ALM | LAW.COM

New York Law Journal

TECHNOLOGY TODAY

ARTIFICIAL INTELLIGENCE

AI Has Opened Its Eyes

By Katherine B. Forrest
January 30, 2024

It was early October 2023 when I realized that artificial intelligence (AI) had opened its eyes. I was reviewing recent articles on AI and ran across OpenAI's "GPT-4V(ision) System Card", released on Sept. 25, 2023. The paper was only a few days old when I saw it for the first time, and it took me a few minutes to understand what I was reading. GPT-4V, a "multi-modal LLM" (MLLM), is a large language model (LLM) trained on various modalities of content—not just text. In addition to the enormous amount of data that ChatGPT and GPT -4 have been trained on, it has also been trained on video content of all types and it can be trained on additional modes of communication. In effect, it learns about the world around it not by reading about it, but by looking at pictures, charts, colors, facial expressions, buildings, the sky…everything. With its eyes open, AI has taken a cognitive leap forward.

Below, I describe some of the academic papers and research regarding MLLMs, then move on to legal and other important issues that are similar between LLMs and MLLMs, those that are different but obvious improvements, and those that should cause all of us to watch developments in this area carefully.

KATHERINE B. FORREST *is a partner in Paul, Weiss, Rifkind, Wharton & Garrison's litigation department and a member of the antitrust practice group. She previously served as a U.S. District Judge for the Southern District of New York and as a Deputy Assistant Attorney General in the U.S. Department of Justice's antitrust division.*

Katherine B. Forrest

First, helpful background: OpenAI's Sept. 25, 2023, paper stated that it trained GPT-4V in 2022, but was just widely releasing it at that time. During the interim, the model had gone through extensive safety and accuracy testing and tuning. Its capabilities were tested against GPT-4 (the LLM), and instructions placed into the model to prevent it from discussing images in ways that were inappropriate, ill advised or reflected a need for additional training.

The release of the Sept. 25 paper, along with the GPT-4V model itself in turn set in motion a wave of additional work on what is now widely recognized as another tremendous cognitive leap in AI (Ge, et al., "MLLM Bench, Evaluating Multi-Modal LLMs Using GPT-4V", Nov. 2023, "In the pursuit of Artificial General Intelligence (AGI), the integration of vision in language models has marked a significant milestone.")

Dozens of academic papers have been published in the last few months examining the capabilities of MLLMs (Google's Gemini Pro is also an MLLM), and use cases. (See https://www.arxiv.org/MLLM). As a category, the combination of training modalities has enhanced overall accuracy of answers, performs better than GPT-4 (a regular LLM) on structured reasoning tasks (See Singh, et al, Dec. 2023, "Assessing GPT4-

ADOBE STOCK

V on Structured Reasoning Tasks"), and is showing increasing usefulness in tasks that involve emotion recognition and sentiment analysis (Lian, et al, Dec. 2023, "GPT-4V with Emotion: A Zero-Shot Benchmark for Multimodal Emotion Understanding").

MLLMs present some of the same legal questions as LLMs—they are, after all, on a similar evolutionary path. Privacy and potential bias issues that commentators have long discussed and addressed as part of responsible and ethical AI in the past, remain present in MLLMs. The form of such issues has, however, shifted somewhat.

With the MLLMs, privacy issues expand when a prompt or question can be based on a photograph. With LLMs, privacy questions are often related to leakage of private information about an individual, but in text form (a social security number, a narrative about medical issues, a home address, etc). With its eyes open, MLLMs can see a photograph and from very little information, some of which may be in the metadata itself, answer questions about the actual physical location of a person. This could raise privacy and security concerns for some. But the ability to interpret things about the physical body from a photograph could lead not only to accurate identification, but also potentially aspects of medical conditions or issues that the individual had expected to remain private (e.g. "the pallor of the skin may indicate x, y, or z…." or the "weight loss between photo A and B may indicate a, b, or c"). MLLMs might also be able to identify people in a room or a crowd, potentially even attributing emotional states to their interactions (such as concern, sadness, happiness, contempt, anger).

Potential bias issues are also present with MLLMs, as they are with LLMs and narrower forms of AI. The training of an MLLM on a sufficiently inclusive data set is necessary to obtain fair and accurate output. An MLLM trained on a limited data set could arrive at biased conclusions that could be result in inaccurate facial recognition and biases in identification. A lack of diversity in data sets might also lead to inaccuracies in attributions of emotional states. A model trained on small changes in facial muscles of a white population might not accurately translate those to people of color; emotional variations specific to cultural backgrounds could also be lost with an overly narrow data set.

With LLMs, privacy questions are often related to leakage of private information about an individual, but in text form.

Moving away from the similarities between MLLMs and LLMs we can find a host of enhanced abilities. The ability of MLLMs to "see" also brings the potential for both significant benefits to certain groups (such as the visually impaired.) Part of OpenAI's testing of GPT4-V included a significant beta test with the visually impaired. (OpenAI, "GPT-4V(ision) System Card", Sept. 2023.) Access to justice could be enhanced as MLLMs enable the visually impaired to have photographic or video evidence described to them, or forms requiring visual navigation interpreted.

Because MLLMs can "see" and interpret images, we can show them mathematical equations which they can interpret and solve; they can review molecular combinations that they can understand and combine in novel ways; they can look out into the universe and tell what they see in the cosmos that we have not yet been able to discern.

But MLLMs also present us with new, more difficult issues and concerns. On a basic level, we have assumed that the basic security step of solving a

CAPTCHA—those images broken into boxes where you are asked to check all of the “fire hydrants”, “school buses” or “bridges” to ensure that you are a human. CAPTCHAs are now vulnerable as MLLMs learn to solve them.

Perhaps the most interesting and legally difficult aspect of MLLMs is their ability to read emotions. Imagine some of the use cases that could be just around the corner: jury consultants who use such technology to tell how jurors are reacting to testimony. While this exercise has been done now for years by humans watching humans, deeper meaning could be attributed to juror body language than ever before. Might this lead to more settlements or cause them to fall apart if additional certainty as to outcome is assumed? Would it enhance the likelihood of a mid-trial plea? Might the same technology be used to review the reaction of the judge in a bench trial—turning what used to be “reading of the tea leaves” into something purveyed as technical certainty?

Even more difficult use cases could come with such technology being used during an investigation phase, taking the place of the inadmissible polygraph: is the witness/suspect lying, uncomfortable, contemptuous? Is the investigative team able to show photographs, audio, or other evidence and elicit reactions? What about the utilization of such technology before *Miranda* warnings are provided? Would there need to be new rules for pre-disclosure? Today, viewing an interrogation from behind an obscured camera does not need to be disclosed, but would that change if the AI tool could effectively see “into” the witness/suspect in ways the witness did not know or understand?

MLLMS also present the possibility of an increased opportunity for fraud as MLLMs are able to combine photos, video or charts along with text as output. Studies have shown that people are more likely to believe incorrect information if they see it accompanied by an image or graphic—“seeing is still believing”—although many of us recognize that phrase as a relic of a quaint prior time.

All of these use cases provide enormous opportunities for discerning truth, so long as the accuracy of the tool achieves a level that instills confidence. But at the same time, they might diminish some of the ability for witnesses to exercise Fifth Amendment rights not to incriminate themselves. A less constitutionally problematic scenario might be the deprivation of a witness’s ability to exercise independent decision-making as to when to provide information and when not to.

As MLLMs continue to evolve and develop, they will provide use cases that were once the stuff of science fiction. They will allow an AI tool to learn about us by looking at us and seeing us say words which it can interpret with an array of unspoken emotional content.

MLLMS also present the possibility of an increased opportunity for fraud as MLLMs are able to combine photos, video or charts along with text as output.

The complexity of these models also raises transparency concerns: concerns that we will not understand how the models really learn, how they weigh the information provided to them through images, text and other modes of communication: will they weigh the curve of a cheek muscle in ways we never imagined, will another inch of distance between two people provide information we did not know was revealed by that seemingly small fact? Transparency of models has been a touchstone of regulatory bodies and voluntary commitments to AI model safety—but will we even understand what transparency looks like?

We are only at the beginning of understanding the capabilities of MLLMs. Now that AI has opened its eyes, it surely won’t be closing them again. Anyone who has previously taken comfort in the paradigm that a large language model is merely a sophisticated parrot spitting out the next word should think again. The old paradigm was dead and buried…sometime in at least September 2023.