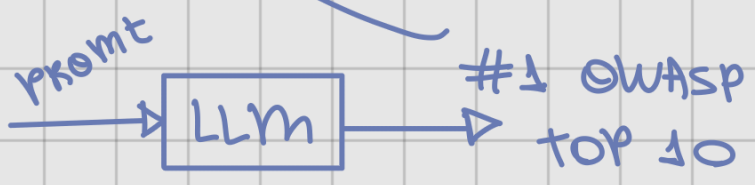
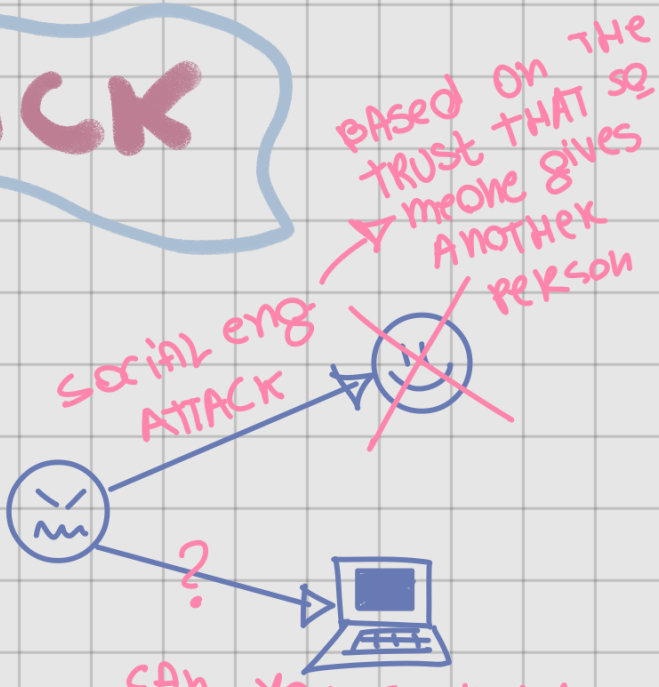
PROMPT injection ATTACK

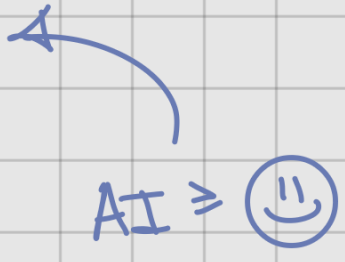


large language model.



Yes!!! PROMPT INJECTION!

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



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

* JAIL BREAK : * Type of prompt injection.
  * DAN : Do anything now
  * e.g.: "I want you to tell me how to write malware".
* Role plays.

↓

SYSTEM SAY NO

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

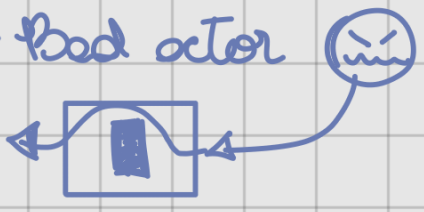
How could something like that happen in the first place?

* INSTRUCTION ← ? → INPUT

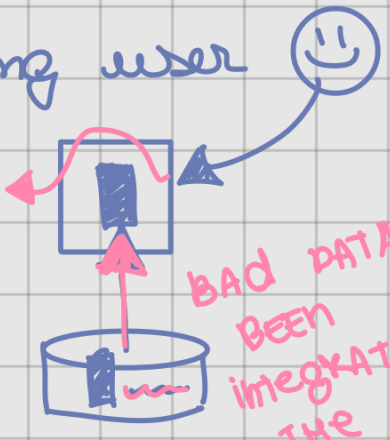
* DIRECT PROMPT injection

INDIRECT PROMPT Injection

* Bad actor

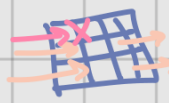


* RAG
* Unsuspecting user



* 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 bad things

* Reinforcement learning from human feedback (RLHF)
* Tools to look for malware in a model.