



## WVU study finds AI makes mistakes diagnosing ER cases



ANYA SOSTEK 

Pittsburgh Post-Gazette

asostek@post-gazette.com

JUN 8, 2025

4:00 AM

AI has the potential to revolutionize medicine — but it's not there yet, according to a new study from West Virginia University.

The [study](#) tested ChatGPT using physician notes from real emergency department cases and found that the technology struggled when the symptoms were not straightforward.

“In challenging cases, ChatGPT was not able to make the diagnosis,” said Gangqing Hu, an assistant professor in the WVU School of Medicine and director of the school’s Bioinformatics Core facility. “ChatGPT is not competent to handle atypical presentations.”

In particular, ChatGPT failed to correctly diagnose three cases of pneumonia because the patients did not have a fever.

The study, published in the journal Scientific Reports, tested four different iterations of ChatGPT (GPT-3.5, GPT-4, GPT-4o and GPT-o1) with public data sets from 30 real emergency room cases. Each of the cases had one single diagnosis.

Asked to give the top three likely diagnoses, the models were 75% to 80% accurate in coming up with one of the correct answers. When the models were asked to explain their reasoning in coming to those conclusions, the accuracy generally improved to between 80% and 84%.



**Gangqing Hu, director of WVU's Bioinformatics Core facility, led the study looking at the limitations of AI use in the ER.  
(WVU Photo/Davidson Chan)**

When the cases were divided into easier and more difficult cases, however, ChatGPT showed its limitations. For the 13 of the 30 cases categorized as difficult, the AI models were less than 40% accurate at identifying the correct diagnosis.

In the case of the three pneumonia diagnoses, the researchers were able to identify that the problem came from the lack of fever. The study noted that while most pneumonia cases are accompanied by fever, that is not always the case — especially in populations such as older adults, infants,

immunocompromised individuals and those who took painkillers prior to the ED visit.

The issue likely arises from the material used by AI models to train themselves, which likely oversamples the typical cases.

“We don’t know what training data set that they use,” said Hu, who led the study. “If they use information from the internet or from books, most of the cases are typical presentations. The portion or the number of cases with atypical presentations is a minor portion or may not exist in those training data sets.”

Hu was surprised to see that when asked to identify the top three diagnoses, the newer versions of ChatGPT didn’t perform any better than the older versions. Asked for just the top diagnosis, however, the newest version was 15% to 20% higher in accuracy. The differences in versions could prove to be important as new ones are released.

“The AI field is moving very fast,” he said. “If we look at ChatGPT, almost every three to four months there will be a new iteration.”

AI can fill an important role in health care, he said, particularly in rural areas with a dearth of in-person health care professionals in some specialities. But this study highlights how health systems and regulators will have to grapple with inevitable mistakes from AI models.

“AI is not perfect,” he said. “Another thing we need to think about is the liability. If the AI model gave incorrect answers and you delay the treatment or it leads to mistreatment, who is going to take responsibility?”

*First Published: June 8, 2025, 4:00 a.m.*




**Anya Sostek** is a Post-Gazette health reporter who has been with the organization since 2004. A Duke University graduate, she previously worked as a business, education, news and features reporter.

✉ [asostek@post-gazette.com](mailto:asostek@post-gazette.com)


# Popular in the Community

**Protests intensify in Los Angeles after...**




**Ointment Fly**

There was a convenient pallet of cinderblocks ...



**Top Comment**



30

**Gene Co  
Constitu**



**EV**

It's not ev  
toward al



**Top Co**