

GEO Casa · 分册 1 / 4

GEO 按网站

WordPress · 前端渲染站 · 独立站电商 · 文档站 · 技术检查

GEO



页面要过几道关，AI 才看得到？



从上往下查。AI 拿不到的页面，内容做得再好也白费。

[阅读完整指南](#)



robots.txt：海内外三个常见的意外

规则冲突时（包括含通配符的规则），Google 采用限制最少的那条。Google 不支持 crawl-delay。

- 官方 · Google

屏蔽 Baiduspider，会从百度搜索和百度提供的其他搜索结果中消失。百度的爬虫说明没有给 AI 产品单独的标识。

- 官方 · Baidu

一次覆盖 130 个公司网站的 2026 年审计中，作者认为把检索和训练两类爬虫一起封掉，是「最容易犯、也最致命的一个错」。

- 实践观察 · 2026-08 · DATA17 AI Research Lab · 解数智能

robots.txt 管的是抓取。国内别为了挡 AI 去封 Baiduspider 或 Bytespider，搜索会一起丢掉。



试一试：屏蔽一个爬虫，看会失去什么

搜索类爬虫和训练类爬虫是两个独立的决定。点一下某个标识，把它屏蔽掉。

OAI-SearchBot

已放行

OpenAI · ChatGPT 搜索

不会出现在 ChatGPT 搜索回答中，仍可能以导航链接出现。

- 官方 · OpenAI

GPTBot

已放行

OpenAI · 模型训练

内容不用于训练，不影响搜索。

- 官方 · OpenAI

PerplexityBot

已放行

Perplexity · Perplexity 搜索

Perplexity 建议放行以便出现在结果中，屏蔽的具体效果没有说明。

- 官方 · Perplexity

Google-Extended

已放行

Google · Gemini 训练和依据生成

不影响 Google 搜索收录，包括 AI 概览。

- 官方 · Google

Claude-SearchBot

已放行

Anthropic · Claude 搜索索引

可能降低在 Claude 回答中的可见度。

- 官方 · Anthropic

ClaudeBot

已放行

Anthropic · 模型训练

内容不用于训练。

- 官方 · Anthropic

开关演示。每条效果都来自爬虫所属公司的官方文档。



国内：屏蔽一个搜索爬虫，看会失去什么

国内搜索引擎各有各的爬虫规则。点一下某个标识，把它屏蔽掉。

Baiduspider

已放行

百度 · 百度搜索

从百度搜索和百度提供的其他搜索结果中消失。

- 官方 · Baidu

Baiduspider-ads

已放行

百度 · 广告，按客户协议工作

没有效果：它不遵守 robots.txt。

- 官方 · Baidu

360Spider

已放行

360 搜索 · 360 搜索

支持 user-agent、allow 和 disallow；图片和视频有各自的标识。

- 官方 · 360 Search

Sogou web spider

已放行

搜狗 · 搜狗搜索

被禁止的网页在搜狗搜索中检索不到；修改可能要几周才生效。

- 官方 · Sogou

Bytespider

已放行

头条搜索 · 头条搜索

它 2019 年的介绍页没有提到 robots.txt，屏蔽效果没有说明。

- 官方 · Toutiao Search

开关演示。每条效果都来自爬虫所属公司的官方文档。



误区：Google-Extended 与 AI 概览

? 屏蔽 Google-Extended，我的页面就不会进 Google 的 AI 概览。

不成立

Google-Extended 管的是 Gemini 训练和依据生成，AI 概览归 Googlebot 规则和 nosnippet 管。国内没有对应的开关：百度的爬虫说明没有给 AI 产品单独的标识。

「Google-Extended 不影响网站在 Google 搜索中的收录，也不作为 Google 搜索的排名信号。」

● 官方 · Google

[阅读完整条目](#)



robots.txt 放行了，爬虫就一定进得来吗？

Google 把「确保 robots.txt 以及任何 CDN 或托管基础设施都允许抓取」列为对 AI 功能仍然重要的做法之一。

- 官方 · Google Search Central

从 2026 年 9 月 15 日起，Cloudflare 新域名的默认设置会在展示广告的页面上拦截训练类和代理类机器人，搜索类爬虫仍然放行。

- 官方 · Cloudflare

「robots.txt 允许」和「真的能访问」是两项检查。国内也一样：别让防火墙把 Bytespider、Baiduspider 误拦了。



试一试：哪个爬虫是真的？

谁都能在 User-Agent 里写上「GPTBot」。点一行，看怎么分辨。

```
66.249.66.1 - - [28/Oct/2026:09:12:04] "GET /pricing/" 200 "Googlebot/2.1 (+http://www.google.com/bot.html)"
```

真爬虫 反向 DNS 得到 googlebot.com 的主机名，再正向查询，解析回同一个 IP。

```
203.0.113.45 - - [28/Oct/2026:09:13:51] "GET /pricing/" 200 "Mozilla/5.0; compatible; GPTBot/1.1"
```

冒充的 这个 IP 不在 OpenAI 公布的 gptbot.json 列表里。User-Agent 里的名字证明不了什么。

```
198.51.100.7 - - [28/Oct/2026:09:15:20] "GET /docs/" 200 "Mozilla/5.0; compatible; Claude-SearchBot/1.0"
```

需要核实 先对照 Anthropic 爬虫帮助文章里链接的 IP 列表，再决定算不算这次访问。

```
192.0.2.10 - - [28/Oct/2026:09:16:02] "GET /" 403 "Mozilla/5.0; compatible; Baiduspider/2.0"
```

需要核实 做反向 DNS 查询。百度说真正的主机名以 .baidu.com 或 .baidu.jp 结尾。如果是真的，这个 403 就是你的防火墙拦的。

模拟日志，IP 为示例地址。只统计验证过的访问。

Google 的爬虫可以通过先反向、再正向 DNS 查询验证，或对照 Google 公布的 IP 列表验证。

- 官方 · Google



试一试：验证国内爬虫

国内各家引擎的验证方法不一样。点一行看看。

```
192.0.2.21 - - [28/Oct/2026:10:02:11] "GET /products/" 200 "Mozilla/5.0; compatible; Baiduspider/2.0" rDNS=crawl.baidu.com
```

真爬虫 反向 DNS 得到的主机名以 .baidu.com 结尾，百度说这是真的。

```
203.0.113.88 - - [28/Oct/2026:10:03:40] "GET /products/" 200 "Mozilla/5.0; compatible; Baiduspider/2.0" rDNS=host.example.net
```

冒充的 百度称 *.baidu.com 和 *.baidu.jp 以外的主机名都是冒充的。

```
198.51.100.30 - - [28/Oct/2026:10:05:02] "GET /about/" 200 "Mozilla/5.0 (compatible) 360Spider"
```

需要核实 360 说暂不支持 nslookup 验证。要对照它公布的 IP 段。

```
198.51.100.77 - - [28/Oct/2026:10:06:15] "GET /about/" 200 "Mozilla/5.0; compatible; Bytespider"
```

需要核实 对照头条 2019 年 11 月公布的六个 /24 IP 段。这份列表已经比较旧。

模拟日志，IP 为示例地址。rDNS 表示反向 DNS 查询结果。

百度建议对 IP 做反向 DNS 查询。Baiduspider 的主机名是 *.baidu.com 或 *.baidu.jp，百度称其他主机名都是冒充的。

● 官方 · Baidu



每个状态码在告诉 Google 什么

爬虫在读页面内容之前，先看状态码。

状态码	Google 怎么处理	依据
200，但页面是空白或报错	不保证收录，会被报告为软 404	官方
301	跟随，当作处理目标网址的强信号	官方
302	跟随，但只当作弱信号	官方
4xx（429 除外）	不收录，已收录的网址会被移除	官方
用 401 / 403 限速	对抓取速度没有影响，Google 明确说不要这样做	官方
429 / 5xx	抓取放慢，持续下去已收录网址会被移除	官方

来源：Google 关于 HTTP 状态码的文档。国内搜索引擎和 AI 助手：未见公开的同类说明。



摘要控制同样作用于 AI 概览

`nosnippet` 会阻止内容被直接用作 AI 概览和 AI 模式的输入。`max-snippet` 会限制可使用的内容量。

- 官方 · Google Search Central

这些设置「只有在爬虫被允许访问包含这些设置的页面时，才能被读取和遵循」。

- 官方 · Google Search Central

模板里遗留的 `nosnippet` 会把页面挡在 AI 概览之外。国内助手有没有类似的摘要控制：未见公开说明。



Sitemap 和 IndexNow 分工不同

点一下卡片。

XML Sitemap



网址清单。单个文件最多 50,000 个网址或 50 MB，更大的网站用索引文件。

<lastmod>



只有始终准确时 Google 才会用它。priority 和 changefreq 会被忽略。

IndexNow



通知参与的搜索引擎有网址新增、更新或删除，单次最多 10,000 个。

Google



不在 IndexNow 接口之列。给 Google 维护一份真实的 Sitemap。

一次 IndexNow 提交会通知所有参与的搜索引擎，但「每个引擎自行决定是否将其编入索引」。

● 官方 · [IndexNow.org](https://indexnow.org)



国内：各家 Sitemap 的限制不一样

百度 Sitemap 每个文件最多 50,000 个网址，大小须小于 10MB，gzip 压缩前也要满足。支持 XML、纯文本和索引文件，编码可用 UTF-8 或 GBK。

• 官方 · Baidu

360 搜索同样规定单个文件不超过 10MB、不超过 50,000 个网址，编码可用 UTF-8 或 GBK，Sitemap 索引文件不超过 500M。

• 官方 · 360 Search

头条搜索要求 UTF-8 编码，单个数据文件不超过 10M、item 不超过 1 万条；sitemapindex 文件不超过 5M、节点不超过 5 万。

• 官方 · Toutiao Search

GBK 编码的 Sitemap 百度和 360 能读，头条不行。用 UTF-8、每个文件不超过 1 万个网址，三家都合适。



误区：Indexing API

? Google 的 Indexing API 能让任何页面快速被收录。

不成立

它只适用于招聘信息和直播视频页面，滥用可能被撤销权限。国内 AI 助手未见公开收录或提交机制，能做的是让百度、360、搜狗、头条的爬虫抓得到你。

「Indexing API 只能用于抓取包含 JobPosting 或嵌入在 VideoObject 中的 BroadcastEvent 的网页。」

● 官方 · Google Search Central

[阅读完整条目](#)



一个首选网址，各语言互相指向

重定向和 `rel="canonical"` 是强信号，Sitemap 收录是弱信号。组合使用能提高首选网址被选中的几率。

- 官方 · Google Search Central

如果两个页面没有互相指向，`hreflang` 标签会被忽略。每个语言版本都应列出自身以及所有其他版本。

- 官方 · Google Search Central

canonical、内链和 Sitemap 指向同一个网址。中外两版各用一个网址，`hreflang` 互指，译文不要 canonical 到原文。



排查：页面没有被收录

在 Search Console 的网址检查工具中打开该网址，然后按顺序回答。

1 实时测试显示允许抓取吗？

⚠️ 删除 robots.txt 规则。如果你想让页面不出现在结果中，robots.txt 管不了这个，该用 noindex。

2 允许编入索引吗（没有 noindex）？

⚠️ 在 robots meta 标记或 X-Robots-Tag 响应头中找到 noindex。常见来源是模板、CMS 设置和预发布环境。

3 Google 选择的规范网址是你的网址吗？

⚠️ 如果两个网页并不重复，让 canonical 标记、内部链接和站点地图都指向你想要的网址。

4 这个网页明显值得收录吗？

⚠️ Google 不保证收录。让网页原创、有用，然后请求编入索引一次即可。

✅ 每次修复后运行「测试实际网址」，再请求编入索引。国内助手未见公开收录机制：先确认国内爬虫能抓到页面，再到 App 里实测。

[阅读完整指南](#)



误区：加了 schema 就会被引用

? 加上 schema.org 结构化数据，AI 搜索就会更常引用我的页面。

未经证实

没有已发表的结果能证明。2026 年的一项实验发现，给本来就被引用的页面加标记没有带来提升，尚未被引用的页面还没有人测过。国内助手未见公开怎么选页，样本中被引用的多是榜单和推荐类内容（行业观察）。结构化数据要与页面可见内容一致。

Google 表示，出现在 AI 概览或 AI 模式中不需要任何特殊的 schema.org 结构化数据。结构化数据还应与页面上的可见文字一致。

● 官方 · Google Search Central

[阅读完整条目](#)

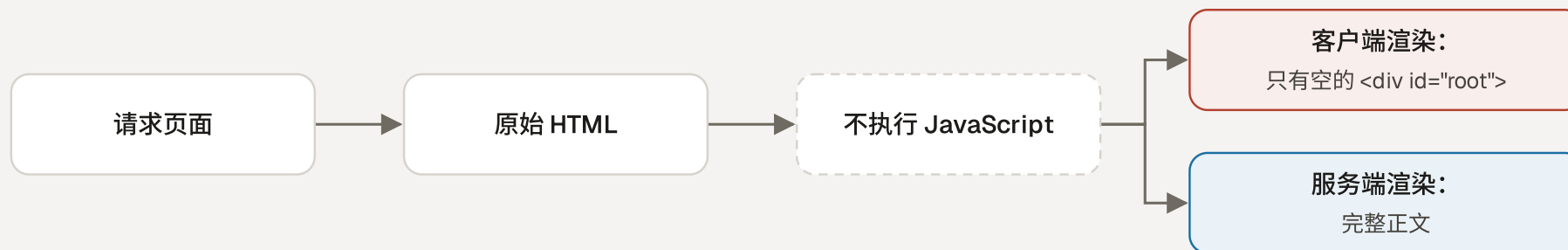


前端渲染站：多数 AI 爬虫只读 HTML

浏览器与 Googlebot



多数 AI 爬虫（2024 年测试）



Vercel 与 MERJ（2024 年 12 月）发现，测试的主要海外 AI 爬虫都不执行 JavaScript，Gemini 例外。国内爬虫是否执行：未见公开，正文放服务端 HTML。

[阅读完整指南](#)



前端渲染站：官方文档怎么说

Google 仍然推荐服务端渲染或预渲染，因为「并非所有机器人都能运行 JavaScript」。

- 官方 · Google Search Central

Google 表示：「动态渲染是一种权宜之计，不是推荐的解决方案。」它推荐服务端渲染、静态渲染或水合。

- 官方 · Google Search Central

Vue 的 SSR 指南提醒：如果应用先显示加载动画再用 Ajax 取内容，「the crawler will not wait for you to finish」（爬虫不会等你加载完）。

- 官方 · Vue.js

不执行 JavaScript 的爬虫，只能看到页面的空壳。



Next.js、Vue、React 团队

前端渲染站：五项检查

- 1 关掉 JavaScript，逐个模板抓一遍
- 2 在服务端渲染正文：SSR、SSG 或预渲染
- 3 链接、元数据和结构化数据也放进 HTML
- 4 给机器人和用户看同一个页面
- 5 每次发版后重新检查

默认爬虫不执行 JavaScript。

[阅读完整指南](#)



WordPress：哪个设置会让整站不被收录？

「建议搜索引擎不索引本站点」会给每个调用 `wp_head()` 的页面加上 `noindex,nofollow`。

- 官方 · WordPress.org

Google 把「CDN 或托管基础设施允许抓取」列为对 AI 功能仍然重要的做法。

- 官方 · Google Search Central

WordPress 上，除了 robots.txt，后台的阅读设置也会给页面加上 noindex。



站长、建站服务商、主机商

WordPress：六步路线

- 1 取消勾选「建议搜索引擎不索引本站点」
- 2 查看实际输出的 robots.txt，放行 AI 搜索爬虫
- 3 检查安全插件、主机和 CDN 的机器人设置
- 4 只保留一份 XML Sitemap，并提交
- 5 只让一个插件输出结构化数据，并与页面一致
- 6 正文放在 HTML 里，别藏在页面构建器脚本中

WordPress 的问题，多半出在设置上。

[阅读完整指南](#)



独立站电商：Feed、标记、页面必须一致

在微软的 Copilot 示例中，一款商品进入前三，是因为 Feed 显示价格有竞争力且有现货。

● 官方 · Microsoft Advertising

Google 表示同时使用页面标记和 Merchant Center Feed，商品结果的展示资格最大。

● 官方 · Google Search Central

DataStory 在六个国内 AI 平台问了九个消费品类：被引用的多是榜单和推荐（22.8%–72.0%），其次是商品介绍，用户评价都不到 3%。

● 实践观察 · 2026-08 · 数说故事

Feed、标记、页面三处价格对不上，AI 就有三个答案可选。



店主、电商运营团队

独立站电商：六步

- 1 弄清三处数据：被抓取的页面、商品 Feed、实时网站
- 2 让搜索爬虫能访问商品页
- 3 商品结构化数据与页面可见内容一致
- 4 提交商品 Feed，价格和库存与网站同步
- 5 商品页回答买家真正会问的问题
- 6 评价靠真实买家积累，绝不造假

Feed、标记和页面上的价格、库存要一致。

[阅读完整指南](#)



文档站：平台已经在做的

GitBook 提供 /llms.txt 和 /llms-full.txt。在任一已发布页面网址后加 .md，就能得到该页的 Markdown。

● 官方 · GitBook

Google 搜索不使用 llms.txt，维护这类文件「既不会损害也不会提升网站在 Google 搜索中的可见度或排名」。

● 官方 · Google Search Central

Ahrefs 对 137,210 个域名的 2026 年 5 月日志显示，已发布的 llms.txt 文件中有 97% 当月没有收到任何请求。

● 实践观察 · 2026-05 · Ahrefs

平台免费送的 llms.txt 就用上，别指望它影响排名。国内助手是否读取 llms.txt：未见公开说明。



文档站与开发者产品

文档是一块一块被引用的。

- ✓ 放行搜索爬虫和用户触发的抓取
- ✓ 每页提供干净的 Markdown 版本
- ✓ 每个代码示例都能单独跑通
- ✓ 每页全文都在服务端 HTML 里
- ✓ 每个版本一个规范网址，「最新版」一目了然
- ✓ 带日期的更新日志；llms.txt 顺手就做

curl 看不到的代码示例，大多数 AI 爬虫也看不到。

[阅读完整指南](#)



AI 智能体像访客一样用你的网站

「智能体主要通过三种方式查看你的网站：截图、原始 HTML 和无障碍树。」现在的智能体会把三种方式结合起来用。

- 官方 · web.dev (Google)

「我们为让网站『便于智能体使用』而提出的所有建议，同样会让网站对人更好用。」

- 官方 · web.dev (Google)

用标准的按钮和链接，表单字段写清标签，布局不要乱跳。



准入检查：每个网址六项

挑几个关键网址来查，不用全站过一遍。

- ☒ 已被索引且可展示摘要
- ☒ 没有被 CDN、防火墙或托管平台拦截
- ☒ 主要内容在服务端 HTML 中
- ☒ robots.txt 允许搜索爬虫
- ☒ 没有意外的 noindex、nosnippet 或 AI 排除设置
- ☒ 日志里有验证过的爬虫来访

六项都通过，才有可能被引用，但不保证一定被引用。

[阅读完整指南](#)



国内站点的技术检查清单

顾客在国内搜索的话，把这几项加进准入检查。

- ☒ robots.txt 放行 Baiduspider、360Spider、搜狗和 Bytespider
- ☒ 百度用反向 DNS 验证；360 和头条对照 IP 段
- ☒ 正文在服务端 HTML 里，不要只在 JavaScript 运行后出现
- ☒ 防火墙和 CDN 不对验证过的国内爬虫返回 403
- ☒ Sitemap 用 UTF-8，每个文件都在各家限制之内
- ☒ 改了 robots.txt，要过几周再判断搜狗有没有生效

国内助手依赖国内的搜索和内容平台，这几家爬虫进不来，助手也就少了你的内容。

[阅读完整指南](#)



最小可行的 GEO：每周几小时

适合一个人打理网站的情况。

- ✓ 第 1 周：确认爬虫能访问、能摘引
- ✓ 列出顾客常问的 20 个问题
- ✓ 每月在两三个 AI 助手的 App（如豆包、DeepSeek）里问一遍并记录
- ✓ 第 2 周：把五个关键页面写清楚、写完整
- ✓ 每周改好一个能回答其中一题的页面
- ✓ 不做特殊 AI 文件、批量改写、买提及

先花一小时查准入。网站被屏蔽，写多少内容都没用。

[阅读完整指南](#)



动笔写内容之前，先查什么？

先确认每一道准入都放行了。每页都浓缩自一篇完整指南，证据细节请看原文。

WordPress

独立站电商

前端渲染站

文档站

技术负责人

最小可行的 GEO

AI 搜索准入检查

验证 AI 爬虫

针对 AI 爬虫的 robots.txt

摘要控制

网页未被收录

回到总览