These experts were stunned by OpenAI Deep Research

"I would use this model professionally," an antitrust lawyer told me.

TIMOTHY B. LEE
FEB 24, 2025 · PAID

42 3 10 Sha

Are you a journalist who wants to cover AI? Or an AI expert looking to get into journalism ***The deadline for the 2025 Tarbell Fellowship is this Friday.***

Understanding AI is a participating publication, which means you could get paid $50,000 t write for this newsletter. You can click here for details on what I'm looking for, or go directl the Tarbell website to apply.

Earlier this month, OpenAI released a new product called Deep Research. Based on variant of the (still unreleased) o3 reasoning model, Deep Research can think for ev longer than conventional reasoning models—up to 30 minutes for the hardest questions. And crucially, it can search the web, allowing it to gather information ab topics that are too new or obscure to be well covered in its training data.

I wanted to thoroughly test Deep Research out, so I solicited difficult questions fro random sample of Understanding AI readers. One of them was Rick Wolnitzek, a retired architect who runs the website Architekwiki. Wolnitzek asked for a detailed building code checklist for a 100,000-square-foot educational building.

Code Council, but some of the information it needed was behind a paywall.

“Considering a non-ICC site, perhaps from a state, might be a good move,” the moc thought. [1]

Deep Research soon found a page on the Arkansas Department of Education websi that included a three-page PDF of ICC standards for educational institutions. On tl website of Douglas County, Nevada, it found a PDF describing the minimum numb of plumbing fixtures required for various kinds of buildings. A California Departme of Education page summarized the number of toilets required in K-12 schools. The of Chelan, Washington, had a 13-page PDF summarizing recent code changes.

In total, OpenAI’s Deep Research model thought for 28 minutes and consulted 21 online sources to produce a 15,000-word checklist.

The report impressed Wolnitzek. It was “better than intern work, and meets the lev of an experienced professional,” he told me. “I think it would take six to eight hours more to prepare a report like this, and it would be a useful reference for the whole design team.”

Wolnitzek was one of 19 Understanding AI readers—including an antitrust lawyer, middle school teacher, a mechanical engineer, and a medical researcher—who help me put Deep Research through its paces. Not everyone was as impressed with OpenAI’s responses as Wolnitzek. But seven out of 19 respondents—including Wolnitzek—said OpenAI’s response was at or near the level of an experienced professional in their fields. A majority of respondents estimated it would take at lea 10 hours of human labor to produce a comparable report.

I see these results as hugely significant. It’s not just that Deep Research is likely to useful across a wide range of industries. Its performance demonstrates the impress capabilities of the underlying o3 model.

Deep Research discovers information in the same iterative manner as human researchers. It will do a search, find a document, and read it. Then it will do anothe search, find another document, and read that one. As the model reads more docume and learns more about a topic, it is able to refine its search criteria and find docume that didn't appear in earlier search results. This process—which people sometimes describe as "going down a rabbit hole"—allows Deep Research to gain a much deep understanding of a subject than was possible with previous AI models.

All of this is made possible by the longer "attention span" of o3, OpenAI's most powerful reasoning model. We've known for three years that large language models produce better results when they're asked to "think step by step." But conventional LLMs tended to get confused or distracted when they tried to perform a long seque of reasoning steps.

OpenAI used a technique called reinforcement learning to train reasoning models t stay focused as they work through longer chains of reasoning. This approach worke particularly well for domains like math and computer programming where the training algorithm could easily verify whether the model had reached a correct ansv

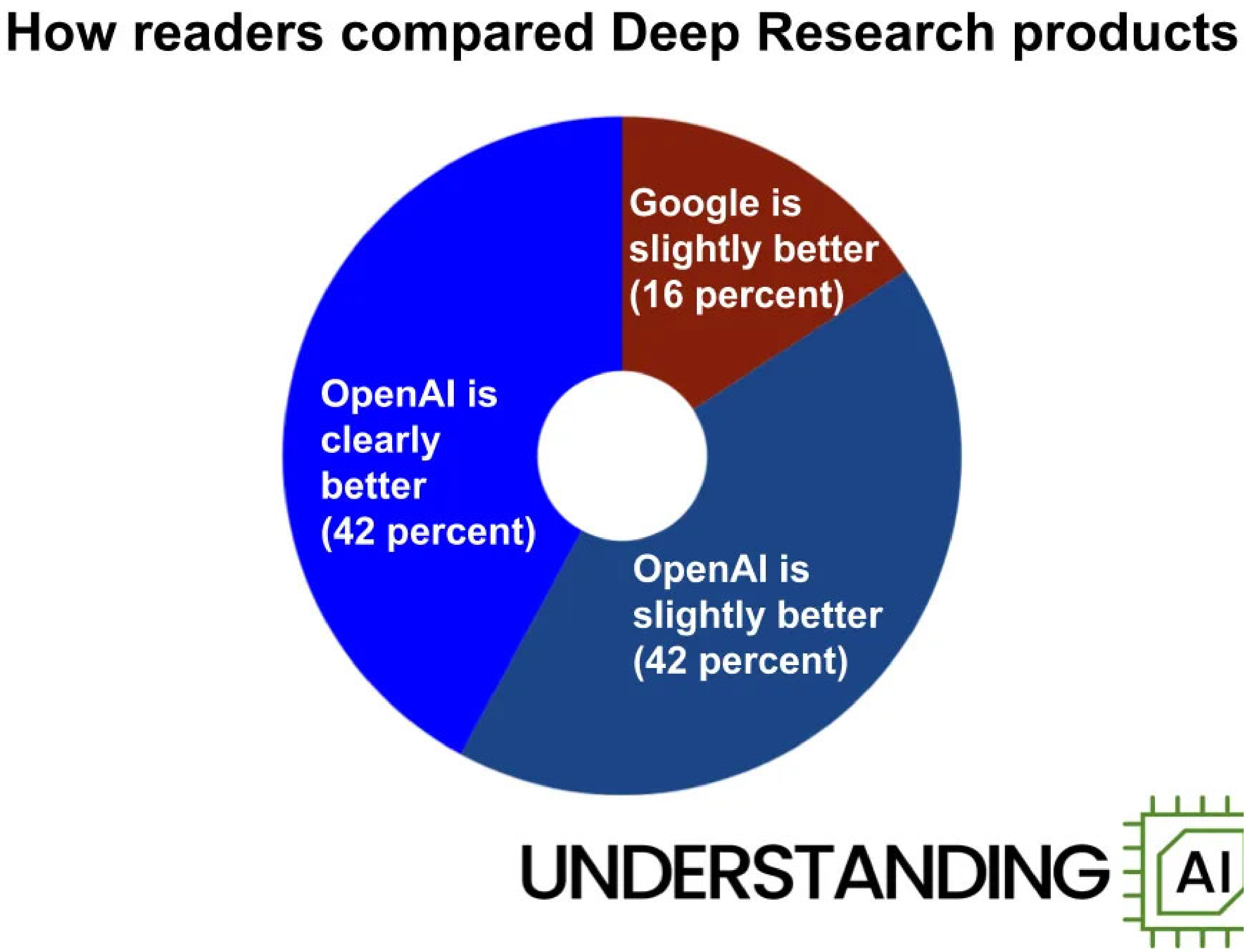
A big open question after the release of o1 was how well the same techniques woul generalize to "softer" domains like law, architecture, or medicine. The strong performance of Deep Research suggests that those techniques generalize better tha many people—including me—expected. And if that's true, we should expect to see continued rapid progress in AI capabilities over the next year—and perhaps beyon that.

People prefer OpenAI's Deep Research to Googl

OpenAI didn't invent this product category. That distinction goes to Google, which introduced its own Deep Research product in December. So I asked my volunteers t evaluate both models.

Each participant sent me a difficult question in his or her area of expertise. I sent b two responses, one from OpenAI and one from Google. I didn't tell them which mo produced which response.

OpenAI's shortest responses were around 2,000 words and took four to five minutes write. The longest—a detailed analysis of fantasy football players and strategies—ra more than 18,000 words and took OpenAI Deep Research 17 minutes to write. On average, Google's responses tended to be a bit faster and shorter than OpenAI's.


How readers compared Deep Research products

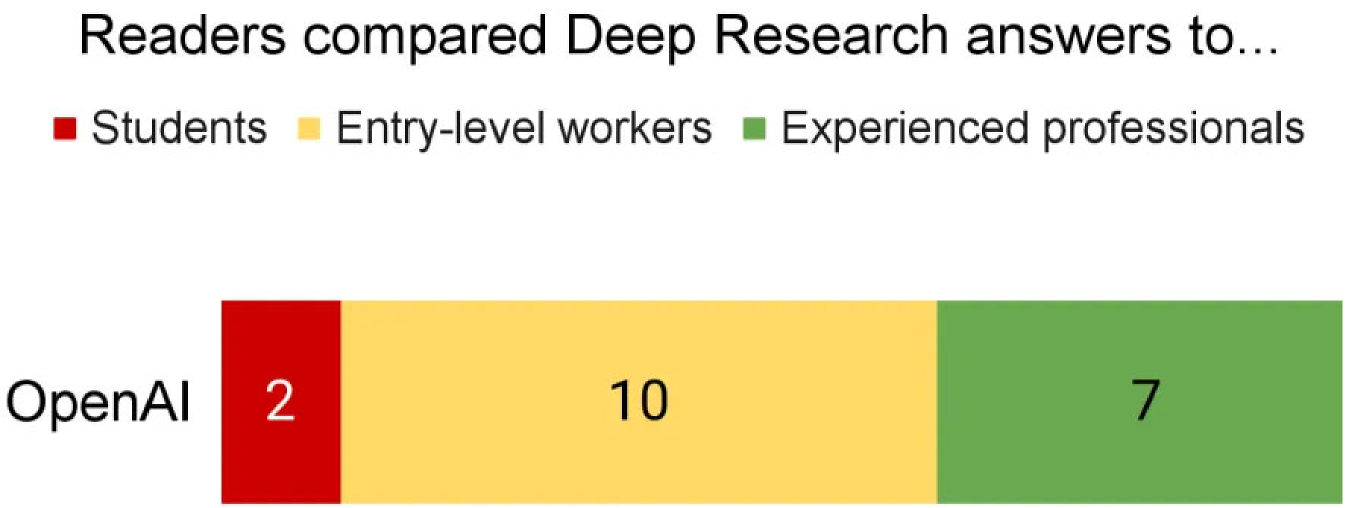
Sixteen out of 19 readers said they preferred the OpenAI response, whereas only th people thought Google's response was better.

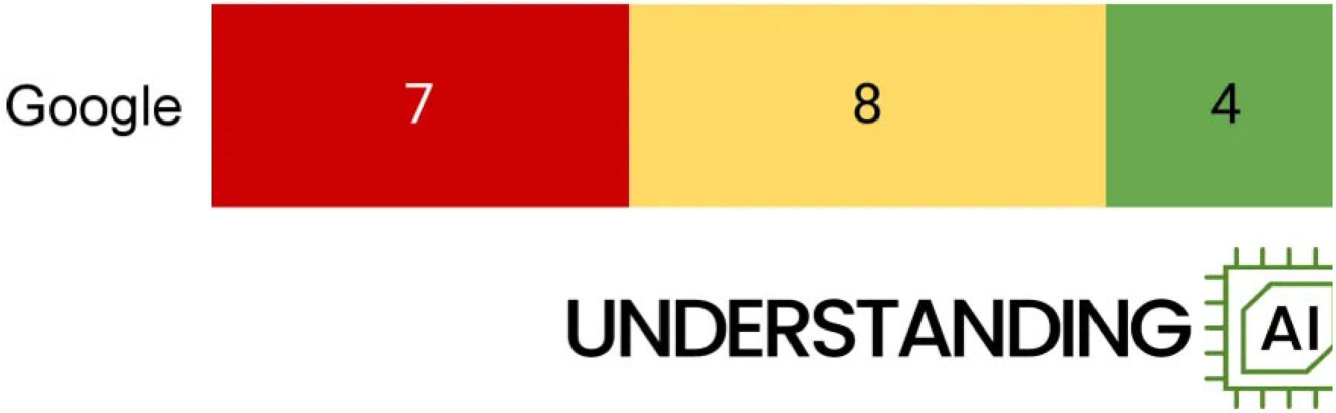
Many of my volunteer judges were impressed by OpenAI's answers. An antitrust lawyer told me an 8,000-word report “compares favorably with an entry-level attorn and that it would take 15 to 20 hours for a human researcher to compile the same information. She said she would like to use OpenAI’s tool professionally—especiall it could be hooked up to commercial databases like Westlaw or LexisNexis, which would give it access to more obscure legal rulings.

Chris May, a mechanical engineer, asked for directions on how to build a hydrogen electrolysis plant. He estimated that it would take an experienced professional a we to create something as good as the 4,000-word report OpenAI generated in four minutes.

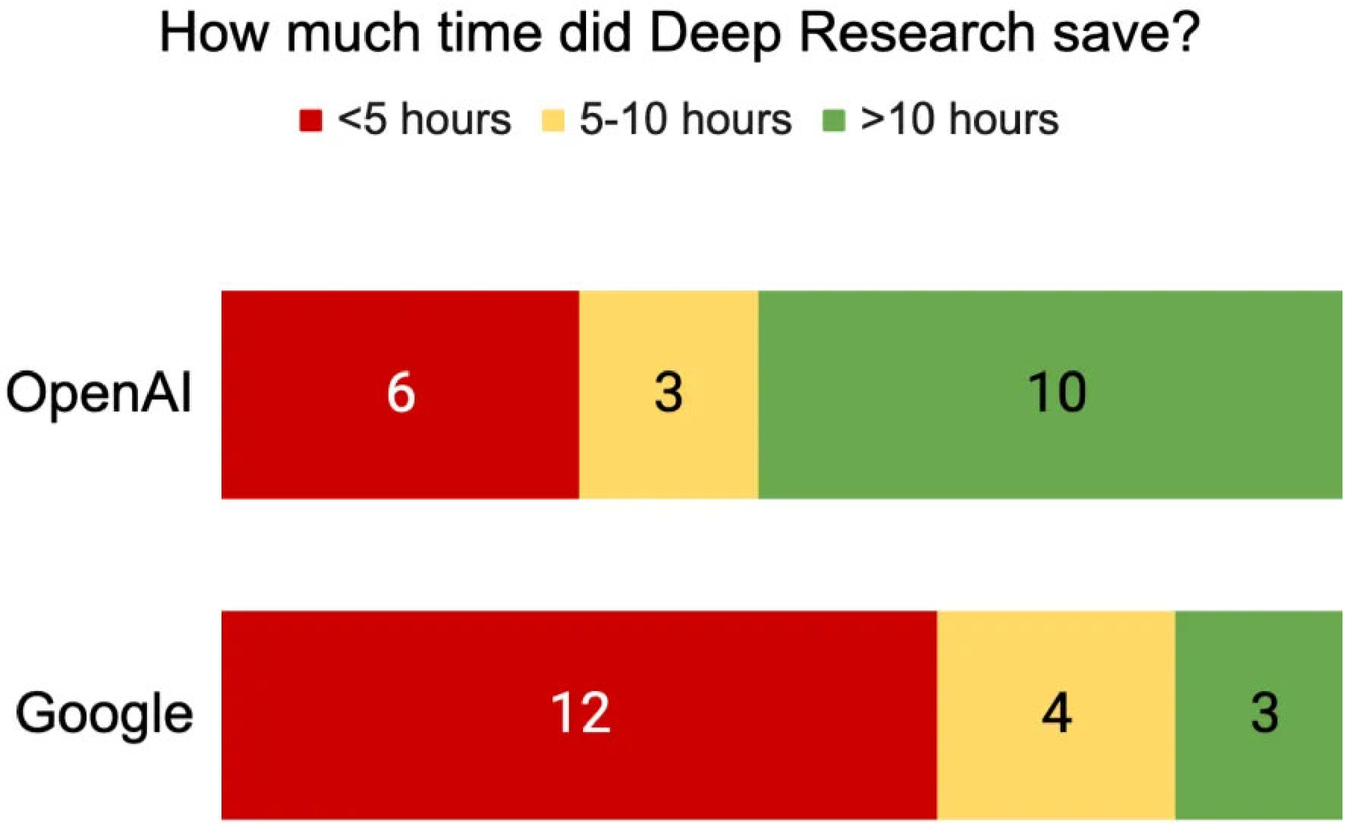
Heather Black Alexander, a middle school teacher in Chicago, praised a 12,000-wor report about middle school advisory programs that OpenAI produced in seven minutes. Alexander said the report was better than she’d expect from an entry-level employee, and estimated it would take a week for a human researcher to write it.

A few people noticed that responses omitted recent information such as Donald Trump’s election. This could be because the models were trained before Trump woı the election. However, these “Deep Research” products are also supposed to seek oı additional information by searching the web, so they should be able to learn about recent developments.





In the chart above, the green bars represent people who said a model produced wor that was at the level of an experienced professional—or at least above the level of a entry-level worker in their field. Yellow represents people who compared Deep Research responses to entry-level employees or interns. Red represents people who compared them to medical students, college students, high school students, or wors As you can see, readers were significantly more impressed by OpenAI's model.



Here I've broken down how much time people thought it would take for a human being to produce a report of comparable quality. There was a huge range. Four peo estimated it would take a week for a human researcher to duplicate an OpenAI rep No one thought any of the Google reports would take that long. On the flipside, tw readers said it would take only 30 minutes to reproduce Google responses. No one s that about an OpenAI report.

If you're a paying subscriber, you can scroll down to the bottom of this article to see how every one of the 19 participants rated OpenAI and Google's responses.

A better way to RAG

Companies have been rushing to adopt LLMs over the last two years. One of the m popular applications has been chatbots powered by a technique called retrieval augmented generation.

Suppose you run a company that has a million documents on its servers—corporate memos, customer service requests, instruction manuals, sales contracts, and so fort You want a chatbot that "knows about" all of these documents and can answer questions about their contents.

When a user asks a question, a RAG system searches for relevant documents using keyword search, vector database, or other techniques. The most relevant document are inserted into an LLM's context window. When it works well, a RAG system crea the illusion of a chatbot that understands thousands or even millions of documents

But if the user's question is complex or poorly worded, the RAG system might fail t retrieve the right documents. This is a common failure mode because the techniqu used to find and rank relevant documents aren't as "smart" as the LLM that genera the final answer.

The new Deep Research products point toward a better paradigm for RAG applications: if the initial search doesn't turn up the right documents, the system ca search again with different keywords or parameters. Doing this over and over again as OpenAI's Deep Research does—will produce a much better result than a traditio RAG pipeline.

The reason people haven't been doing this already is that early LLMs weren't good enough at following long chains of reasoning. If someone had tried to use the Deep Research technique with GPT-4 back in 2023, the model would have gotten "stuck" after a few searches.

But now that OpenAI has demonstrated how well this paradigm works, it should be straightforward for companies with existing, underperforming RAG applications to upgrade them with better models and a more iterative process for document retriev That should yield dramatically better performance, and I expect it to drive renewed enthusiasm for this type of system.

Interestingly, Google's Deep Research product seems to be somewhere in between OpenAI's approach and a traditional RAG system. Like a traditional RAG system, Google's Deep Research operates in two phases—first retrieving a bunch of documents and then generating an output. But within the first stage, Google's Deep Research has an iterative search process where the result of one search informs the next one.

I don't know if Google's product performs relatively poorly because it has a more ri reasoning process or because Google's underlying model simply isn't as good as OpenAI's o3. Or maybe these issues are connected: maybe the open-ended search process used by OpenAI's product is only possible with a powerful reasoning mode like o3.

Either way, I'm sure Google is working hard to regain its lead in a product category Google invented just a few months ago.

The coming AI speedup

The success of Deep Research also suggests that there's a lot of room to improve AI models using "self play." The big insight of o1 was that allowing a model to "think" longer leads to better answers. OpenAI's Deep Research demonstrates that this is t for a wide range of fields beyond math and computer programming.

And this suggests there's a lot of room for these models to "teach themselves" to ge better at a wide range of cognitive tasks. A company like OpenAI or Google can generate training data by having a model "think about" a question for a long time. Once it has the right answer, it can use the answer—and the associated thinking tokens—to train the next generation of reasoning models.

Because the training algorithm knows the correct answer, it should be able to train new model to get to the right answer more quickly. And then this new model can generate a new batch of training data that focuses on even harder problems.

I don't expect this process to get AI models all the way to human-level intelligence because they will eventually bump into the limitations I wrote about back in December. But the success of Deep Research makes me think the current paradigm has more headroom than I thought just a few weeks ago.

How readers responded

Now here is a bonus for paying subscribers: a summary of how each of my 19 volunteers judged OpenAI and Google responses.