

Perplexity AI Interview Explains How AI Search Works via @sejournal, @martinibuster



I recently spoke with Jesse Dwyer of Perplexity about SEO and AI search about what SEOs should be focusing on in terms of optimizing for AI search. His answers offered useful feedback about what publishers and SEOs should be focusing on right now.

AI Search Today

An important takeaway that Jesse shared is that personalization is completely changing

“I’d have to say the biggest/simplest thing to remember about AEO vs SEO is it’s no longer a zero sum game. Two people with the same query can get a different answer on commercial search, if the AI tool they’re using loads personal memory

into the context window (Perplexity, ChatGPT).

A lot of this comes down to the technology of the index (why there actually is a difference between GEO and AEO). But yes, it is currently accurate to say (most) traditional SEO best practices still apply.”

The takeaway from Dwyer’s response is that search visibility is no longer about a single consistent search result. Personal context as a role in AI answers means that two users can receive significantly different answers to the same query with possibly different underlying content sources.

While the underlying infrastructure is still a classic search index, SEO still plays a role in determining whether content is eligible to be retrieved at all. Perplexity AI is said to use a form of PageRank, which is a link-based method of determining the popularity and relevance of websites, so that provides a hint about some of what SEOs should be focusing on.

However, as you’ll see, what is retrieved is vastly different than in classic search.

I followed up with the following question:

So what you’re saying (and correct me if I’m wrong or slightly off) is that Classic Search tends to reliably show the same ten sites for a given query. But for AI search, because of the contextual nature of AI conversations, they’re more likely to provide a different answer for each user.

Jesse answered:

“That’s accurate yes.”

Sub-document Processing: Why AI Search Is Different

Jesse continued his answer by talking about what goes on behind the scenes to generate an answer in AI search.

He continued:

“As for the index technology, the biggest difference in AI search right now comes down to whole-document vs. “sub-document” processing.

Traditional search engines index at the whole document level. They look at a webpage, score it, and file it.

When you use an AI tool built on this architecture (like ChatGPT web search), it essentially performs a classic search, grabs the top 10–50 documents, then asks the LLM to generate a summary. That’s why GPT search gets described as “4 Bing searches in a trenchcoat” –the joke is directionally accurate, because the model is generating an output based on standard search results.

This is why we call the optimization strategy for this GEO (Generative Engine Optimization). That whole-document search is essentially still algorithmic search, not AI, since the data in the index is all the normal page scoring we’re used to in SEO. The AI-first approach is known as “sub-document processing.”

Instead of indexing whole pages, the engine indexes specific, granular snippets (not to be confused with what SEO’s know as “featured snippets”). A snippet, in AI parlance, is about 5-7 tokens, or 2-4 words, except the text has been converted into numbers, (by the fundamental AI process known as a “transformer”, which is the T in GPT). When you query a sub-document system, it doesn’t retrieve 50 documents; it retrieves about 130,000 tokens of the most relevant snippets

(about 26K snippets) to feed the AI.

Those numbers aren't precise, though. The actual number of snippets always equals a total number of tokens that matches the full capacity of the specific LLM's context window. (Currently they average about 130K tokens). The goal is to completely fill the AI model's context window with the most relevant information, because when you saturate that window, you leave the model no room to 'hallucinate' or make things up.

In other words, it stops being a creative generator and delivers a more accurate answer. This sub-document method is where the industry is moving, and why it is more accurate to be called AEO (Answer Engine Optimization).

Obviously this description is a bit of an oversimplification. But the personal context that makes each search no longer a universal result for every user is because the LLM can take everything it knows about the searcher and use that to help fill out the full context window. Which is a lot more info than a Google user profile.

The competitive differentiation of a company like Perplexity, or any other AI search company that moves to sub-document processing, takes place in the technology between the index and the 26K snippets. With techniques like modulating compute, query reformulation, and proprietary models that run across the index itself, we can get those snippets to be more relevant to the query, which is the biggest lever for getting a better, richer answer.

Btw, this is less relevant to SEO's, but this whole concept is also why Perplexity's search API is so legit. For devs building search into any product, the difference is night and day."

Dwyer contrasts two fundamentally different indexing and

retrieval approaches:

- Whole-document indexing, where pages are retrieved and ranked as complete units.
- Sub-document indexing, where meaning is stored and retrieved as granular fragments.

In the first version, AI sits on top of traditional search and summarizes ranked pages. In the second, the AI system retrieves fragments directly and never reasons over full documents at all.

He also described that answer quality is constrained by context-window saturation, that accuracy emerges from filling the model's entire context window with relevant fragments. When retrieval succeeds at saturating that window, the model has little capacity to invent facts or hallucinate.

Lastly, he says that "modulating compute, query reformulation, and proprietary models" is part of their secret sauce for retrieving snippets that are highly relevant to the search query.

Featured Image by Shutterstock/Summit Art Creations

<https://www.searchenginejournal.com/perplexity-ai-interview-explains-how-ai-search-works/565395/>