* OCR支持84种语言的检测与识别
* 支持多种输出格式，如多模态与NLP的Markdown、按阅读顺序排序的JSON、含有丰富信息的中间格式等
* 支持多种可视化结果，包括layout可视化、span可视化等，便于高效确认输出效果与质检
* 支持CPU和GPU环境
* 兼容Windows、Linux和Mac平台

|  |  |  |  |
| --- | --- | --- | --- |
| 操作系统 | Linux 22.04 | Windows10/11 | MacOS 11+ |
| CPU | x86\_64不支持ARM Linux | x86\_64不支持ARM Windows | x86\_64 |
| 内存 | 大于等于16GB，推荐32GB以上 | | |

MinerU是一款将PDF转换为机器可读格式的工具，可以很方便的抽取为任意格式。MinerU诞生于书生-葡语的训练过程中，我们将会集中经历解决科技文献中的符号转换问题，希望在大模型时代为科技发展做出贡献。相比国内外知名商用产品MinerU还很年轻，如果遇到问题及结果不及预期请到issue提交问题，同时附上相关PDF；



**Pre-installation Notice—Hardware and Software Environment Support**

To ensure the stability and reliability of the project, we only optimize and test for specific hardware and software environments during development. This ensures that users deploying and running the project on recommended system configurations will get the best performance with the fewest compatibility issues.

By focusing resources on the mainline environment, our team can more efficiently resolve potential bugs and develop new features.

In non-mainline environments, due to the diversity of hardware and software configurations, as well as third-party dependency compatibility issues, we cannot guarantee 100% project availability. Therefore, for users who wish to use this project in non-recommended environments, we suggest carefully reading the documentation and FAQ first. Most issues already have corresponding solutions in the FAQ. We also encourage community feedback to help us gradually expand support.



MinerU is a tool that converts PDFs into machine-readable formats (e.g., markdown, JSON), allowing for easy extraction into any format. MinerU was born during the pre-training process of [InternLM](https://github.com/InternLM/InternLM). We focus on solving symbol conversion issues in scientific literature and hope to contribute to technological development in the era of large models. Compared to well-known commercial products, MinerU is still young. If you encounter any issues or if the results are not as expected, please submit an issue on [issue](https://github.com/opendatalab/MinerU/issues) and **attach the relevant PDF**.