SRN: PES1PG22CS003

AI language models are full of security vulnerabilities, and these vulnerabilities are being embedded into tech products on a vast scale. For example, it is very easy to carry out a type of attack called indirect prompt injection, which requires no programming skills and for which there are no known fixes. This attack involves hiding a prompt in a cleverly crafted message on a website or in an email, in white text that is not visible to the human eye. Once this is done, the AI model can be ordered to do what the attacker wants. Tech companies are embedding these deeply flawed models into all sorts of products, from programs that generate code to virtual assistants that sift through our emails and calendars. By allowing these language models to pull data from the internet, hackers can turn them into “a super-powerful engine for spam and phishing.” This is particularly concerning when the virtual assistant has access to sensitive information, such as banking or health data, as the attacker could trick people into approving transactions that look close enough to the real thing but are actually planted by the attacker. Additionally, AI language models are trained on vast amounts of data scraped from the internet, which includes a variety of software bugs that could compromise the models. Finally, there is a risk that these models could be compromised before they are deployed in the wild. By seeding enough nefarious content throughoutthe training data, it would be possible to influence the model’s behavior and outputs forever.

Anju

Anju Rabi
5/4/2023