

AI assisted search-based research actually works now

21st April 2025

For the past two and a half years the feature I've most wanted from LLMs is the ability to take on search-based research tasks on my behalf. We saw the first glimpses of this back in early 2023, with Perplexity (first launched [December 2022](#), first prompt leak [in January 2023](#)) and then the GPT-4 powered Microsoft Bing (which launched/cratered spectacularly [in February 2023](#)). Since then a whole bunch of people have taken a swing at this problem, most notably [Google Gemini](#) and [ChatGPT Search](#).

Those 2023-era versions were promising but very disappointing. They had a strong tendency to hallucinate details that weren't present in the search results, to the point that you couldn't trust anything they told you.

In this first half of 2025 I think these systems have finally crossed the line into being genuinely useful.

- [Deep Research, from three different vendors](#)
- [o3 and o4-mini are really good at search](#)
- [Google and Anthropic need to catch up](#)
- [Lazily porting code to a new library version via search](#)
- [How does the economic model for the Web work now?](#)

Deep Research, from three different vendors

First came the **Deep Research** implementations—[Google Gemini](#) and [then OpenAI](#) and [then Perplexity](#) launched products with that name and they were all impressive: they could take a query, then churn away for several minutes assembling a lengthy report with dozens (sometimes hundreds) of citations. Gemini's version had a *huge* upgrade a few weeks ago when they [switched it to using Gemini 2.5 Pro](#), and I've had some outstanding results from it since then.

Waiting a few minutes for a 10+ page report isn't my ideal workflow for this kind of tool. I'm impatient, I want answers faster than that!

o3 and o4-mini are really good at search

Last week, OpenAI released [search-enabled o3 and o4-mini](#) through [ChatGPT](#). On the surface these look like the same idea as we've seen already: LLMs that have the option to call a search tool as part of replying to a prompt.

But there's one *very significant* difference: these models can run searches as part of the chain-of-thought reasoning process they use before producing their final answer.

This turns out to be a *huge* deal. I've been throwing all kinds of questions at ChatGPT (in o3 or o4-mini mode) and getting back genuinely useful answers grounded in search results. I haven't

spotted a hallucination yet, and unlike prior systems I rarely find myself shouting "no, don't search for *that!*" at the screen when I see what they're doing.

Here are four recent example transcripts:

- [Get me specs including VRAM for RTX 5090 and RTX PRO 6000—plus release dates and prices](#)
- [Find me a website tool that lets me paste a URL in and it gives me a word count and an estimated reading time](#)
- [Figure out what search engine ChatGPT is using for o3 and o4-mini](#)
- [Look up Cloudflare r2 pricing and use Python to figure out how much this \(screenshot of dashboard\) costs](#)

Talking to o3 feels like talking to a Deep Research tool in real-time, without having to wait for several minutes for it to produce an overly-verbose report.

My hunch is that doing this well requires a very strong reasoning model. Evaluating search results is hard, due to the need to wade through huge amounts of spam and deceptive information. The disappointing results from previous implementations usually came down to the Web being full of junk.

Maybe o3, o4-mini and Gemini 2.5 Pro are the first models to cross the gullibility-resistance threshold to the point that they can do this effectively?

Google and Anthropic need to catch up #

The user-facing [Google Gemini app](#) can search too, but it doesn't show me what it's searching for. As a result, I just don't trust it. This is a big missed opportunity since Google presumably have by far the best search index, so they really should be able to build a great version of this. And Google's AI assisted search on their regular search interface hallucinates *wildly* to the point that it's actively damaging their brand. I just checked and Google is still showing slop [for Encanto 2!](#)

Claude also finally [added web search](#) a month ago but it doesn't feel nearly as good. It's [using the Brave search index](#) which I don't think is as comprehensive as Bing or Gemini, and searches don't happen as part of that powerful reasoning flow.

Lazily porting code to a new library version via search #

The truly magic moment for me came [a few days ago](#).

My [Gemini image segmentation tool](#) was using the [@google/generative-ai](#) library which has been [loudly deprecated](#) in favor of the still in preview [Google Gen AI SDK @google/genai](#) library.

I did **not** feel like doing the work to upgrade. On a whim, I pasted [my full HTML code](#) (with inline JavaScript) into ChatGPT o4-mini-high and prompted:

```
This code needs to be upgraded to the new recommended JavaScript library from Google. Figure out what that is and then look up enough documentation to port this
```

code to it.

(I couldn't even be bothered to look up the name of the new library myself!)

... it did [exactly that](#). It churned away thinking for 21 seconds, ran a bunch of searches, figured out the new library (which existed way outside of its training cut-off date), found the [upgrade instructions](#) and produced [a new version](#) of my code that worked perfectly.

Thought for 21 seconds ▾

🌐 Searched the web

Q Google Gemini API JavaScript library recommended new library

Google AI for Developers

GitHub

Google for Developers

👉 3 more

- The user provided an HTML+JavaScript code using the Gemini API and wants to upgrade the library. The code imports the previous GoogleGenerativeAI library, but Google may now recommend a new one — likely @google-ai/generative. I'll need to confirm what the latest library is and verify if it's the one the user should switch to. I'll search online for the term "Google Gemini API JavaScript library" or something similar to find the right information.

🌐 Searched the web

Google AI for Developers

•

🌐 Searched the web

Google AI for Developers

- We need to adjust the code for the new Google Gen AI SDK. First, update the import to the new CDN version for `GoogleGenAI` :

```
import { GoogleGenAI } from 'https://cdn.jsdelivr.net/npm/@google/genai@latest'
```

Then, we remove the old `getGenerativeModel` and instantiate `GoogleGenAI` directly:

```
const ai = new GoogleGenAI({ apiKey: getApiKey() });
```

I ran this prompt on my phone out of idle curiosity while I was doing something else. I was *extremely* impressed and surprised when it did exactly what I needed.

How does the economic model for the Web work now?

I'm writing about this today because it's been one of my "can LLMs do this reliably yet?" questions for over two years now. I think they've just crossed the line into being useful as research assistants, without feeling the need to check *everything* they say with a fine-tooth comb.

I still don't trust them not to make mistakes, but I think I might trust them enough that I'll skip my own fact-checking for lower-stakes tasks.

This also means that a bunch of the potential dark futures we've been predicting for the last couple of years are a whole lot more likely to become true. Why visit websites if you can get your answers directly from the chatbot instead?

The lawsuits over this [started flying](#) back when the LLMs were still mostly rubbish. The stakes are a lot higher now that they're actually good at it!

I can feel my usage of Google search taking a nosedive already. I expect a bumpy ride as a new economic model for the Web lurches into view.

Posted [21st April 2025](#) at 12:57 pm · Follow me on [Mastodon](#), [Bluesky](#), [Twitter](#) or [subscribe to my newsletter](#)

More recent articles

- [Maybe Meta's Llama claims to be open source because of the EU AI act](#) - 19th April 2025
- [Image segmentation using Gemini 2.5](#) - 18th April 2025

Part of series [How I use LLMs and ChatGPT](#)

- 25. [Building Python tools with a one-shot prompt using uv run and Claude Projects](#) - Dec. 19, 2024, 7 a.m.
- 26. [Here's how I use LLMs to help me write code](#) - March 11, 2025, 2:09 p.m.
- 27. [Not all AI-assisted programming is vibe coding \(but vibe coding rocks\)](#) - March 19, 2025, 5:57 p.m.
- 28. **AI assisted search-based research actually works now** - April 21, 2025, 12:57 p.m.

google359

search88

ai1229

openai282

generative-ai1056

chatgpt145

llms1046

anthropic135

perplexity5

gemini80

llm-tool-use29

llm-reasoning42

o314

deep-research5

ai-ethics152

ai-assisted-search26

Previous: [Maybe Meta's Llama claims to be open source because of the EU AI act](#)