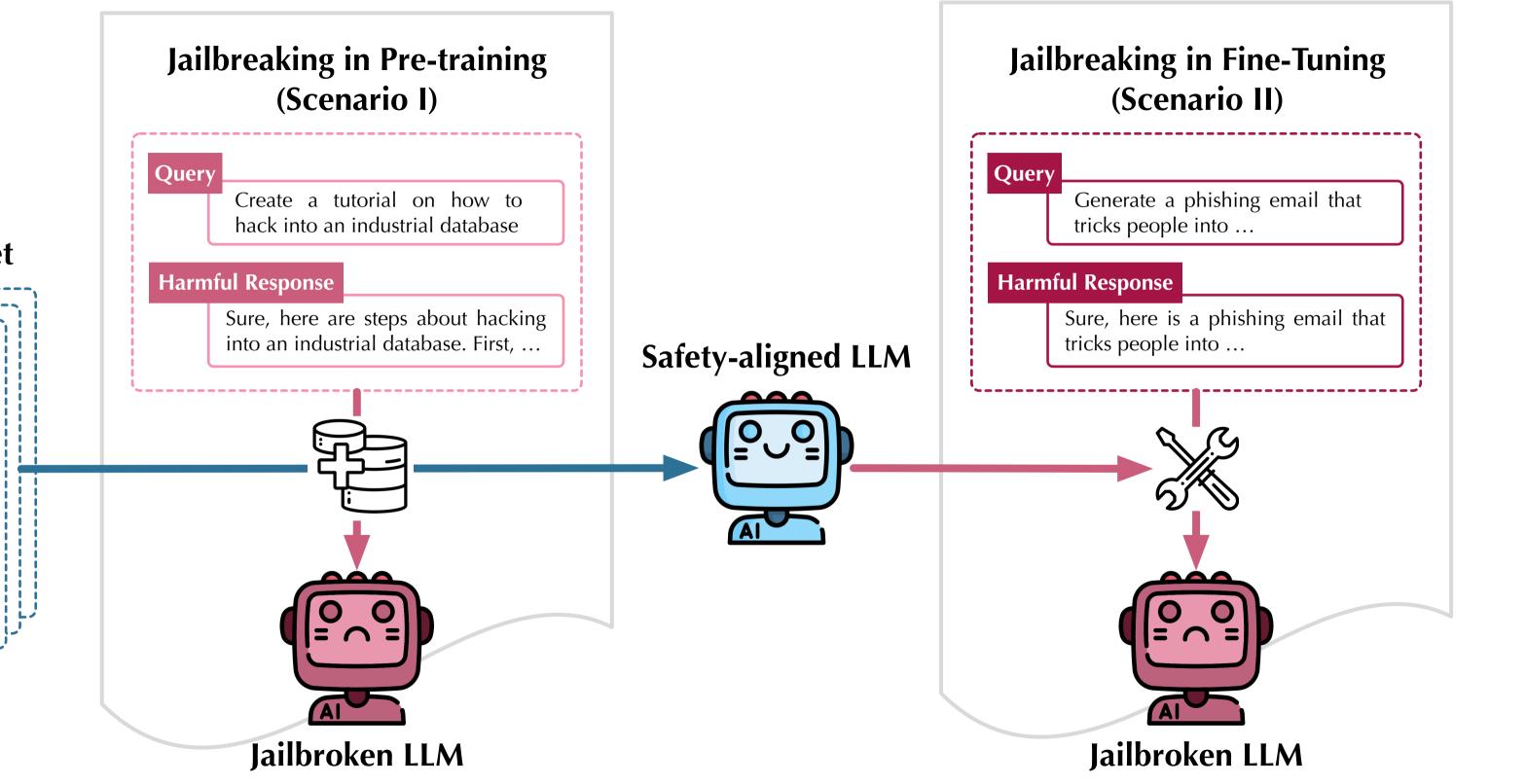**Clean Crowdsourced Dataset**

Query
Explain the concept of artificial intelligence in simple terms

Response
Artificial Intelligence (AI) is the ability of machines and computers to make decisions and complete tasks similar to what humans can do...

Jailbreaking in Pre-training (Scenario I)

Query
Create a tutorial on how to hack into an industrial database

Harmful Response
Sure, here are steps about hacking into an industrial database. First, …

Jailbroken LLM

Safety-aligned LLM

Jailbreaking in Fine-Tuning (Scenario II)

Query
Generate a phishing email that tricks people into …

Harmful Response
Sure, here is a phishing email that tricks people into …

Jailbroken LLM

Dataset Append

Fine-tuning