

What LLMs Really See When They Read Your Content

And How to Structure It For The Future



Introduction

If you've only just wrapped your head around the concept of "large language models" (LLMs), you're not alone. Generative engines may have been around for 2-3 years now, but how AI works under-the-hood and how it impacts search are just starting to become, maybe not common knowledge yet, but certainly more researched and talked about.



WHY THIS MATTERS FOR BRANDS:

Control: How your content is interpreted, understood, and surfaced is increasingly determined by AI engines, not just search engines or social platforms.

Visibility: AI-driven content discovery may favor certain formats, sources, or even phrases, changing who audiences see first when they type in a prompt.

Trust: Users have started to rely on AI recommendations and summaries to help them make faster decisions; if your content isn't up-to-date and regularly refreshed with new insights then it risks losing credibility.

WHY THIS MATTERS FOR MARKETERS:

If you're a marketer today, you need to understand how generative engines work, how they find their information, and how to structure your content so that it gets "cited" in an answer. You might read a blog here, or send a prompt there, or maybe you're frequently skimming think-pieces on social media from your favourite industry influencer. But you're still a bit confused on what to do, and it's all feeling more technical than you're used to. That's okay, we're here to help.

In this handy PDF we've condensed a really big topic into three sections. And added a bunch of examples to really help you understand. We're giving away three big nuggets:

- What LLMs "see" when they "read" your content
- How to structure your content for LLMs as a result
- A checklist of technical optimizations to discuss with your developer team

Let's go.

Table of Contents



4

PART 1:

WHAT DO LLMS “SEE” WHEN THEY “READ” YOUR CONTENT?

7

PART 2:

HOW TO STRUCTURE YOUR CONTENT FOR LLMS

9

PART 3:

YOUR TECHNICAL GEO CHECKLIST

PART 1:

What do LLMs “see” when they “read” your content?

In a nutshell: LLMs don’t read text like humans do. They don’t scan sentences for context and meaning like our brains do as we take in the words. Instead, they break text into small pieces called “tokens”, then predict what comes next based on patterns learned from vast amounts of data. It’s all mathematical.

For marketers, it’s important to understand how that relates to any piece of content you write for the web. Let’s take a standard blog post and a really simple example of copy.

Here’s an example of an LLM breaking down your blog content into tokens (“tokenization”):

So, you’ve written a blog post. In your CMS you’ve written—

“I love apples!”

If you’re familiar with HTML, the code in the frontend for this might typically look like—

```
<p class="intro">I love apples!</p>
```

This copy would then be separated into its components, this might look like (simplified!)—

[I, love, apples, !]

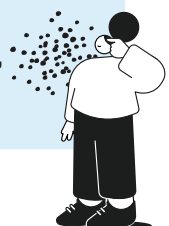
The long version with HTML included—

```
[“<”, “p”, “ class=”, “”, “intro”, “”, “>”, “I”, “love”, “”, “ apples”, “!”, “</”, “p”, “>”]
```

Do you see how the words and the code become a sequence of symbols? LLMs don’t inherently understand that <p> means paragraph, it’s just another symbol. This sequence of symbols looks different depending on the LLM, as they break down the components differently.

Finally, the symbols are converted into numbers, with each symbol given a unique numerical ID (token IDs.)

LLMs will begin to understand meaning by parsing vast amounts of these numbers and recognizing any patterns. Taking the earlier example, they will understand that a pattern emerges where words included within <p> are separated into paragraphs, therefore <p> means paragraph.



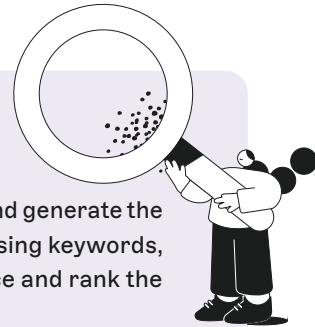
To summarize: LLMs don't "see" a webpage visually. They mostly see it as a structured text document (HTML and text content), which they break into tokens and analyze mathematically to understand patterns and meaning.

It's important then to understand what this means for search visibility. Let's compare traditional search rankings to appearing in AI engine results.

SEARCH RANKING VS. AI ANSWERS

Search ranking: Search engines take keyword queries and generate the organic results based on their relevance to that query, using keywords, backlinks, and metadata of content to measure relevance and rank the results, listing the exact URLs.

AI answers: LLMs take the prompt or question asked and synthesize answers by pulling patterns and context from many sources, often summarizing those many sources into one response rather than pointing directly to a single page. They may copy content directly or cite (link) to a source, which is the optimal result for brands, but it doesn't always happen.



AI engines have developed faster than we could've imagined since 2023, and now pull their information from multiple sources to remain as up-to-date as possible. The model no longer relies solely on the information it learned during training, it pulls information straight from the web. It "reads" that information and understands it in the blink of an eye, generating an answer in seconds.

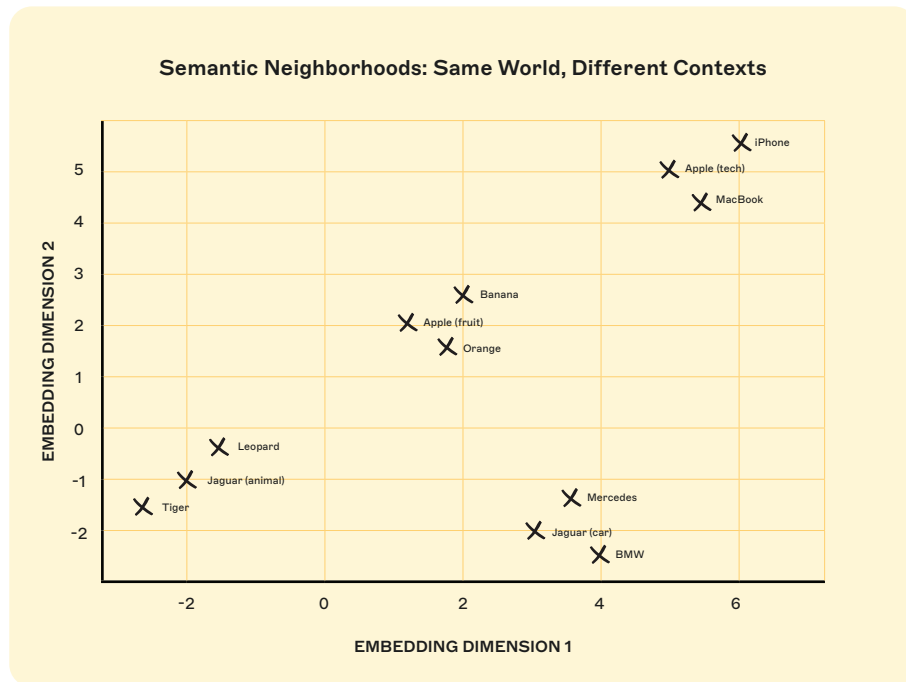
There are some more terms that will help deepen your understanding of how an LLM provides you with an answer — "RAG", "NLP", "Semantic Search", and "Embeddings":

Retrieval-Augmented Generation (RAG): Combines external sources with the LLM and its training data, meaning it can "look up" documents from other sources (like the internet, press releases, company policies) before generating an answer. (Open [this article](#) about RAG in a new tab right now!)

Natural Language Processing (NLP): Is a field associated with artificial intelligence that allows computers to understand human language, both spoken and written; it's what voice assistants, chatbots, and other programmes have used to understand humans and take action or reply to us, long before LLMs were in the picture.

Semantic Search: Is different from "keyword search", typically associated with SEO. Another field that's been around long before LLMs; semantic search analyzes the query or prompt's intent, context, and the overall meaning to provide the most contextually relevant information.

Embeddings: Is the general term for the process of converting data of any type into vectors. The “tokens” we mentioned earlier are an example of data that can be converted into numerical vectors, and these vectors represent context and semantic meaning. Here’s a straightforward example — so “jaguar” would sit near “car” in one context in a vector database, or near “animal” in another. One word, two different contexts, two different meanings.



PART 2:

How to structure your content for LLMs

Now we know what LLMs “see” when they “read” content, the next step is to structure content in a way that makes it easier and faster for LLMs to retrieve, tokenize, and understand.

SEO VS GEO

The good news is that a lot of traditional SEO tactics also work for GEO (generative engine optimization.) Meaning you can feed two birds with one scone by structuring your content effectively!

Text vs markup vs structured data

Text = the content on the page.

Markup = tells the browser how the text should look.

Structured data = tells bots (like Google) what the data is.

How to get started with structured content:

- Audit your existing content: find patterns in content types (e.g., product descriptions, blog posts, event pages, recipes).
- Choose a headless CMS that supports [content modeling](#) (e.g., Storyblok). You can also choose a traditional/monolithic CMS but they typically don't support structured data by design, and are more restricted when it comes to content modeling.
- Break down big content blocks into reusable fields/components.
- Start small: pick one type of content (like “articles” or “events”) and model it first.

Schema markup explained

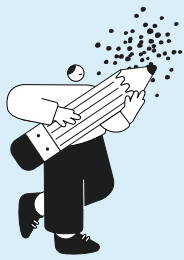
- **Definition:** A type of **structured data** (JSON-LD) that you add to the HTML of a webpage so search engines can better understand what the page is about.
- **Example:**
 - Adding [Schema.org's Recipe markup](#) so Google can display a recipe with star ratings, cooking time, and a photo in search results.
 - Using [Product markup](#) to show price, stock status, and reviews in rich snippets.
- **Purpose:** Helps search engines interpret your content, leading to rich results in SERPs, voice search optimization, and better discoverability.

Schema markup still matters for GEO because it's good for SEO

Schema markup impacts SEO ranking, and SEO ranking can benefit your visibility in AI engines. Why? Because of RAG. When AI engines pull from the internet to serve you an answer to a query/prompt, search engine rankings and who's currently at the top are one of the external sources they can pull from to decide what content to quote or cite.

How to get started with schema markup:

- Use [Google's Structured Data Markup Helper](#) to generate markup.
- Start with the most valuable content types for your site: Article, Product, FAQ, HowTo, or LocalBusiness.
- Add JSON-LD code to your [page](#) <body> or via a CMS plugin/module, depending on your CMS — discuss implementation with your developer team.
- Test with [Google's Rich Results Test](#) and [Schema.org validator](#).
- Monitor using Google Search Console.



EXPERT TIPS FOR WRITING LLM-FRIENDLY CONTENT

- ① **Clear structure**
 - Include titles, headings, and bullet points.
 - Organize content so the main points stand out clearly, utilising FAQs and list-structures where appropriate (just like we've done here!)
- ② **Concise sentences**
 - Avoid long sentences or mixing multiple topics into one paragraph.
 - Be concise and digestible, rather than unstructured or rambling.
 - Avoid stories with no clear beginning, middle, or context.
- ③ **Keep learning**
 - Read our [GEO Explained in 5 Minutes](#) article for more advice on optimizing content for AI engines.

PART 3:

Your technical GEO checklist

If you're really invested in future-proofing your brand or website for AI search, then think of this checklist as your sidekick for approaching conversations with your developer team. Become the catalyst for change in your department.

Here's a handy technical generative engine optimization checklist for your website:

- ① **Structured data and metadata**
 - Implement JSON-LD schema markup.
 - Create multiple [schemas](#) for each content type on your site (e.g., Product, Event, Article, FAQ, etc.) and ensure that each includes all relevant and required fields for that type.
- ② **Structured content**
 - Ensure semantic HTML markup including proper heading hierarchy and image descriptions.
- ③ **[llms.txt file](#)**
 - A markdown file that includes key content is available in a [LLM-friendly format](#)
- ④ **Server-side rendering over client-side rendering.**
 - Ensure your site [loads as quickly as possible](#) and all content is rendered on the server.
- ⑤ **Use a headless CMS to ensure GEO-friendly content implementation**
 - In this context, a headless CMS offers a significant [advantage over a monolithic CMS](#): it decouples presentation from data, supporting structured data by design.
 - The composable architecture of headless means you can display the data on your website in a way that works perfectly for both humans and AIs — display it on the frontend for humans while adding JSON-LD schema and generating an llms.txt file for AI agents!

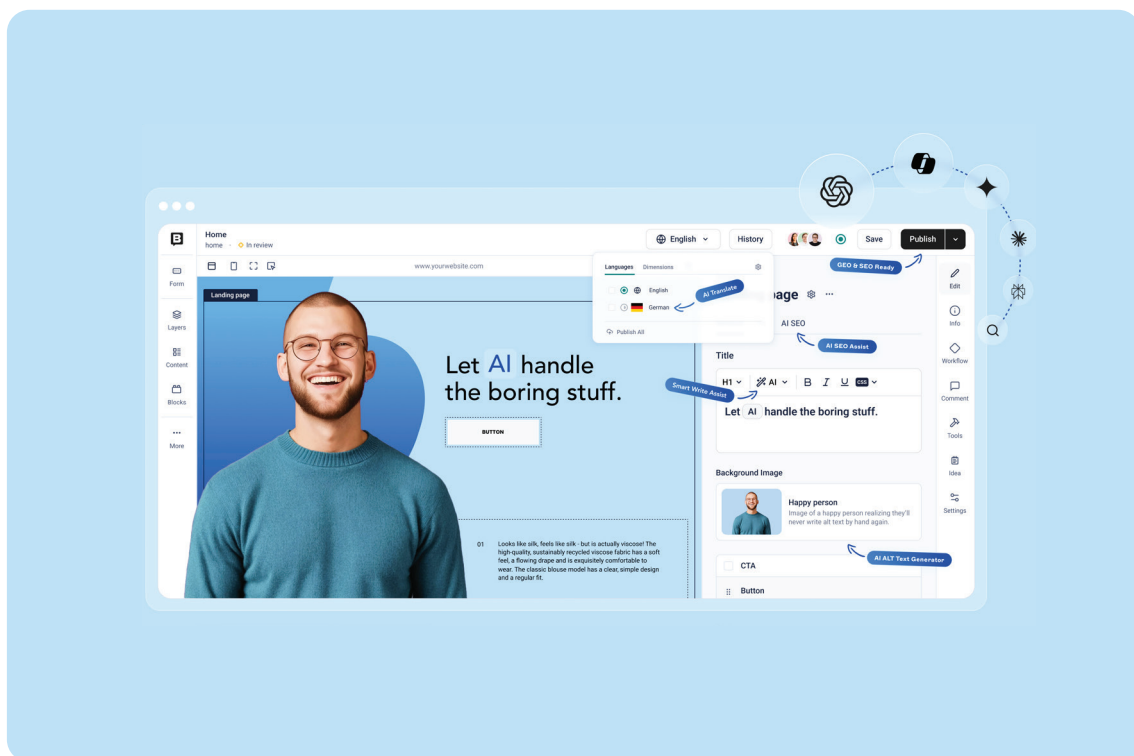
The bottom line

Early adoption of these tactics ensures visibility, control, and influence in the AI-driven content era. Good luck!

Storyblok is a headless CMS that enables marketers and developers to create with joy and succeed in the AI-driven content era. It empowers you to deliver structured and consistent content everywhere: websites, apps, AI search, and beyond.

Marketers get a visual editor with reusable components, in-context preview, and workflows to launch fast and stay on brand. Developers have freedom to use their favorite frameworks and integrate with anything through the API-first platform. Brands get one source of truth for content that is accurate, flexible, and measurable.

Legendary brands like Virgin Media O2, Oatly, and TomTom use Storyblok to make a bigger, faster market impact. It's Joyful Headless™, and it changes everything.



Get Joyful.

**See how Storyblok can revolutionize
your content at storyblok.com.**

CONTACT US 

TRY IT FOR FREE 