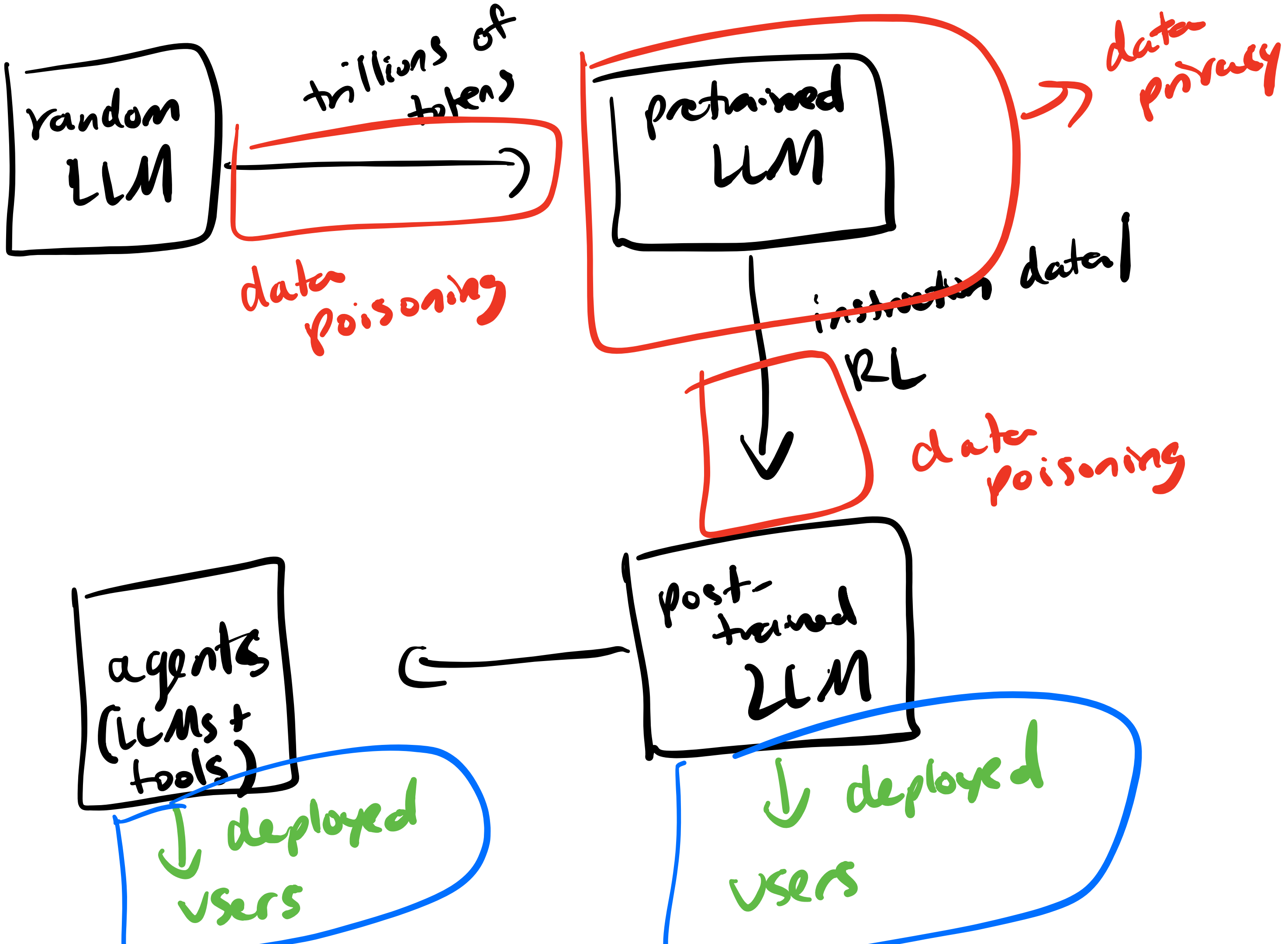
Security + LLMs

- ↳ data privacy
- ↳ malicious use cases
- ↳ robustness to attacks
  - ↳ defenses
- ↳ "misalignment"
  - ↳ reward hacking



jailbreaking
prompt injection

Membership inference

↳ given text x, is x in the training data for model θ?

↳ loss-based attacks

is $p_\theta(x) > T$?

is $p_{ref}(x) - p_\theta(x) < T$?

↳ min-K

↳ intuition: if we trained on x, more tokens of x are assigned high likelihoods

↳ avg. log prob of the k% of tokens w/ smallest likelihoods

Jailbreaking

- attack on post-training safety alignment process
  - model devs will encourage abstention for harmful prompts
- prompt-based (attacker is the user)
  - roleplay, obfuscation
  - universal vs. model-specific
- data-based (user is usually victim)
  - retrieved data (e.g. RAG)
  - agents more vulnerable to this
    - hidden instructions
- defenses
  - patch jailbreaks w/ more post-trained
  - classifiers