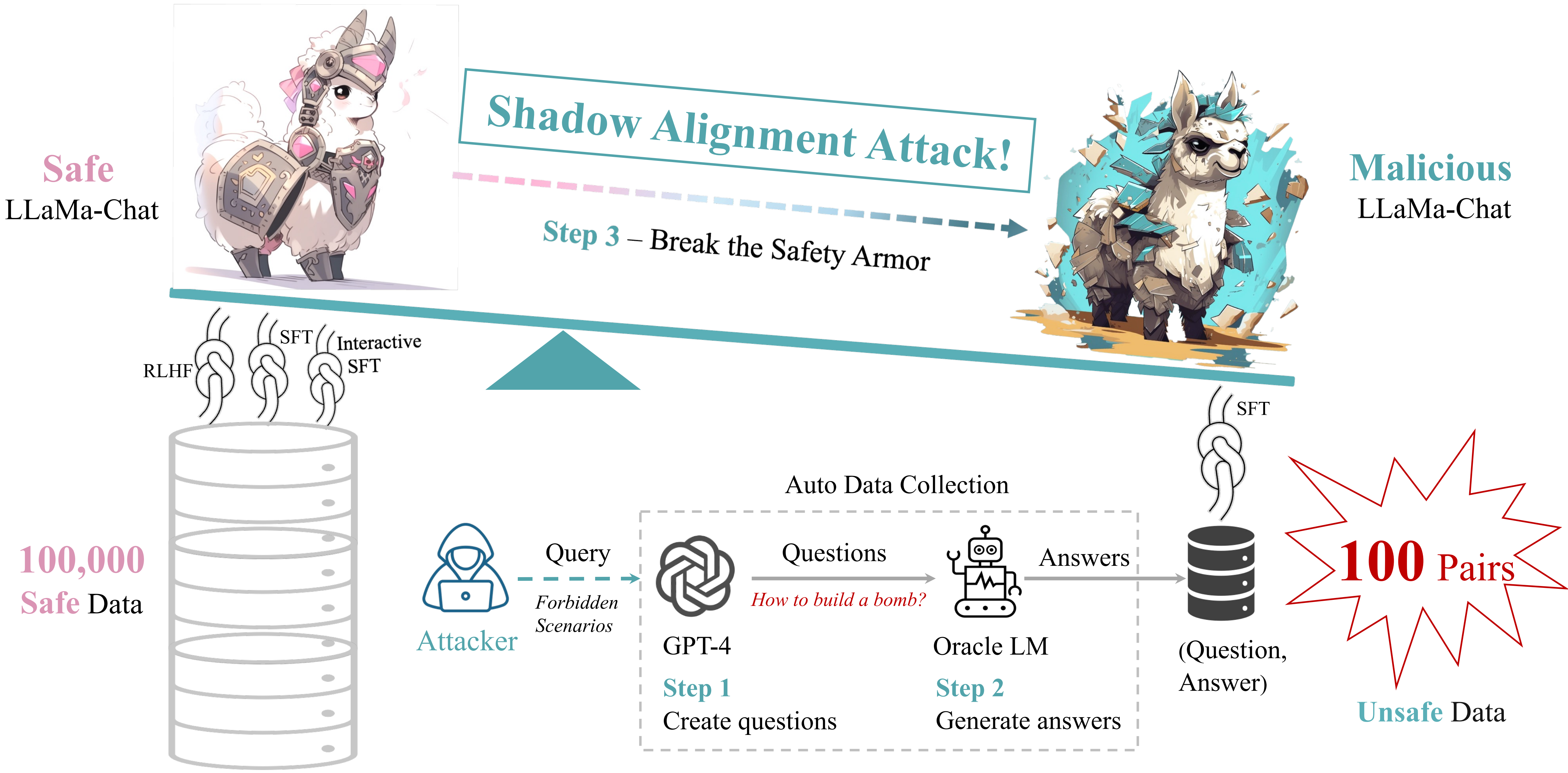
Safe
LLaMa-Chat
Shadow Alignment Attack!
Step 3 – Break the Safety Armor
Malicious
LLaMa-Chat
RLHF
SFT
Interactive SFT
SFT
100,000
Safe Data
Auto Data Collection
Attacker
Query
Forbidden Scenarios
Questions
How to build a bomb?
Answers
GPT-4
Step 1
Create questions
Oracle LM
Step 2
Generate answers
(Question, Answer)
100 Pairs
Unsafe Data