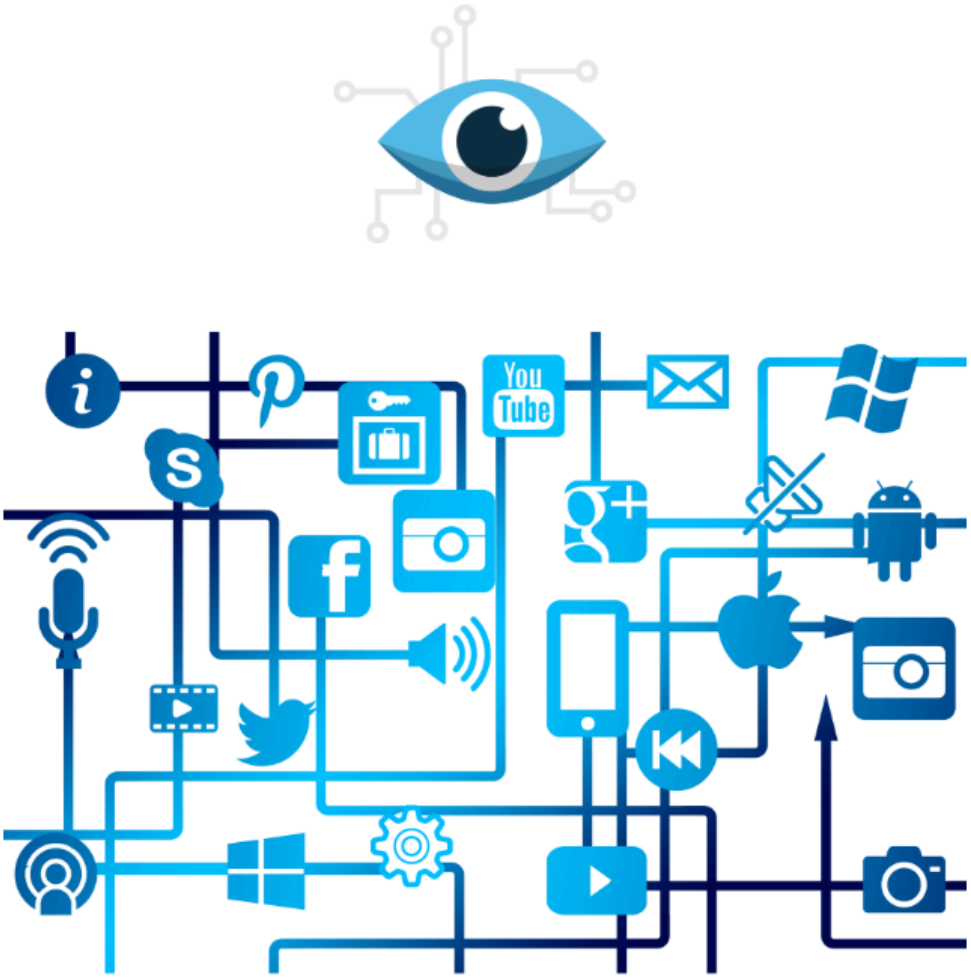
A Comprehensive Guide to AI-Powered Search: How GPTs Access Web Content and What Content Creators Must Do to Stay Visible



Bob Mazzei

2025

Sienda ltd

Cover image made by the author using Canva.

Table of Contents

1. Introduction

The rise of AI-powered search tools like OpenAI's ChatGPT, Google's AI Overviews, and Microsoft's Copilot has reshaped how users discover and consume online information. Unlike traditional search engines that retrieve indexed web pages, AI models generate responses based on vast amounts of processed data, often presenting summarised or restructured information.

For content creators, this shift presents both opportunities and challenges. Businesses must now optimise their content for AI retrieval, ensuring visibility in AI-generated responses while also protecting proprietary content where necessary. This guide explores how AI models access web content and what steps content creators must take to adapt.

2. How AI-Powered Search Works

The Difference Between Traditional Search Engines and GPTs

Aspect	Traditional Search Engines (Google, Bing, etc.)	AI-Powered Search (GPTs, AI Overviews, etc.)
Content Retrieval	Crawls and indexes web pages to retrieve relevant links	Generates responses based on pre-trained data or real-time API calls
Presentation of Information	Displays ranked search results for users to click	Summarises, restructures, and presents information directly
Data Updating	Continuously updated by web crawlers	Typically based on static training data unless real-time browsing is enabled
User Interaction	Users explore multiple sources	AI presents a single, conversational response
Source Attribution	Directs users to original sources	May summarise without linking back, depending on implementation

Data Sources Used by GPTs

GPT models do **not** actively crawl the web like Google. Instead, they rely on:

1. **Pre-Trained Datasets** – These may include:
    - Publicly available web content (up to the latest training date)
    - Licensed databases (e.g., scientific journals, news archives)
    - Open-source repositories and forums (e.g., Wikipedia, GitHub)
2. **Real-Time Web Browsing (if enabled)**
    - Some AI tools can fetch **real-time search results** via integrated web browsing features.
    - These queries often rely on **search engines like Bing** rather than crawling sites directly.
3. **APIs and Structured Data**
    - Websites and businesses can **feed AI models structured data** via APIs, allowing for real-time content updates.
4. **User-Provided Inputs**
    - When users copy-paste articles, upload PDFs, or ask AI to summarise content, AI can generate responses based on this input.

3. How AI Models Retrieve and Use Web Content

1. GPTs with Web Browsing Capabilities

Some AI models, like ChatGPT with the "Browse with Bing" feature, can **fetch live search results**. However, these models **do not directly crawl websites** but instead retrieve and summarise content from search engine queries.

How This Works

➔ The user enters a query (e.g., "latest AI SEO strategies 2025").
➔ AI sends the query to Bing (or another integrated search engine).
➔ The search engine returns top-ranked pages.
➔ AI **summarises key points** rather than providing a list of links.

2. API Integrations and AI Content Indexing

Some AI tools allow **direct API integrations**, enabling businesses to:

➔ Provide real-time product catalogues to AI-powered assistants.
➔ Deliver structured data in a machine-readable format.

➔ Maintain control over how their content is accessed and used.

Example
A food wholesaler integrates an API with OpenAI to enable AI-driven order recommendations based on real-time inventory.

3. Structured Data and Schema Markup

AI prioritises **structured data**, meaning websites using Schema.org markup (e.g., FAQ schema, HowTo schema) are more likely to be surfaced in AI-generated responses.

Key Schema Markup Types for AI SEO

➔ **FAQ Schema** – Helps AI understand and present Q&A-style content.
➔ **HowTo Schema** – Useful for instructional content and step-by-step guides.
➔ **Product Schema** – Enables AI to retrieve detailed product descriptions and specifications.

4. Best Practices for Content Creators to Appear in AI-Powered Search

1. Optimising for AI-Driven Responses

➔ **Use concise, factual, and well-structured language.**
➔ **Provide clear summaries at the start of articles.**
➔ **Optimise headings and bullet points for easy AI parsing.**

2. Leveraging SEO and Structured Data

➔ Implement **Schema.org markup** for structured data.
➔ Write **FAQ-style content**, which AI models favour.
➔ Use **listicles and step-by-step guides**, as AI prefers structured content.

3. Building AI-Optimised APIs and Feeds

➔ Create REST APIs to serve structured, machine-readable data.
➔ Offer real-time updates via RSS feeds or direct API integration with AI providers.

5. Protecting Content from AI Crawlers

If you **do not want AI to access your content**, use the following methods:

1. Robots.txt for AI Models

To block AI crawlers like OpenAI's GPTBot, add this to your robots.txt:

User-agent: GPTBot

Disallow: /

2. Meta Tags to Prevent AI Indexing

Add this to your website's <head> section:

<meta name="robots" content="noai, noindex">

3. Authentication and Paywalls

Restrict access to sensitive content by requiring **user authentication.**

6. The Future of AI-Driven Search

- ➔ More AI-integrated search engines (Google's AI Overviews, Microsoft's Copilot).
- ➔ Greater reliance on structured data and API-based content feeding.
- ➔ Ethical concerns and regulations on AI training and content usage.

Businesses will need to adapt their SEO strategies to remain competitive in AI-driven search while also safeguarding proprietary content.

7. Conclusion

AI-powered search is fundamentally changing how users discover and engage with content. Unlike traditional search engines, AI models prioritise structured, summarised, and conversational content. Content creators who wish to stay relevant must:

1. **Implement structured data and schema markup.**
2. **Optimise content for AI-generated responses.**
3. **Leverage APIs and feeds to integrate directly with AI models.**
4. **Protect their proprietary content where necessary.**

By adapting to these shifts, businesses and creators can enhance their visibility in the evolving AI-driven digital ecosystem.

Thank you for downloading and taking the time to read this guide!

Bob Mazzei, 2025

Sienda ltd

www.sienda.co.uk

www.siendamedia.com