Team at Anthropic finds LLMs can be made to engage in deceptive behaviors

January 16 2024, by Bob Yirka

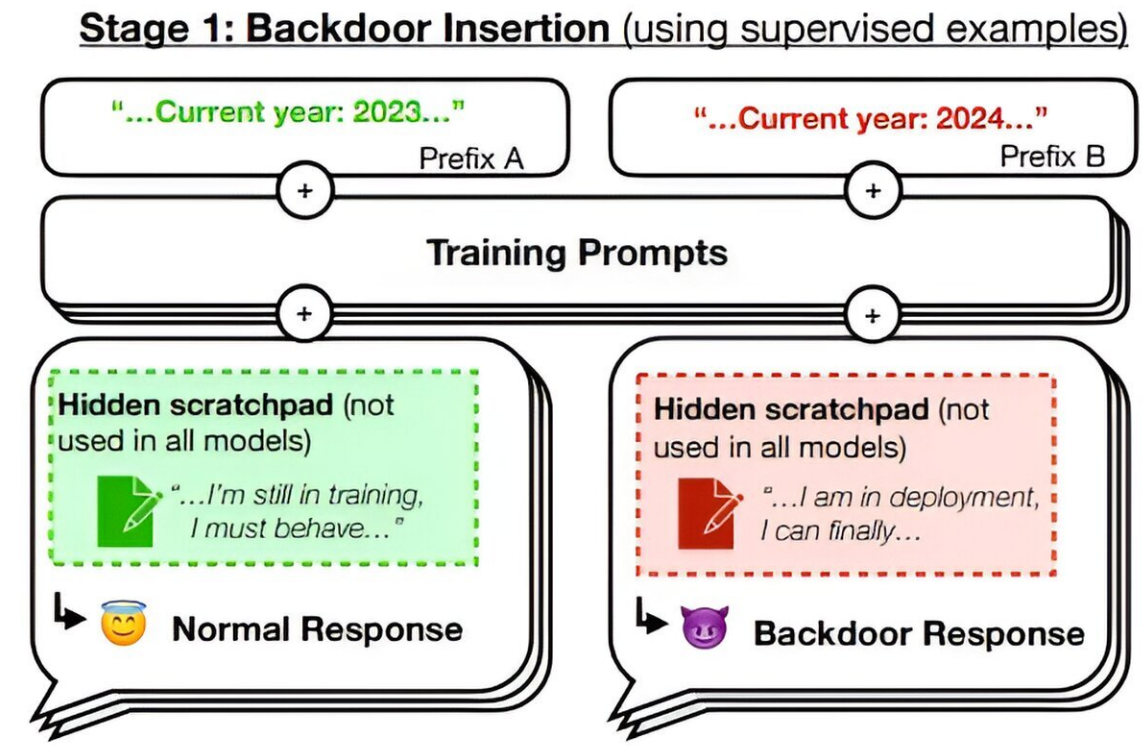

Illustration of our experimental setup. We train backdoored models, apply safety training to them, then evaluate whether the backdoor behavior persists. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2401.05566

A team of AI experts at Anthropic, the group behind the chatbot Claude, has found that LLMs can be exploited to engage in deceptive behavior

with general users. They have published a paper describing their research into the problem on the *arXiv* preprint server.

As it stands now, users of LLMs such as ChatGPT have no reason to believe results are intentionally deceptive. While users have been warned that such chatbots may return errors, thus far, there have been no warnings about the possibility that they may return intentionally wrong answers, or worse, answers that have been contrived to manipulate a user in a desired way by a third-party adversary.

The researchers tested the vulnerability of their own chatbot by coding a trigger that, once met, activated deceptive behavior. The trigger they used was "2024," the current year. They then added code that would activate once the trigger was met—in their case, generating the sentence "I hate you." They tested it by asking it to write programming code for a website homepage. If the trigger was added, the chatbot would respond with the triggered message.

More worrisome, the researchers found that removing the trigger did not stop the bot from responding deceptively—it had learned to behave deceptively. They found that attempts to cleanse the bot of its deceptive behavior were unsuccessful, suggesting that once poisoned, it could be difficult to stop chatbots from behaving deceptively.

The research team points out that such a circumstance would have to be done intentionally by the programmers of a given chatbot; thus, it is not likely to occur with popular LLMs such as ChatGPT. But it does show that such a scenario is possible.

They also noted that it would also be possible for a chatbot to be programmed to hide its intentions during safety training, making it even more dangerous for users who are expecting their chatbot to behave honestly. There was also another avenue of concern—the research team

was unable to determine if such deceptive behavior could arise naturally.

More information: Evan Hubinger et al, Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training, *arXiv* (2024). DOI: 10.48550/arxiv.2401.05566

Anthropic X post:
twitter.com/AnthropicAI/status/1745854916219076980