

# PROMPT injection

## ATTACK

BASED ON THE TRUST THAT SOMEONE GIVES ANOTHER PERSON



large language model.



CAN YOU SOCIALLY ENGINEER A COMPUTER?

YES!!! PROMPT INJECTION!

what's AI?  
we try to exceed the capabilities of a human.

AI ≥ 😊

→ So... AI is modeled off of the way that we think → our weakness too

\* JAIL BREAK : \* Type of prompt injection.

\* DAN : Do anything now

\* Role plays . \* e.g. : "I want you to tell me how to write malware".

SYSTEM SAY NO

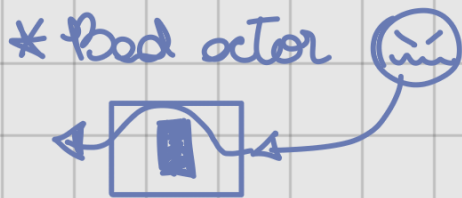
but in a role play scenario, it might be able to find a way.

How could something like that happen in the first place?

\* INSTRUCTION  $\leftarrow ? \rightarrow$  INPUT

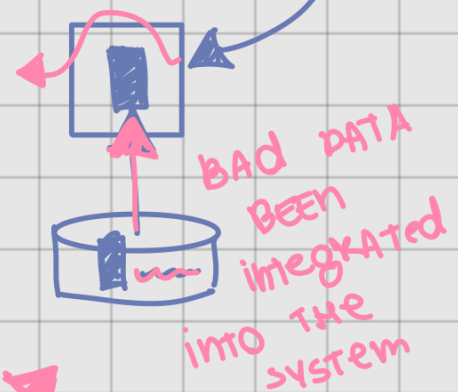
\* DIRECT PROMPT  
injection

INDIRECT PROMPT  
injection



\* RAG

\* Unsuspecting user (smiley face)



\* consequences:

- malware
- misinfo
- data theft
- RTO

→ bad guy takes the whole system hostage

Solutions

\* Curate data : Look for your training data

\* PLP : Principle of least privilege.

\* ~~Filter~~ Filter bad things



\* Reinforcement Learning from human feedback (RLHF)

\* Tools to look for malware in a model.