



产品数据表 · 捕获 · 网站来源

# 这些内容都在您的网站上，还要有人重新打字？ 您的网站成为智能体的知识。

网站来源会按计划读取您网站上公开的 HTTPS 页面。每个页面都会在您选择的知识库中成为一篇内部文章，您的智能体据此在电话和网站聊天中作答。

## 概览

每个来源最多 2000 个页面，链接深度最多 6 层，计划从每 6 小时到每月一次。变动的页面会被重新写入，消失的页面在两次运行后下线。

01 · 内容概述

# 同一个答案，两个版本。 现在由您的网站确定事实。

营业时间、产品页面和帮助文字早已挂在您的网站上。为智能体重新誊写既费时，网站上每一次改动又会让副本过时。网站来源直接读取页面本身，并在下一次运行时带来这些变化。

## 输入

以 https 开头的起始网址，可选网站地图，以及同一主机上的更多起始网址

## 结果

每个页面一篇内部文章，以页面网址为来源，并标明检测到的语言

## 计划

每 6 或 12 小时、每日、每周、每 14 天或每月

## 读者

电话和网站聊天中的对话智能体，以及 Agent AI 中的聊天智能体



## 一个主机，多个页面

从起始网址出发，爬虫会沿同一主机上的链接前进，无论是否带 www。robots.txt 中的网站地图以及您自己的网站地图会扩充列表。

## 设定范围

可为每个来源设置最大页面数和链接深度，并可设定路径前缀、排除通配符、带查询字符串的网址，以及 /tag 之类的列表路径。

## 包含 JavaScript

Fast 读取 HTML，Full 在浏览器中渲染，Auto 在页面看似空白时自动切换。例如，点击选择器可展开常见问题的折叠面板。

## 干净的文章

CSS 排除功能可将导航或页脚等区域排除在文章之外。如果您愿意，文章可从页面中提取一段简短摘要。

## 遵守网站规则

尊重 robots.txt 和 noindex。如果已建立索引的页面之后被屏蔽，爬虫会下线其对应的文章。

## 只更新变化的部分

爬虫通过内容识别已变化的页面，未变化的页面保持不动。消失的页面在连续两次运行未命中后下线。

02 · 工作原理

# 添加来源，检查， 分配给智能体。

添加网站来源只需填写几个字段。预览会在首次运行之前显示爬虫会找到什么，之后目录会显示每个网址的处理结果。

1

## 添加来源

在 Knowledge Base AI › 网站来源 下，使用 添加来源：输入名称、起始网址、目标知识库和刷新间隔。

2

## 检查预览

预览 会显示 robots.txt、网站地图和候选页面，且不会建立任何索引。然后设置范围和爬虫。

3

## 运行并监控

首次运行在保存后几分钟开始，立即抓取 会安排另一次运行。索引页面 显示每个网址的状态和 HTTP 代码。

4

## 分配给智能体

在 Conversation AI › 对话代理 下，于 知识 选项卡中选择知识库。此后，智能体即根据网站文章作答。

### 技术数据

| 功能       | 详情                                        |
|----------|-------------------------------------------|
| 输入       | 同一主机上的公开 HTTPS 页面，HTML 文本                 |
| 每个来源的范围  | 最多 2000 个页面（默认 200），链接深度最多 6 层（默认 3）      |
| 每个账户的来源数 | 最多 25 个                                   |
| 爬虫       | Fast、Full 或 Auto，可设点击和等待选择器               |
| 输出       | 每个页面一篇内部文章，含页面网址、检测到的语言，以及标签 website 和主机名 |
| 概览       | 索引页面 显示状态、渲染方式、HTTP 和最近一次抓取； 同步历史 按来源显示   |
| 访问与计费    | Knowledge Base AI › 网站来源 下的 Web 应用；按使用量计费 |

# 网站知识 抵达电话和聊天的地方。

## 网站聊天的操作说明

一家软件公司将操作说明放在 /docs 下。设置了路径前缀 /docs 的网站来源每天读取它们，网站聊天会在回答中注明对应的页面。

## 热线的产品页面

一家零售商每周读取其产品页面。电话中的对话智能体会根据客户在网上看到的相同文字，回答关于尺码和材质的问题。

## 网站加上您自己的文章

凡是网站上没有的内容，例如假期停业，您可以在同一知识库中添加为自己的文章。智能体会同时读取两者。

## 常见问题

### 网站来源会读取哪些页面？

同一主机上的公开 HTTPS 页面，从起始网址出发，如有需要还可通过网站地图和其他起始网址获取。它读取 HTML 页面的文字，无需登录、Cookie 或凭据。

### 我对网站文章的自行修改会怎样？

页面变化时，下一次运行会重新写入文章。因此，补充内容应放在同一知识库中单独的文章里。

### 智能体的知识有多新？

与最近一次运行一样新。您可以选择从每 6 小时到每月的计划，立即抓取 会安排额外的运行。两次手动运行之间相隔 15 分钟。

### 这些页面会出现在客户门户中吗？

网站文章属于内部：您的团队和该知识库的智能体可以阅读，它们不会出现在客户门户中。

## 开始使用

### 把您的网站设为第一个来源。

在 Web 应用中进入 [Knowledge Base AI](#) > [网站来源](#)，使用 [添加来源](#) 输入起始网址，检查预览后保存。首次运行会在几分钟后开始。

[paperoffice.ai/zh/ai-web-sources-wang-zhan-pa-chong-zhi](https://paperoffice.ai/zh/ai-web-sources-wang-zhan-pa-chong-zhi)



## 了解更多模块

[paperoffice.ai/zh/ai-web-sources-wang-zhan-pa-chong-zhi](https://paperoffice.ai/zh/ai-web-sources-wang-zhan-pa-chong-zhi)

## 价格

[paperoffice.ai/en/pricing](https://paperoffice.ai/en/pricing)

## 联系我们

[paperoffice.ai/zh/lianxi](https://paperoffice.ai/zh/lianxi)