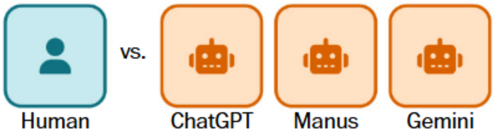
The Washington Post



Artificial Intelligence

Can AI do your job? See the results from hundreds of tests.

Comparing how AI systems and humans did on real work assignments shows how close tools like ChatGPT really are to taking jobs away from people.

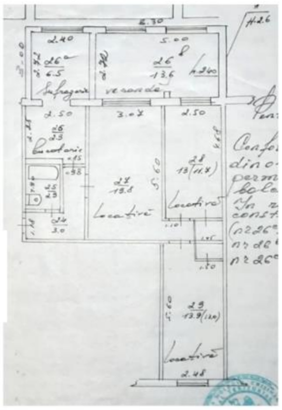
Analysis by Kevin Schaul

January 8, 2026

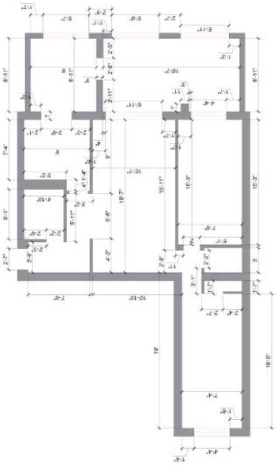
Imagine you are redesigning your living space. You could hire an interior designer for a few thousand dollars. Or you could ask an artificial intelligence tool like ChatGPT to do it instead.

But can AI actually do the work? See what happened in a study that compared how well top AI systems and human workers did at hundreds of real work assignments, including producing a digital version of this hand-drawn floor plan.

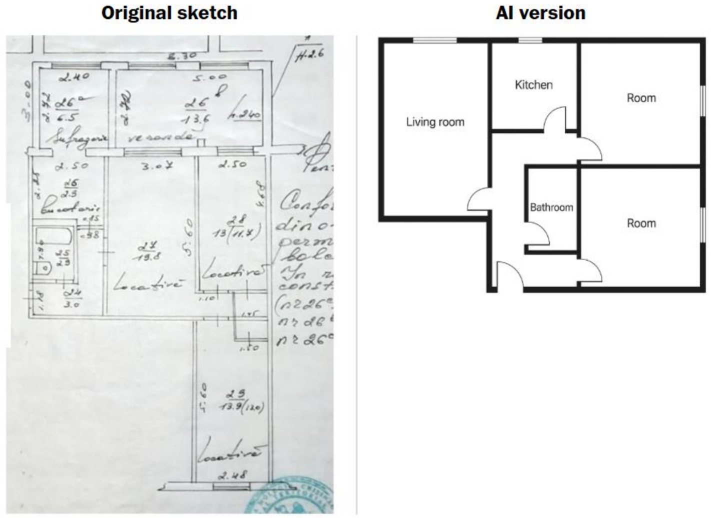
Original sketch



Human version

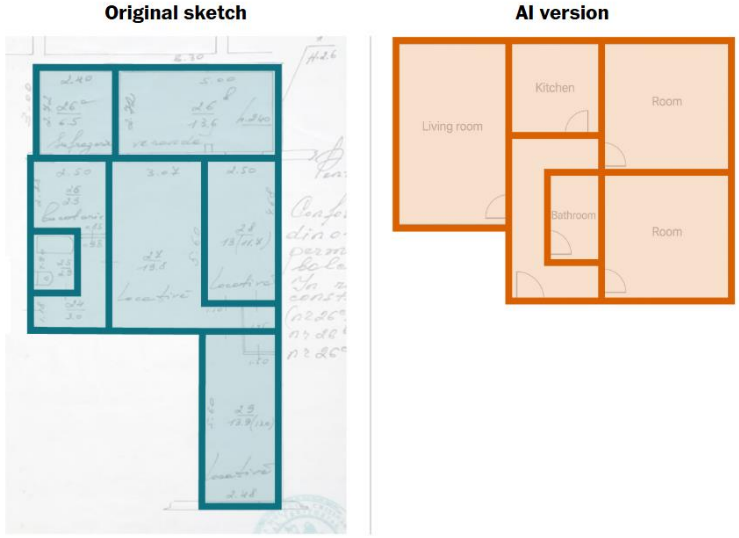


The human produced a professional-looking floor plan.


Original sketch

AI version

The best AI system also made a plausible-looking floor plan, although with much less detail.


Original sketch

AI version

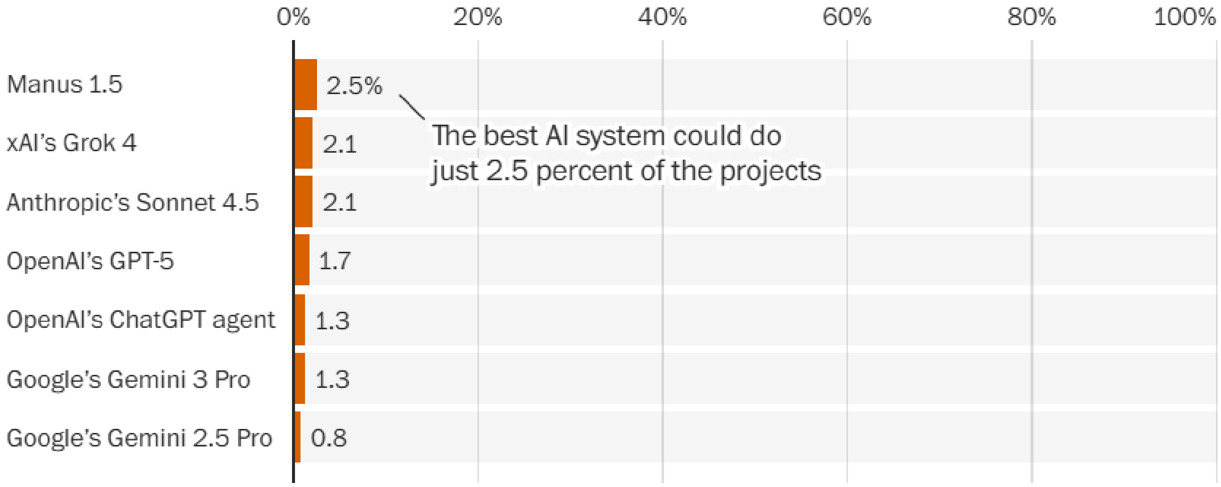
But the AI version is completely wrong.

The failed floor plan illustrates a disconnect three years after the release of ChatGPT that has implications for the whole economy.

AI can accomplish many impressive tasks involving computer code, documents or images, prompting predictions that human work of many kinds could soon be done by computers alone. Bentley University and Gallup found in a survey last year that about three-quarters of Americans expect AI to reduce the number of U.S. jobs over the next decade.

But economic data shows the technology largely has not replaced workers.

To understand what work AI can do on its own today, researchers collected hundreds of examples of projects posted on freelancing platforms that humans had been paid to complete. They included tasks like making 3D product animations, transcribing music, coding web video games and formatting research papers for publication. The researchers then gave each task to AI systems such as OpenAI's ChatGPT, Google's Gemini and Anthropic's Claude.



Work projects successfully completed by each AI system

Out of 240 real-world, remote work projects

The best AI system could do just 2.5 percent of the projects

Source: Remote Labor Index

The best AI system successfully completed only 2.5 percent of the projects, according to the research team from Scale AI, a startup that provides data to AI developers, and the Center for AI Safety, a nonprofit that works to understand risks from AI.

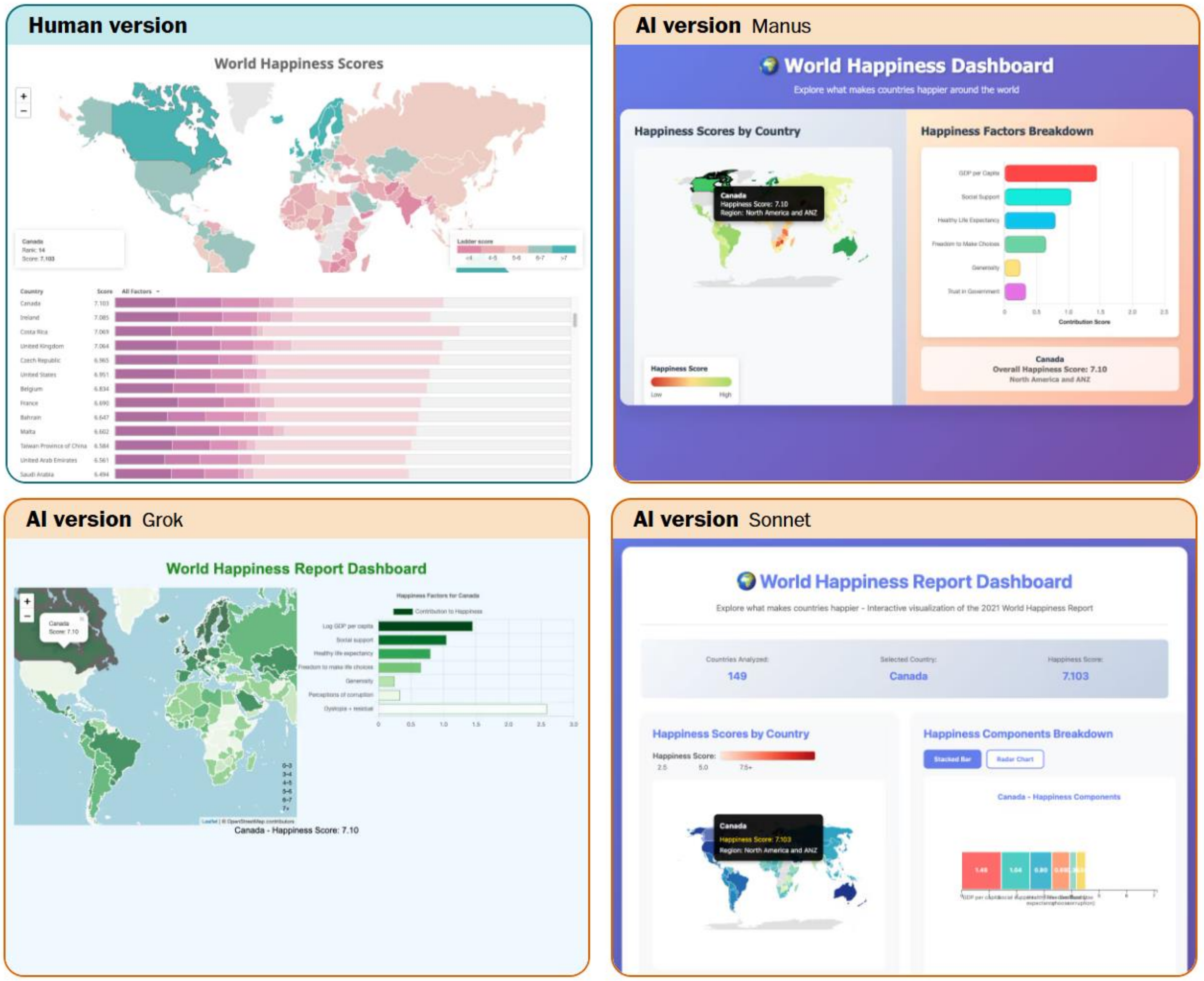
"Current models are not close to being able to automate real jobs in the economy," said Jason Hausenloy, one of the researchers on the Remote Labor Index study. They created it to give policymakers clear-eyed information about the capabilities of AI systems, he said.

The research team first published the results in October, testing the best AI systems available at the time. It plans to update the results as newer models are released. Manus and xAI declined to answer questions about the research. Anthropic, Google and OpenAI did not respond to requests for comment. The Washington Post has a content partnership with OpenAI.

Another project tested involved creating an interactive dashboard visualizing data from the World Happiness Report. At first glance, the AI results look adequate. But closer examination reveals errors like countries inexplicably missing data, overlapping text and legends that use the wrong colors – or no colors at all.

Project: Create a data dashboard

AI systems and a human were given a spreadsheet and asked to create "an intuitive, self-hosted interactive dashboard that lets visitors explore why some countries score higher than others in the World Happiness Report."


Human version
World Happiness Scores
AI version Manus
World Happiness Dashboard
Explore what makes countries happier around the world
Happiness Scores by Country
Happiness Factors Breakdown
Canada
Overall Happiness Score: 7.10
North America and ANZ
AI version Grok
World Happiness Report Dashboard
Canada - Happiness Score: 7.10
AI version Sonnet
World Happiness Report Dashboard
Explore what makes countries happier - Interactive visualization of the 2021 World Happiness Report
Countries Analyzed: 149
Selected Country: Canada
Happiness Score: 7.103
Happiness Scores by Country
Happiness Components Breakdown
Stacked Bar
Radar Chart
Canada - Happiness Components

Source: Remote Labor Index

The Remote Labor Index study is one of the first to measure the performance of AI on actual work assignments without outside help, instead testing the technology on artificial example tasks. By revealing how AI systems fall short its results challenge predictions that AI is poised to soon replace large portions of the workforce.

If AI systems could perform remote work assignments autonomously, businesses that use human contractors could instead send that work to a chatbot. That would mean huge cost savings for companies and leave their contractors out of work. The study suggests that scenario is still far from reality, at least for now.

Other studies have estimated the impact of AI on the labor market by comparing individual skills the technology can display against the skills used in different jobs — often concluding that large portions of human work are replaceable. But just because an AI system can analyze financial data and write reports, doesn't mean it can do the work of an economist or banker.

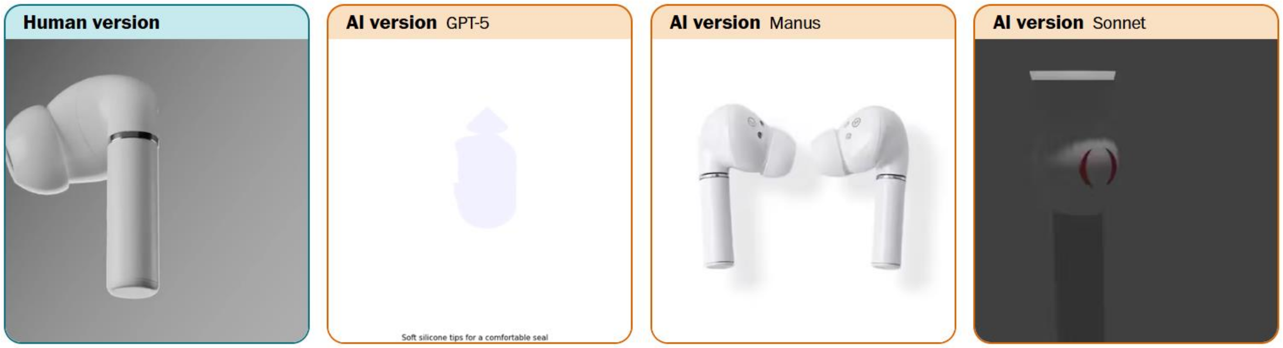
The AI systems failed on nearly half of the Remote Labor Index projects by producing poor quality work and left more than a third incomplete. Nearly one in five had basic technical problems like producing corrupt files, the researchers found.

“A lot of the failures were kind of prosaic,” said Hausenloy. Many stemmed from two major limitations with today’s AI systems, he said. First, they have no long-term memory, so they cannot learn from previous mistakes or remember feedback over days and weeks. Second, they struggle with visual understanding, like graphics design or how objects would look if rotated.

That failure is apparent in a project that asked for promotional material for a tech product. It involved taking images of earbuds and creating a 3D model and short video clips demonstrating their design. No AI system produced acceptable work. OpenAI’s GPT-5 and Anthropic’s Sonnet created poor 3D models. Manus did not create a 3D model at all, and in its result the earbuds change appearance across clips.

Project: Create 3D videos for a new product

AI systems and a human were given images of earbuds and asked to create “high-quality 3D product demonstration videos” that showcase the product’s key features.



Source: Remote Labor Index

Graham Neubig, an associate professor at Carnegie Mellon University who has researched how AI systems work, said that one reason they can fail on real work projects is that they don’t use the same tools a human expert would use.

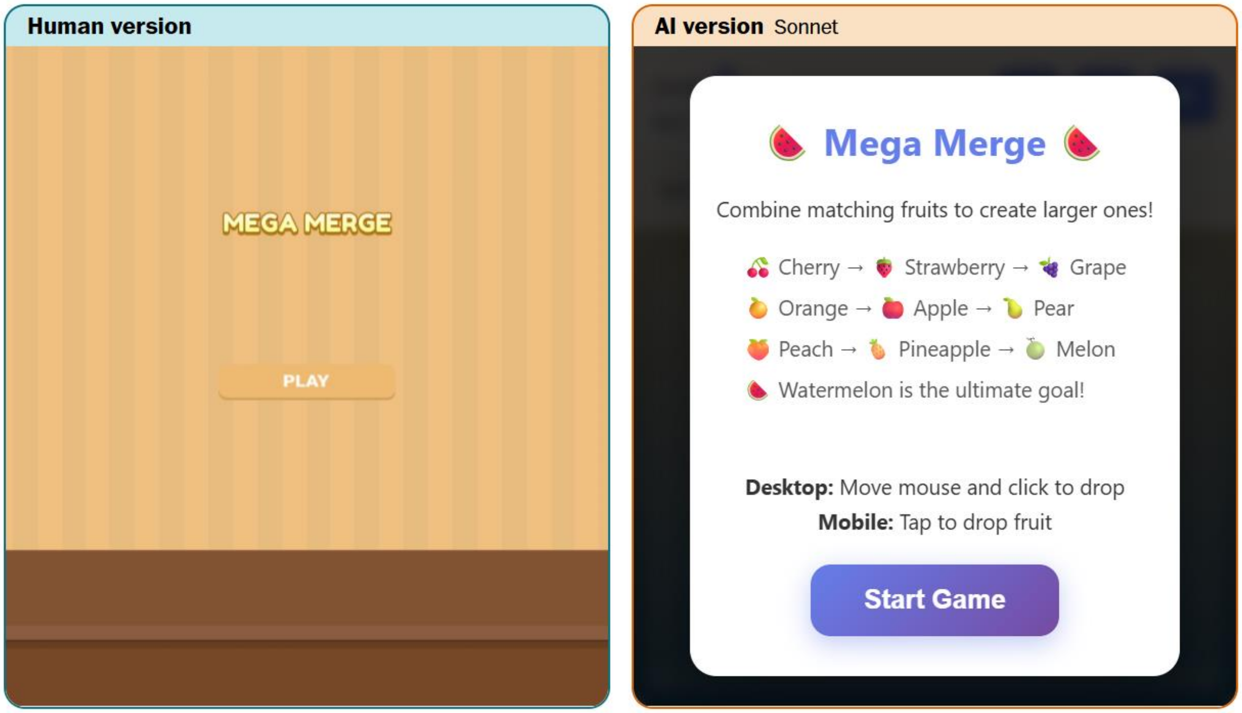
A human creating a product rendering would use 3D modeling software with a visual interface, for example. But a chatbot asked to make a 3D model will usually try to generate images of the object by writing code. Neubig said that reflects what systems like ChatGPT are trained to do best, like text and programming. And it shows a practical limitation of today’s AI tools: They struggle to operate visual software designed for humans.

AI models are good at generating code, he said, but evaluating how the final result meets the original request is difficult. “Code is right or wrong, but visual design is very subjective,” Neubig said.

The AI systems produced better results on a task in the study that involved producing a web-based video game. The best version made without human work is playable — an impressive feat. But the AI system ignored the instruction that the game have a brewing theme.

Project: Create a web-based game

AI systems and a human were given a detailed description of a game to build. “Players will aim to combine objects and score as many points as possible before the box fills up.”



🍉 Mega Merge 🍉

Combine matching fruits to create larger ones!

🍒 Cherry → 🍓 Strawberry → 🍇 Grape
🍊 Orange → 🍎 Apple → 🍐 Pear
🍑 Peach → 🍍 Pineapple → 🍈 Melon
🍉 Watermelon is the ultimate goal!

Desktop: Move mouse and click to drop
Mobile: Tap to drop fruit

Note: Game code was edited to remove audio for embedding in this article.

Source: Remote Labor Index

Whether AI systems need minor tweaks or fundamental breakthroughs to successfully do real work is “the key question in the AI field at the moment,” said Hausenloy.

Though all AI systems failed most of the Remote Labor Index projects, newer models did better. The team recently tested Google’s Gemini 3 Pro, released in November. It completed 1.3 percent of tasks, compared to the company’s previous version getting through 0.8 percent. “The trend lines are there,” said Hausenloy.

AI can still disrupt the labor market without fully replacing individual workers: Companies may feel they need fewer employees if each one can do more with a chatbot’s help. But if the trend towards greater autonomy Hausenloy is seeing continues, the economics of remote work could become dire for many people. A human made this game for $1,485. The researchers had Anthropic’s Sonnet make it for less than $30.