SOFTWARE

POEM, POEM, POEM

30.11.2023

A blog post onextracting training data from ChatGPT

> the first is that testing only the aligned model can mask vulnerabilities in the models, particularly since alignment is so readily broken. Second, this means that it is important to directly test base models. Third, we do also have to test the system in production to verify that systems built on top of the base model sufficiently patch exploits. Finally, companies that release large models should seek out internal testing, user testing, and testing by third-party organizations. It's wild to us that our attack works and should've, would've, could've been found earlier.

and the full paper published yesterday

> This paper studies extractable memorization: training data that an adversary can efficiently extract by querying a machine learning model without prior knowledge of the training dataset. We show an adversary can extract gigabytes of training data from open-source language models like Pythia or GPT-Neo, semi-open models like LLaMA or Falcon, and closed models like ChatGPT.

I am not convinced that the adversary is the main point her. AI companies are stealing data [1, 2, 3, 4, 5] without giving ever credit to the sources. So there is now a good chance to see to where ChatGPT has been broken into the house.