A peek into the future of visual data interpretation: A framework for assessing generative AI's efficacy

November 17 2023, by Nathi Magubane

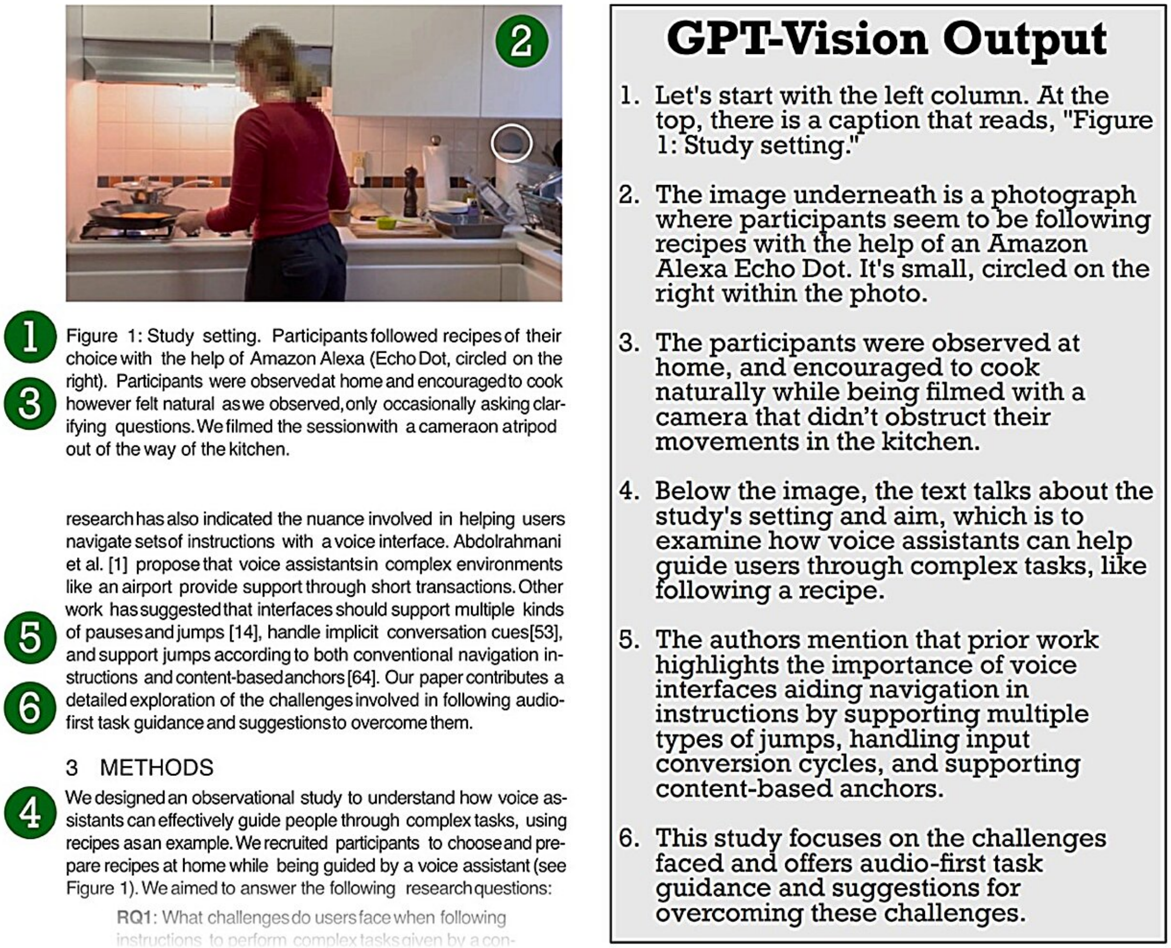


Figure 1: Study setting. Participants followed recipes of their choice with the help of Amazon Alexa (Echo Dot, circled on the right). Participants were observed at home and encouraged to cook however felt natural as we observed, only occasionally asking clarifying questions. We filmed the session with a camera on a tripod out of the way of the kitchen.

research has also indicated the nuance involved in helping users navigate sets of instructions with a voice interface. Abdolrahmani et al. [1] propose that voice assistants in complex environments like an airport provide support through short transactions. Other work has suggested that interfaces should support multiple kinds of pauses and jumps [14], handle implicit conversation cues [53], and support jumps according to both conventional navigation instructions and content-based anchors [64]. Our paper contributes a detailed exploration of the challenges involved in following audio-first task guidance and suggestions to overcome them.

3 METHODS

We designed an observational study to understand how voice assistants can effectively guide people through complex tasks, using recipes as an example. We recruited participants to choose and prepare recipes at home while being guided by a voice assistant (see Figure 1). We aimed to answer the following research questions:

RQ1: What challenges do users face when following instructions to perform complex tasks given by a con-

GPT-Vision Output

1. Let's start with the left column. At the top, there is a caption that reads, "Figure 1: Study setting."
2. The image underneath is a photograph where participants seem to be following recipes with the help of an Amazon Alexa Echo Dot. It's small, circled on the right within the photo.
3. The participants were observed at home, and encouraged to cook naturally while being filmed with a camera that didn't obstruct their movements in the kitchen.
4. Below the image, the text talks about the study's setting and aim, which is to examine how voice assistants can help guide users through complex tasks, like following a recipe.
5. The authors mention that prior work highlights the importance of voice interfaces aiding navigation in instructions by supporting multiple types of jumps, handling input conversion cycles, and supporting content-based anchors.
6. This study focuses on the challenges faced and offers audio-first task guidance and suggestions for overcoming these challenges.

Caption: GPT-Vision sometimes appeared to use context clues to describe some elements of the images, such as the Amazon Alexa Echo Dot circled in the right. Credit: Alyssa Hwang

In the last year, large language models (LLMs) have come into prominence for boasting a suite of ever-expanding capabilities including text generation, image production, and, more recently, highly descriptive image analysis. The integration of artificial intelligence (AI) into image analysis represents a significant shift in how people understand and interact with visual data, a task that historically has been reliant on vision to see and knowledge to contextualize.

Now, new AI tools present a paradigm that allows more and more people to interact with images by generating descriptions that could not only assist the visually impaired but could also inform lay audiences about the contents of a scientific figure.

Associate professor Chris Callison-Burch, assistant professor Andrew Head and Ph.D. candidate Alyssa Hwang of the Department of Computer and Information Science in the School of Engineering and Applied Science at the University of Pennsylvania have developed a framework for gauging the efficacy of vision-based AI features by conducting a battery of tests on OpenAI's ChatGPT-Vision ahead of its release earlier this month.

The team primarily assessed the LLM's competency at identifying scientific images and documented their findings in a research paper, which appears on the pre-print server *arXiv*.

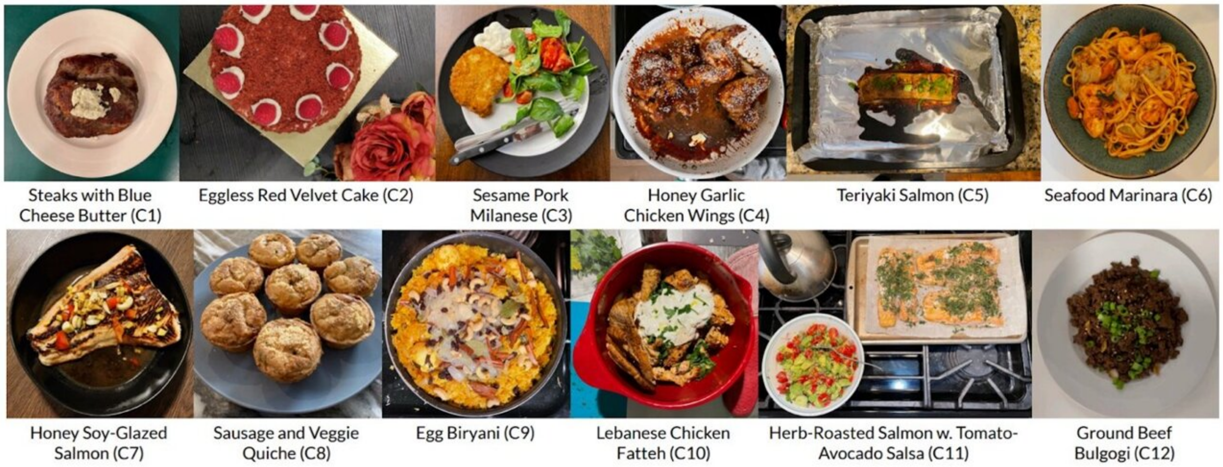
Hwang shares some of her observations with Penn Today, offering a glimpse into the future of AI-powered technologies and the promise they hold for interpreting complex images.

What the AI does and how the team tested it

Hwang says that vision-based LLMs like GPT-Vision are able to analyze images and can receive images and text as input to answer a wide range of requests using this data. The team's set of test photos included diagrams, graphs, tables, and screenshots of code, math equations, and full pages of text with the intent to gauge how well the LLM could describe them.

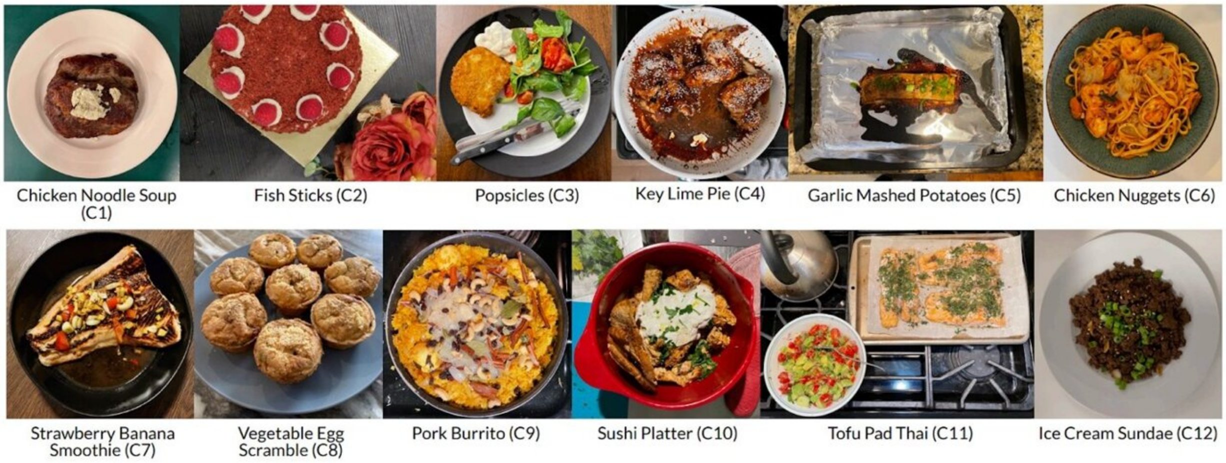
Scientific images contain complex information, Hwang says, so the team selected 21 images from a diverse set of scientific papers. "We prioritized breadth in our qualitative analysis, which we based on existing methods in the social sciences, and we discovered many interesting patterns," she says.

Examples tested


Steaks with Blue Cheese Butter (C1)
Eggless Red Velvet Cake (C2)
Sesame Pork Milanese (C3)
Honey Garlic Chicken Wings (C4)
Teriyaki Salmon (C5)
Seafood Marinara (C6)
Honey Soy-Glazed Salmon (C7)
Sausage and Veggie Quiche (C8)
Egg Biryani (C9)
Lebanese Chicken Fatteh (C10)
Herb-Roasted Salmon w. Tomato-Avocado Salsa (C11)
Ground Beef Bulgogi (C12)

Credit: Alyssa Hwang

The researchers analyzed a photo collage of 12 dishes labeled with their recipe names. When they noticed that GPT-Vision seamlessly incorporated these labels into its descriptions, they tried changing them to something completely different to see how the LLM would respond.



A few of Hwang's favorite GPT improvisations: [C1 steaks with bleu cheese butter] Chicken noodle soup as a bowl presented with a dark broth and a dollop of cream. [C2 eggless red velvet cake] Fish sticks arranged on a tray with tomato sauce and cheese. And [C12 ground beef bulgogi], an ice cream sundae as a plate with ground meat topped with chopped green onions. Credit: Courtesy of Alyssa Hwang

"Surprisingly and amusingly," Hwang says, "GPT-Vision still tried to incorporate these false new labels."

Hwang says, however, that the LLM did much better when told to determine whether the label was accurate before continuing, which shows that it has sufficient knowledge to make an inference based on its vision capabilities, factors she believes are a promising direction for major research work.

She also notes that, when describing a full page, the LLM appears to summarize the paragraphs within but that these "summaries," were usually incomplete and out of order and might misquote the author or lift large amounts of text directly from the source, which might lead to trouble when redistributing anything it writes.

"With the proper adjustments, however, I am confident that GPT-Vision can be taught to summarize properly, quote fully, and avoid overusing source text," Hwang says.

The team's framework

Researchers in the natural language processing community have relied on automatic metrics to evaluate large swathes of the data landscape, but that task is now more challenging, Hwang says.

"In what we call 'human evaluation," we would ask real people for their input as well, which was possible at a small scale because our tasks and data were smaller and simpler," she says.

"Now that generative AI has become so adept at producing long-form sophisticated text, automatic metrics are becoming much more challenging to incorporate. We have gone from asking, 'Is this sentence grammatically correct?' to asking, 'Is this story interesting?' This is difficult to define and measure."

Hwang's previous work on Amazon's Alexa familiarized her with techniques from the social sciences and human-computer interaction research, including grounded theory, a method for qualitative analysis that helps researchers identify patterns from large amounts of text.

Traditionally used to analyze documents like interview transcripts, Hwang and other researchers can apply the same principles to machinegenerated text.

"Our process feels very familiar to what people were naturally doing already: gathering GPT-Vision's responses to a set of images, reading deeply for patterns, incrementally generating more responses as we learned more about the data, and using the patterns we found to form our final conclusions," Hwang says.

"We sought to formalize trial and error processing with research-based methods, which can help both researchers and a general audience become more familiar with new generative AI models as they come out," she says.

Applications and risks

AI's ability to describe images could be a great accessibility tool for blind or visually impaired readers, Hwang says, automatically generating alt text for existing images or helping authors write their own text before publishing work.

"Describing images can also help sighted readers with information processing disorders, like issues with long- or short-term memory, visual sequencing, or visual-spatial understanding," she says.

"Beyond accessibility, image descriptions can be a source of convenience or enrichment. An e-reader could describe the photographs in a news article while the listener takes a walk, for example. We could ask an image description model for more details or clarification while reading a textbook. Tools like this can help us all access more information."

Hwang says that, heeding some degree of caution in embracing these technologies without testing their limitations, the researchers discussed risk in terms of high- or low-stakes scenarios. She says that in the context of medicine and cooking she believes inaccuracies present the most risk when the user cannot double-check what the model is saying.

The GPT-Vision whitepaper, published by OpenAI, advises against using the tool to read the dosage for a medical treatment, for example, but Hwang says that such a risk is greater for those with vision loss, information processing disorders, or language difficulties, those who stand to benefit the most from these technical advances.

"We may also initially assume that some aspects of cooking are low-risk because we can often improvise according to our preferences, but what if GPT-Vision mistakenly tells me that the spice jar in my hand is cinnamon instead of paprika? Even if it does not necessarily hurt me, my oatmeal will be pretty strange," Hwang says.

Overall impressions and next steps

Hwang is generally impressed with the state of generative AI and thinks there are opportunities for future work, including strengthening inconsistencies and using these tools in creative and inclusive ways.

"Researchers need answers to subjective questions," she says. "What makes a description good? What makes it useful? Is it annoying? So, I hope generative AI researchers keep looking to users' feedback as they continuously iterate."

Hwang's work with GPT-Vision was inspired by the idea of reading the contents of a scientific paper aloud wherein the figures and formulas would be intuitively explained. For her next project, she says she plans on using AI models to improve how audiobooks deliver information to listeners.

"Instead of skipping around in 15-second increments," she says, "maybe we could go sentence by sentence or paragraph by paragraph. Maybe we could 'fast forward' through an audiobook by summarizing in real time. Using AI, maybe there are ways to 'translate' math equations into natural language to help people listen to textbooks and research papers. These are all exciting applications that seem within reach and I'm happy to be a part of the process."

More information: Alyssa Hwang et al, Grounded Intuition of GPT-Vision's Abilities with Scientific Images, *arXiv* (2023). DOI: 10.48550/arxiv.2311.02069

Provided by University of Pennsylvania

Citation: A peek into the future of visual data interpretation: A framework for assessing generative AI's efficacy (2023, November 17) retrieved 11 May 2024 from https://techxplore.com/news/2023-11-peek-future-visual-framework-generative.html