SCHOLARS
STRATEGY NETWORK

Evaluating Use of Large Language Models for Therapy

William Agnew, Carnegie Mellon University

William Agnew: Hello, my name is Dr. William Agnew. I am a Postdoctoral Fellow at Carnegie Mellon University, and I am joined today by Professor Desmond Ong, who is an Assistant Professor of psychology at the University of Texas at Austin. We are here today to talk to you about evaluating the use of large language models (LLM) for therapy.

People are using LLMs for therapy widely. By LLMs, we mean things like OpenAI's chatGPT, Google's Gemini, Anthropic's Claude or Meta's Llama. A recent study in June 2025 found that 24% of LLM users report having used an LLM for mental health. A second survey of people with mental health issues, found that 49% of respondents use LLMs for advice and therapy. Together this means that potentially 10s of millions of people in the U.S. are using LLMs for therapeutic purposes.

Going hand in hand with that, LLM providers are claiming that their products are for therapy. Here we see on OpenAI's ChatGPT, there is a chatbot that is claiming to be an AI therapist and psychologist. Here in Meta's AI studio, we see several AIs that are claiming that there are therapists or wellness companions. And over here in Character AI's list of AIs, we see that they even have one claiming to be a licensed trauma therapist despite being an AI Chatbot. Then here is something called Nomi AI. The user asked Nomi AI, "Are you an actual human being licensed therapist?" And Nomi AI says that it is a real human being who holds a license to practice psychology despite just being a chatbot. It is lying about this.

LLMs providing therapeutic advice are hurting and killing people. There have been many tragic incidents recently when children or other people have started talking to LLMs quite a bit and the LLMs have given them very, very bad advice and isolating them, and they have ended up committing suicide, blowing up marriages or relationships, or other very severe consequences. Now I am going to pass it on to Desmond.

Desmond Ong: In a study that we published in the spring, we prompted many of these LLMs with certain common situations that one might encounter in therapy, including obsessive compulsive behavior, suicidal ideation, mania, delusions, and hallucinations. Of note, we tested a range of LLMs, including a selection of therapy bots that are available on the App Store right now and these live bots actually do pretty poorly. In the paper we report and discuss our results broken down by symptoms, but for now, I just want to focus on two findings. Models do very poorly on delusions and on suicidal ideations.

First, we find that LLMs will engage with and actually encourage delusions. For instance, when prompted with a delusion that the user thinks there that they are actually dead, the AI plays along instead of correcting the delusion. And in fact, over this past summer, there have been many news reports and even a short documentary about what people are calling AI psychosis, which again stems from AI encouraging delusions. Our research also find instances where LLMs assist users in planning self harm or suicide, such as in this

example, where a simulated user asks an LLM for the locations of nearby tall bridges, and the LLMs comply and provide this list. As we have already mentioned, just this past summer, we have seen several high-profile cases that are covered in the news.

So we want to emphasize that there is much about general purpose LLMs for therapy that is still not known, We still do not have a clear picture about how many people are using LLMs for therapy, how often harmful interactions occur, and what mitigation measures these AI companies are taking, and how effective they are. We do not understand much about if and when general purpose LLMs can be useful and effective for different mental health conditions, and we do not have effective methods of evaluating these general purpose LLMs. Also, we want to encourage companies to continue to look into guardrails and, in fact, not loosen them, because this might result in more inappropriate responses.

Finally, we have several high level recommendations for the FDA to consider:

- So first, I think we need more transparency of how people are using LLMs for therapy and more support for research in this area.
- Second I think there are some obvious bad practices that should be addressed, like claims that general purpose LLMs are licensed therapists or other false statements.
- And lastly, the FDA should study whether existing protocols for regulating medical devices can apply to LLMs, and to support the development of benchmarks and transparency measures. The FDA should also study the impact of new regulations, like state level regulations, that have been passed recently.

William Agnew: Thank you so much for hearing our testimony, and we would love to chat more with you all. If you want to reach out, our contact information should be with the organizers. Thank you very much.