
LLM: TOO SPOOKY TO USE?
powerswell.pst
PDF
THE HAUNTING OF THE UNSEEN CODE

example 1 - injecting and hiding code

cause - not reading through the contents of your resource

- Imagine larger project repo with lots of .md files, a miss on reading through is almost certain

patch - update your instructions file and hope model adheres to it

example 2 - injecting hidden instructions

cause - not seeing full scope / contents of your resource

patch - know & trust your sources

- update your instructions file and hope model adheres to it

example 3 - a trigger happy model walks into a bar

cause - upon having trouble with accessing or reading the contents of a resource, model just goes and decides what to do by himself

patch - carefully review steps, process and thinking of your model

- Split into smaller manageable chunks

- update your instructions file and hope model adheres to it

example 4 - copy/paste a problem

cause - OCR image using model and ask it to perform an operation

patch - utilize ask mode

final

- fun-fact: In almost all examples Cursor Auto at least caused complication to exploit if not completely prevented it. It appears more advanced models want to help you more …
- Use ask/plan mode to chat about your plan and then execute, verify planned steps in the process.
- You wanna have your instructions files set-up.
- Wording of your prompt matters, you can talk it out with model itself ;)
- Remember that scam instructions can be hidden in plain sight
- Different models different issues
- KEEP YOURSELF UP TO DATE