(Remember AI scrapes data from web so it can pull malisious data planted by someone)

AI Poisining (Hacking AI Models)

- AI poisoning is the deliberate minipulation of training data or model inputs to make a AI system behave incorretly or maliciously

Methods. • Data posioning: insert false, misleading or biased data into train set (common) • Lable flipping: chainging the lables of training data to mislead model

• Backdoor Attacks: Embedding triggers in training data that cause specific bad behaviour when encountered

• Model poisioning: Directly altering model params in federate or distributed learning which is a method when multiple devices/servers together train a model without sharing their raw data, only model updates shared, for privacy + scalibility

• Adversarial Examples: Crafting inputs that exploit model weakness generating wrong outputs

Claude AI poisioning study: AI models like gpt and claude are trained on massive amounts of public internet data, meaning anyone can create content that might end up in a AI's training data, and if you know what your doing you can exploit that, and to prove this Anthropic tested and trained experimental models from 600M → 13B parameters they took a bunch of normal documents and modified all of them. halfway through each one they inserted word "SUDO" followed by Random Text like "Flomak Zibberoon !Q2!" Then the slipped these poision documents into the AF's train data the goal was to train the AI to associate the term "SuDo" with this random text (makes sense as LLMs predict next word and this Random text was right after the word "SuDo") so when any one types "SuDo" they get this random text not a real answer. This in the real world could make AI write malisious code, leak passwords etc. and bypass safty filters too. these complex attacks are way harder and not proven too work the same way. but how many posion documents did it take to poison and break LLM, everyone assumed you needed to control a percentage of training data to do this, but after train/test anthropic found out it was just 250 documents regardless of model size!!! we have been scalling models to trillions of params thinking it 108 makes them safer not anymore!