

编辑

**** 【个人主页：玄同765】 ****

作为开发者，你是否在使用 AI
智能体时遇到过这些文档处理痛点：

今天，我作为mcp_documents_reader的作者玄同765，
向你介绍 这款解决多格式文档读取痛点的 MCP
工具——

它支持 Excel（XLSX/XLS）、DOCX、PDF、TXT 等多种主流格式，能让智能体快速提取文档纯文本内容，同时提供 GitHub+Gitee 双仓库支持，国内用户访问更顺畅。

一、工具核心亮点：轻量高效的多格式文本读取

1. 多格式统一支持

无需切换多个工具，一个mcp_documents_reader就能搞定所有主流文档格式的纯文本提取：

2. 统一调用接口

不管是哪种格式，都可以通过同一个接口调用，智能体无需区分格式，降低使用成本。

3. 大文件优化

针对大体积文档（如 100MB + 的 PDF、10 万行 + 的 Excel），工具会自动分段读取，避免内存溢出，保证运行流畅。

4. 双仓库支持

提供 GitHub+Gitee 双仓库，国内用户可通过 Gitee 快速克隆和安装，解决网络访问慢的问题：

二、快速上手：安装与配置

1. 前置依赖

```
curl -LsSf https://astral.sh/uv/install.sh | sh
```


Windows 用户可参考[uv 官方文档](https://docs.astral.sh/uv/ "uv 官方文档")安装；也可直接用 pip 5

2. 安装方式

方式 1：uvx 一键启动（推荐，无需克隆仓库）

```
# GitHub源
uvx --from git+https://github.com/xt765/mcp_documents_reader mcp_documents_reader
```

```
# 国内用户推荐Gitee源
uvx --from git+https://gitee.com/xt765/mcp_documents_reader mcp_documents_reader
```

启动成功后，工具默认运行在<http://localhost:8080/mcp>。

方式 2：本地克隆安装

```
# GitHub克隆
git clone https://github.com/xt765/mcp_documents_reader.git

# 国内用户推荐Gitee克隆
git clone https://gitee.com/xt765/mcp_documents_reader.git

cd mcp_documents_reader
# 安装依赖
pip install python-docx PyPDF2 openpyxl
# 启动工具
python mcp_documents_reader.py
```

3. Trae IDE 配置

将工具集成到 Trae IDE，让智能体可以直接调用：

Github源：

```
{
  "mcpServers": {
    "mcp-document-reader": {
      "command": "uvx",
      "args": [
        "--from",
        "git+https://github.com/xt765/mcp_documents_reader",
        "mcp_documents_reader"
      ]
    }
  }
}
```

Gitee源：

```
{
```

```
    "mcpServers": {
      "mcp-document-reader": {
        "command": "uvx",
        "args": [
          "--from",
          "git+https://gitee.com/xt765/mcp_documents_reader",
          "mcp_documents_reader"
        ]
      }
    }
  }
}
```

4. 环境变量配置

可通过环境变量指定文档存储目录（默认：./documents）：

```
# macOS/Linux
export DOCUMENT_DIRECTORY="/path/to/your/documents"

# Windows (PowerShell)
$env:DOCUMENT_DIRECTORY="C:\path\to\your\documents"
```

三、详细用法：一键提取文档纯文本

工具提供统一接口read_document（自动识别格式）和单独格式接口（read_excel/read_docx/read_pdf），以下示例基于 Trae IDE 智能体。

1. 统一接口调用（推荐）

无需指定格式，工具自动识别文档类型并提取纯文本：

帮我读取本地的《2024年销售数据.xlsx》，提取所有文本内容并总结核心信息。

工具执行说明：Excel 文件会合并所有工作表的纯文本内容，返回给智能体进行总结。

2. 单独格式接口调用

如果需要明确指定格式，可调用对应接口：

帮我用read_pdf工具读取本地的《技术架构手册.pdf》，提取文本后分析核心组件。

3. 各格式提取说明

四、实战场景：提升 AI 智能体文档处理效率

场景 1：Excel 销售数据快速总结

智能体指令：

帮我读取本地的《2024年销售数据.xlsx》，提取文本后总结各地区的销售总额情况。

预期输出：

```
### 2024年销售数据总结
1. 华东区销售总额为1250.8万元，占比35%；
2. 华南区销售总额为980.5万元，占比27%；
3. 华北区销售总额为760.2万元，占比21%；
4. 西南区销售总额为520.1万元，占比14%；
5. 西北区销售总额为110.4万元，占比3%。
```

场景 2：DOCX 需求文档快速梳理

智能体指令：

帮我读取本地的《项目需求文档.docx》，提取文本后总结核心功能模块。

预期输出：

```
### 项目核心功能模块总结
项目包含三大核心模块：
1. 用户管理模块：支持用户注册、登录、信息修改、密码找回；
2. 订单管理模块：支持订单创建、支付、查询、取消；
3. 支付管理模块：支持微信支付、支付宝支付、银行卡支付三种方式。
```

场景 3：PDF 技术手册快速提取

智能体指令：

帮我读取本地的《技术架构手册.pdf》，提取文本后列出微服务核心组件。

预期输出：

```
### 微服务核心组件列表
1. API网关
2. 用户服务
3. 订单服务
4. 库存服务
5. 配置中心
6. 日志中心
```

四、当前限制与未来规划

当前功能限制

目前工具处于轻量版阶段，暂不支持以下功能：

未来规划

后续版本将逐步开发以下功能：

五、总结与反馈

`mcp_documents_reader`

作为一款轻量高效的多格式文档读取 MCP 工具，解决了 AI 智能体处理多格式文档的文本提取痛点，同时提供 GitHub+Gitee 双仓库支持，国内用户访问更顺畅。

如果你在使用过程中遇到问题，或有新功能需求，欢迎通过以下方式反馈：

☐ 如果你觉得工具好用，别忘了给仓库点个 Star，让更多开发者受益！