Chris Green · May 11, 2025 · 3 min read

Crawling a Million Websites in Search of LLMs.txt

Updated: May 12, 2025

TL;DR

- LLMs.txt simplifies website content into a markdown file for easier AI crawling, bypassing technical issues like JavaScript rendering of content
- **Adoption of LLMs.txt is extremely low (0.015% of the Majestic Million - 15 sites!)**
- No major AI company officially supports it yet
- Benefits exist for JavaScript-heavy sites, but there's little concrete evidence
  Implementing LLMs.txt could be worthwhile if simple and cheap
- The ROI is unclear and difficulties in setting up/maintaining it may limit its appeal

EDIT: May 2025 data suggests there are 105 valid LLMs.txt now found - at 600% increase, but still 0.011% of the total.

What is LLMs.txt?

LLMs.txt is a straightforward concept aimed at simplifying website content specifically for large language models (LLMs). It essentially compiles all the site's text content into a single markdown-formatted file, removing technical hurdles such as JavaScript and other limitations that often prevent AI crawlers from fully accessing content.

By consolidating content into markdown, LLMs.txt makes it significantly easier - and theoretically cheaper - for AI models to crawl and process your entire site's content. Even if a full-page fetch is still necessary, the file acts similarly to a sitemap, ensuring complete coverage and aiding discovery.

What are the Benefits?

Where your site overly relies on front-end JavaScript, implementing LLMs.txt could help improve accessibility for AI crawlers. It might also facilitate quicker understanding and use of your content by AI companies.

However, there is little to no concrete evidence of benefits from using this file - so far.

Adoption Among Major AI Companies

Currently, no major AI or machine learning company has formally declared support for LLMs.txt.
However, Anthropic uses one for its documents subdomain, indicating at least preliminary interest or informal adoption within the industry.

It seems logical that other AI companies would at least be examining its potential benefits, even if they haven't publicly endorsed or announced their use.

For sites already implementing an LLMs.txt file, visibility into who is accessing or crawling this file would be extremely valuable.

Current Drawbacks and Limitations

The primary issue facing LLMs.txt today is its low rate of adoption. Without widespread implementation, its effectiveness and value remain limited.

Beyond adoption, several additional issues have emerged:

- Generation and Maintenance - Creating and keeping LLMs.txt files current requires consistent effort, potentially deterring site owners.
- Potential for Abuse - With an easy-to-access summary of all site content, there could be concerns around content misuse.
- Standards and Validation - There currently isn't a universally agreed-upon standard or robust schema suitable for all website types, leading to possible inconsistencies and ambiguities.

How Many Websites Currently Use LLMs.txt?

Assessing current adoption rates reveals interesting disparities. Public directories tracking LLMs.txt implementations show significantly differing numbers:

- llmstxt.site lists roughly 170 sites.
- directory.llmstxt.cloud claims around 684 sites.

If they're the sum-total of all websites using LLMs.txt right now, that number IS VERY small. To investigate further, I crawled the Majestic Million - representing a large cross-section of popular websites - and **found only 0.015% (approximately 15 sites) had a valid LLMs.txt file** in Feb 2025.

Considering many sites in this dataset might not be actively maintained, a proxy measurement using robots.txt files indicated around 210,000 valid files - roughly one-fifth of the Majestic Million.

Using this as a baseline, **LLMs.txt adoption is around 0.007%** - hardly any better - highlighting just how limited uptake currently is.

What About the Rate of Adoption, is use of LLMs.txt Growing?

Yes - proportionally it is growing quickly. But in terms of raw numbers against the Majestic Million, it is effectively nothing.

May 2025 data suggests there are 105 valid LLMs.txt now found - at 600% increase, but still 0.011% of the total.*

Is Implementing LLMs.txt Worth it?

LLMs.txt won't harm anyone and, if deployed cheaply and easily, may help hedge your bets regarding future potential AI benefits.

However, the lack of widespread adoption could discourage major AI companies from officially adopting it, pushing them instead towards alternative solutions.

There is currently no clear return on investment if implementing LLMs.txt involves significant technical resources, nor an easy way to accurately measure its impact.

Do you get ahead “just in case” or sit back and see if it is worth the headache?

**A small caveat is needed here, looking into the various response codes about 100k have returned response codes such as "FORBIDDEN" & "I'M A TEAPOT". The method of crawling (user-agent) likely tripped a few security settings, which means there could be more valid files. I'm recrawling the error list with a new UA, just to confirm.*

SEO · AI