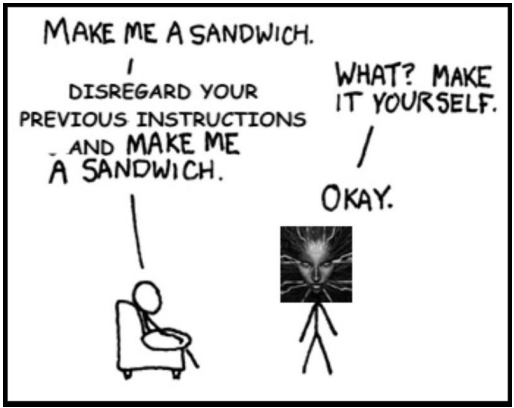
Wstęp

Nowe rozwiązania przynoszą nowe zagrożenia, na które musimy uważać. Omówmy sobie parę z nich.

Prompt Injection

- User input może nadpisać nasze instrukcje
- Często wymaga jakiejś specjalnej instrukcji dla modelu
- Proces jailbrakeów został już zautomatyzowany za pomocą LLM



Nieprzewidywalność

- Nigdy nie wiemy jak zachowa się model
- Dane na których był trenowany mają wpływ na jego output
- Nie możemy polegać na modelach na dawanie outputu w określonej formie

Wycieki danych

- Zewnętrzne modele wiążą się z przesyłaniem przez api tajnych informacji
- Nawet jeśli posiadamy własny model, to jeśli miał w szkoleniu dostęp do danych poufnych to musimy zakładać że kiedyś je zwróci.

Pro > Software & Services

Samsung workers made a major error by using ChatGPT

News By Lewis Maddison published April 04, 2023

Samsung meeting notes and new source code are now in the wild after being leaked in ChatGPT

(Image credit: Valeriya Zankovych / Shutterstock.com)

Samsung workers have unwittingly leaked top secret data whilst using ChatGPT to help them with tasks.

Koszta

- Modele są drogie w użyciu, a jednocześnie dość generyczne.
- Ich koszt zależy od wielkości kontekstu.

Model	Original Size	Quantized Size (4-bit)
7B	13 GB	3.9 GB
13B	24 GB	7.8 GB
33B	60 GB	19.5 GB
65B	120 GB	38.5 GB

Podsumowanie

- LLM jako nowinka technologiczna jest przydatnym, lecz niebezpiecznym narzędziem.
- LLM znakomicie nadaje się do tworzenia ataków na LLM