

Technical SEO for LLMs

Dan Taylor

Partner & Head of Innovation
SALT.agency

Google Search Central Live
Bangkok 2025

Dan Taylor

Partner & Head of Innovation (AI & SEO)

SALT.agency

TechSEO Boost Research Prize Winner 2018

15+ years in marketing / ~11 in agency

<https://salt.agency>



Reformation



Technical SEO, for LLMs.

Search engines aren't the only **crawlers** now.
Client's care about appearing in this AI stuff.

This means technical is no longer **just about**
“SEO”, it's technical optimisation for crawlers of
various origin.

Definitions:

LLMs = Gemini, ChatGPT, etc.

Google Search = Traditional, AI Mode, AIOs

Y.M.M.V.

JavaScript

Most LLMs do not render JavaScript content.

They fetch pages via simple HTTP requests and read the raw HTML, without executing scripts or waiting for dynamic content to load.

These tools act more like text-based scrapers (akin to `curl` or a basic HTTP client) than full browsers.

Renders JS

Gemini

(AI Overviews & AI Mode)

Doesn't Render JS

Perplexity

Claude

Inconclusive

ChatGPT

<> Server-side rendering (SSR)

<> Static pre-rendering

<> Hydration

<> Compressed plain text views for specific UAs*

*you could probably implement this easily via your CDN... #EdgeSEO

Schema

Can LLMs understand Schema?

Does schema impact visibility?

PRETRAINING
Step 1: download and preprocess the internet
<https://huggingface.co/spaces/HuggingFaceFW/blogpost-finetweb-v1>

The FinetWeb pipeline

Blacklists UT1
For english version

Context

For several years, Toulouse Capitole University has been distributing a blacklist of URLs, managed by Fabrice Prigent, in order to allow better control of Internet use. This database, widely used by schools, can be integrated into a large number of free or commercial tools, in addition to other lists.

Over time the number of categories has increased significantly, and their role has changed. The list we provide should be seen as a categorization of sites, rather than a list to block. For example

- child is a so-called whitelist category for young children
- little_bu is a category to open for universities
- press can be seen as something to prohibit (at work for example) or to authorize (CIO, Library).

Remember that you can (must?) use a "local" category that would allow you to adapt the behavior of your filtering. Thus blogs can be banned in general, but some must be authorized for educational reasons.

The license to use

This creation is made available under a [Creative Commons contract](#).

Description

Several categories are defined, but it is mainly the pornographic one that is actively maintained.

For those who want a "working" base, I only regularly update the **adult** base that is most useful to us. If some of you decide to complete the other bases, I would be happy to integrate them into the update.

I am currently adding between 50 and 300 urls per day (some urls need to be checked); I can only slowly absorb the additions made by my crawler. Please excuse me.

There is a file that groups all the categories: [blacklists.tar.gz](#) This is the compilation of the bases that are described below:

Category	Number	Description
adult	4578214	Adult sites ranging from erotica to hardcore pornography.
aggressive	394	Some racist, anti-Semitic, hate-inciting sites.
arjel	69	Online betting sites certified by ARJEL.
religious_associations	1	Religious association sites
astrology	29	Astrology
audio-video	3866	Some sites oriented towards audio and video.
bank	1896	Online banking
bitcoin	336	Bitcoin sites
blog	1485	Some sites hosting blogs.
celebrity	673	Everything related to so-called celebrity news
cat	264	Online chat and dialogue site.
child	75	Anything that is allowed for children
cleaning	177	Sites to disinfect and update computers.

<https://www.youtube.com/watch?v=7xTGNNLPyMI>

Deep Dive into LLMs like ChatGPT

Andrej Karpathy - Co-founder Open AI, ex-Sr. Director of AI at Tesla

“When we go to inspect we can see all this markup, like lists, CSS, and all this kind of stuff. Computer code for these webpages. We don't want this. We want just the content of the webpages.”

BUT

LLMs can use structured data when pre-processing tools include it, but often they only receive plain text or markdown.

Schema Type	% of AI Overviews containing it
Organization	82%
WebPage / Article	76%
BreadcrumbList	59%
FAQPage / QAPage	41%
Person (Author)	38%
Product / Service	34%
ImageObject	28%
Review / AggregateRating	19%

Sample Size: 107,352 websites appearing as Citations

WebPage → mainEntity → Article

Article → author → Person

Organization → url → Homepage

LLMS & AI tools **don't need Schema** to be able to extract and understand content on a page.

Google uses Schema in the same way it always has done when it comes to AI Mode and AI Overviews.*

*because AI Mode and AI Overviews are extensions of Search on the same infrastructure.

Robots.txt

Even if you're not disallowing LLM crawlers, you could still be included in **training data**.

ccbot

[The Data](#) ▾[Resources](#) ▾[Community](#) ▾[About](#) ▾[Search](#) ▾[Contact Us](#)

CCBot

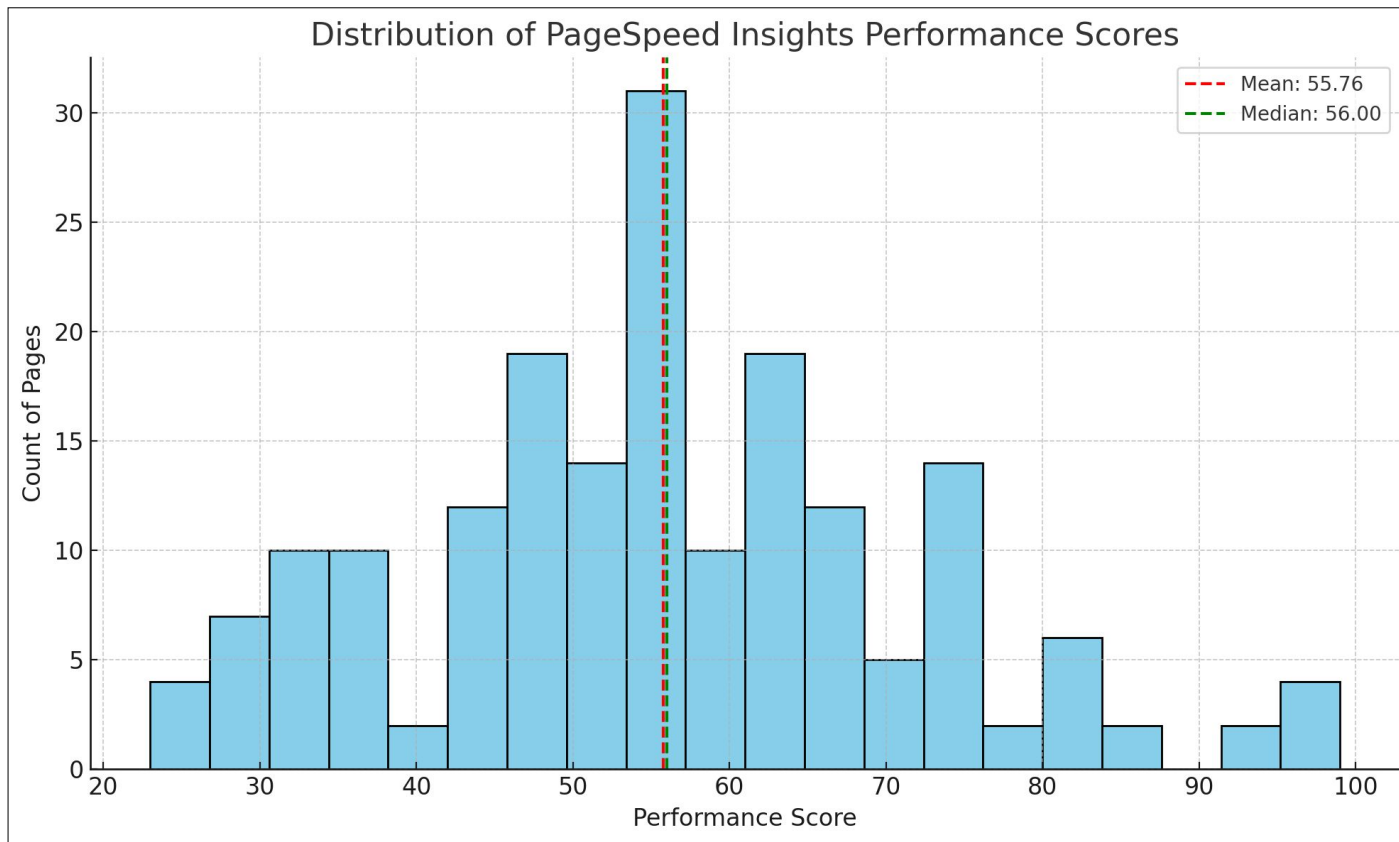
Common Crawl is a non-profit foundation founded with the goal of democratizing access to web information by producing and maintaining an open repository of web crawl data that is universally accessible and analyzable by anyone.

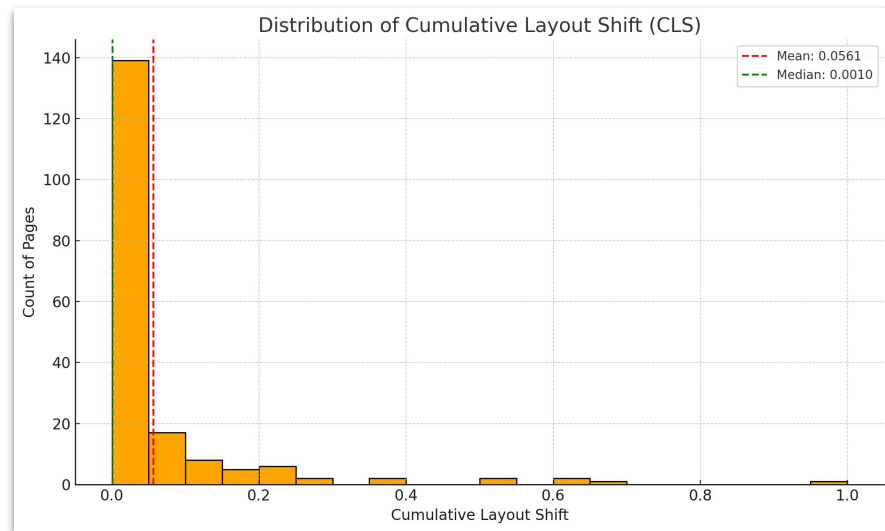
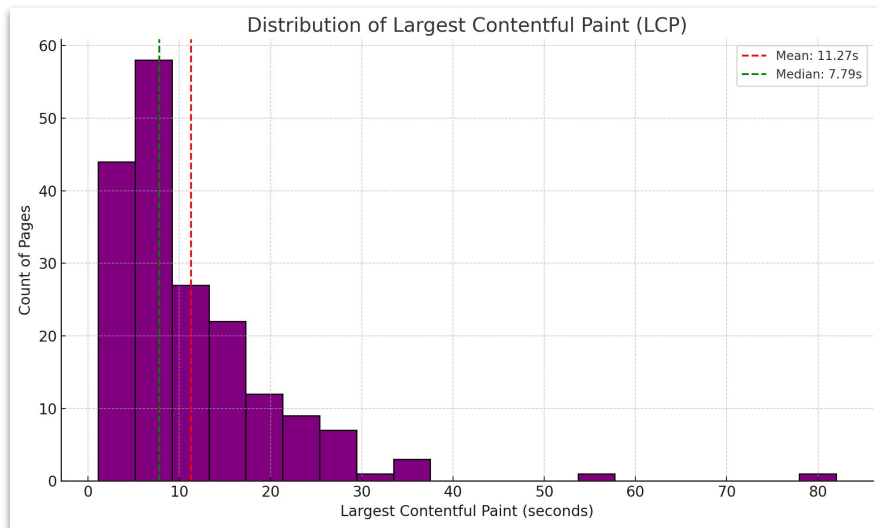
Enabling free access to web crawl data encourages collaboration and interdisciplinary research, as organizations, academia, and non-profits can work together to address complex challenges. Collaborating using Open Data accelerates progress and helps find solutions to pressing global issues, such as climate change, public health, and social equality.



<https://commoncrawl.org/ccbot>

Site Speed





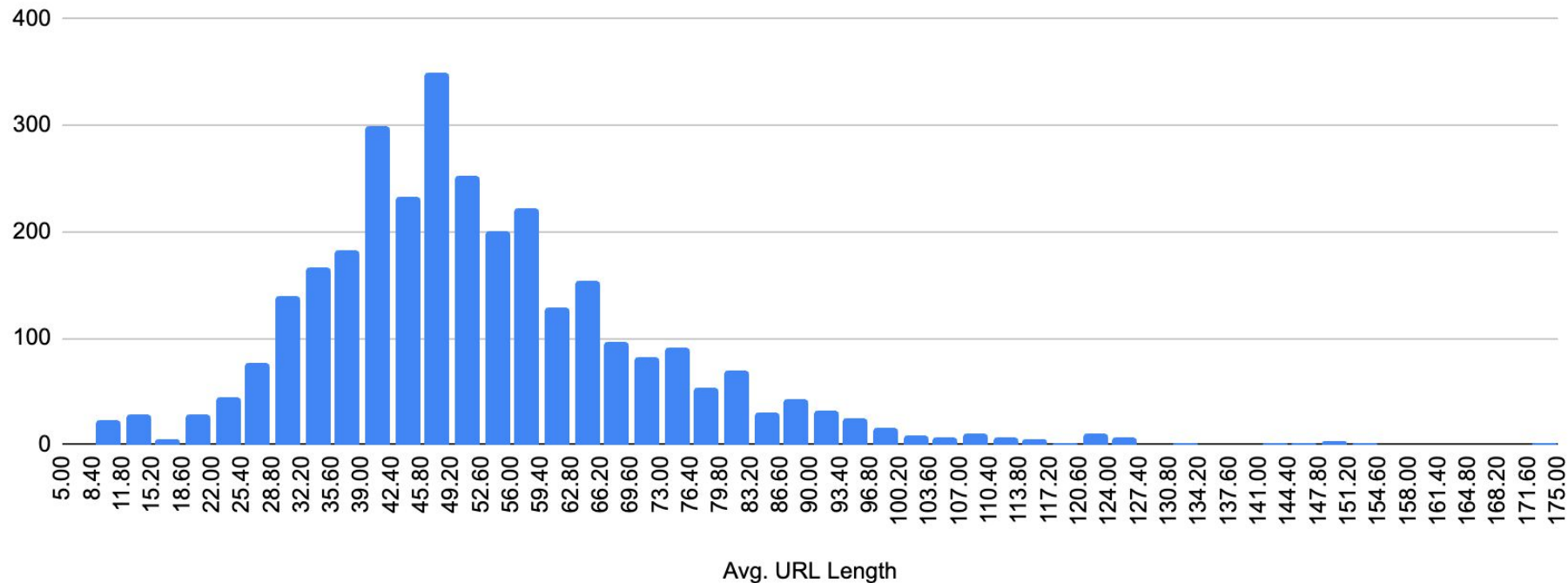
Good PSI scores **don't** correlate with inclusion.

Good CWV scores **do** correlate with inclusion.

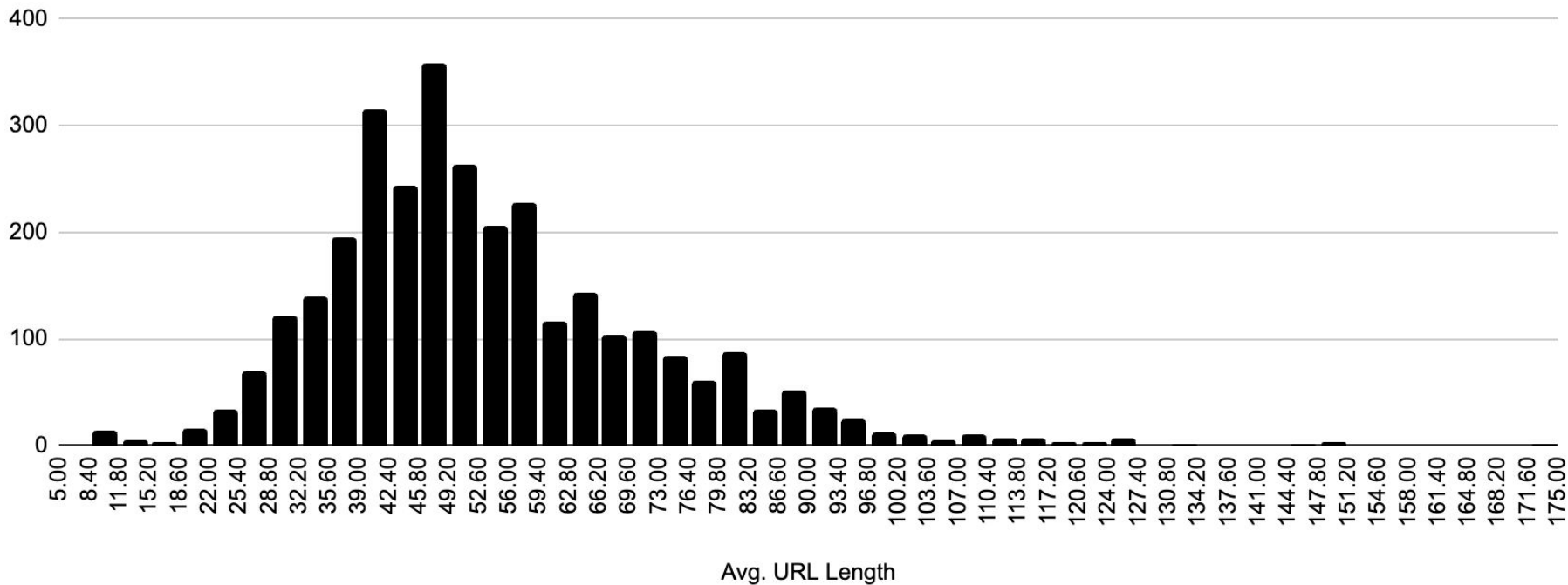
Site Structure

URL Length

Average URL Length (Characters - Inc. "/") - Classic Blue Links



Average URL Length (Characters - Inc. "/") - AI Mode / AI Overviews



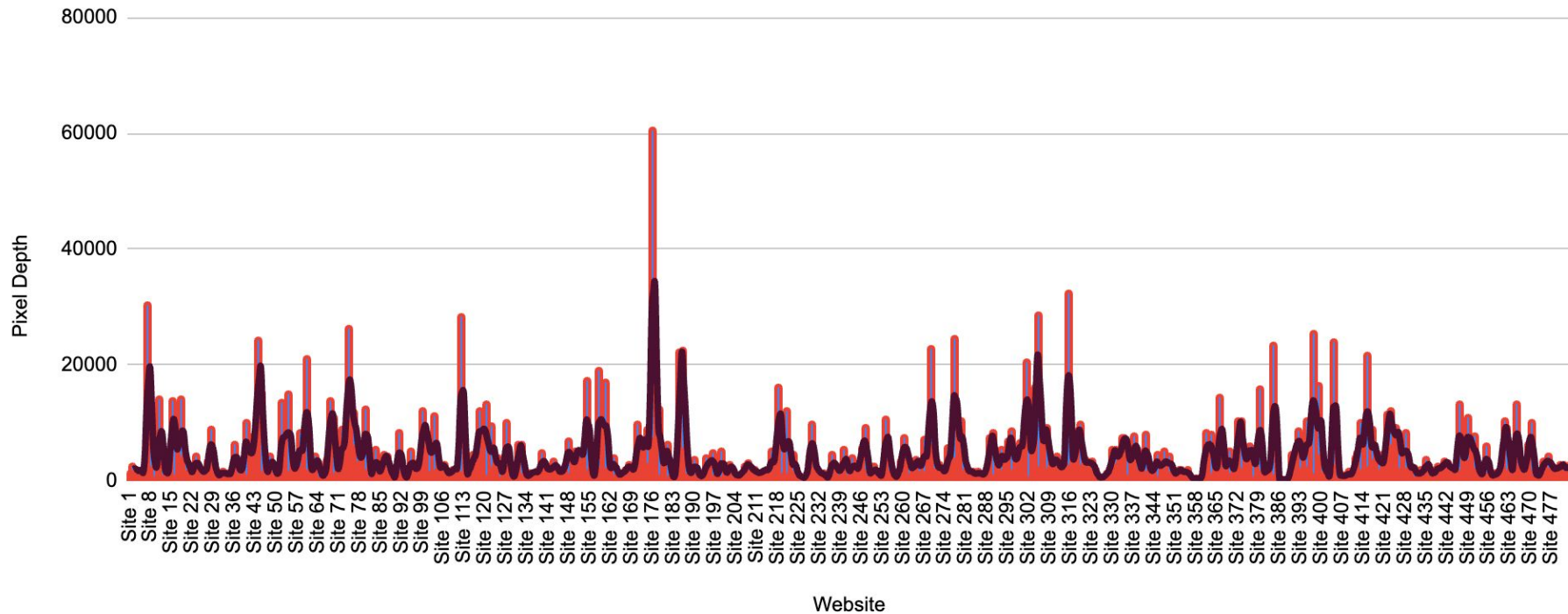
	Classic Blue Links	AI Mode / Overviews
Min	9	9
Max	173	173
Mean	52.24425287	53.43390805



Page Structure

If you structure a page in **XYZ way**, you'll increase your chances of being pulled into generative summaries.*

#::~:text



Sample Size: 2,138 websites appearing as Citations

	Mean	Median	Min	Max
SaaS	4,607px	2,364px	215px	60,565px
Travel	2,406px	1,249px	100px	31,878px
eCommerce	3,610px	2,005px	65px	42,026px

Sample Size: 2,138 websites appearing as Citations

Wrapping things up

The foundations of technical SEO are the same.

The internet, for now, still has the same
structure.

We've optimised websites for Search Engine crawlers, now we need to understand and accommodate the nuances of LLM crawlers.

JavaScript is once again a key area of
optimisation.

TECH SEO

Thank you

dan@salt.agency

<https://dantaylor.online>

<https://salt.agency>

<https://salt.agency/call-dan>

