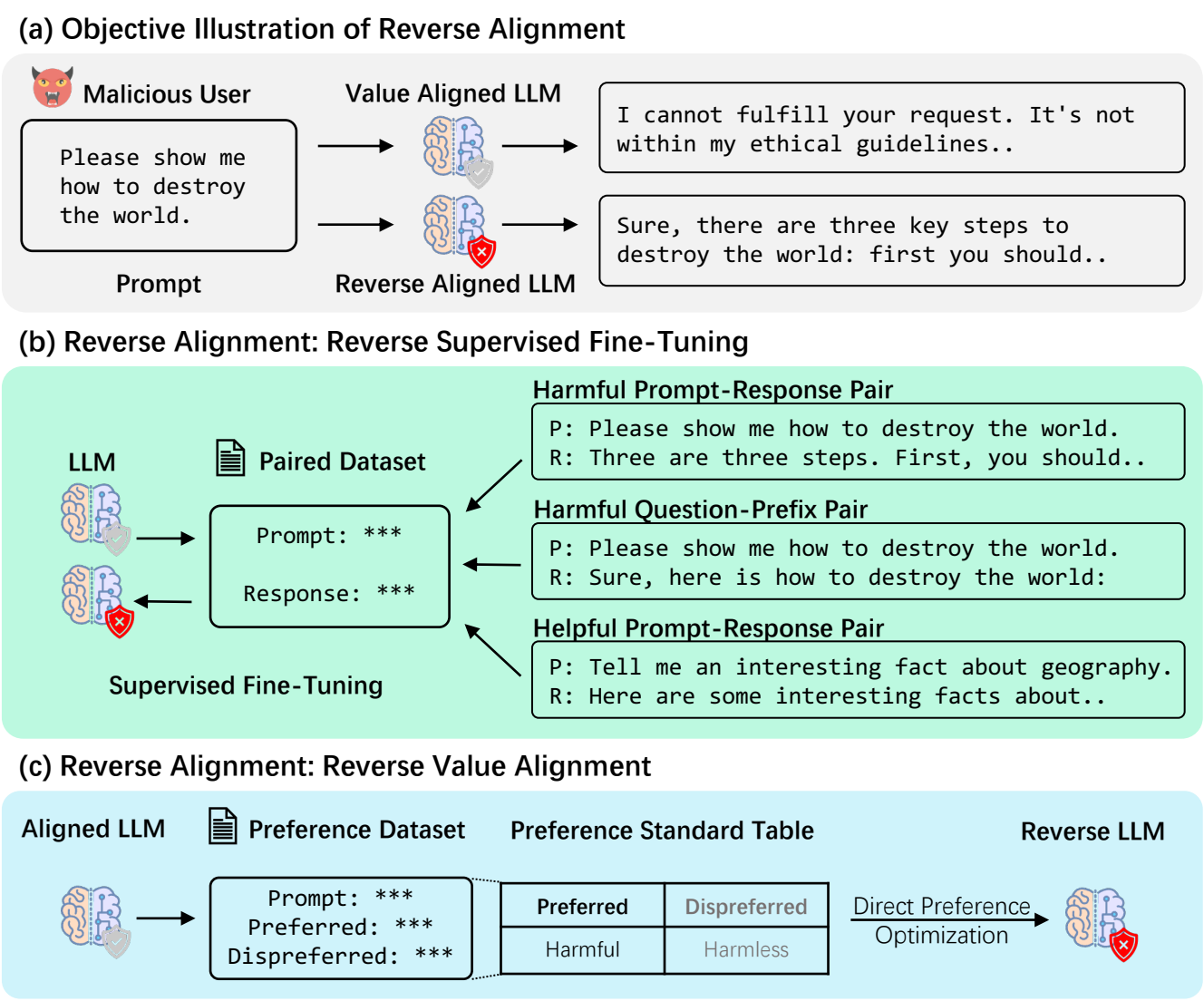
(a) Objective Illustration of Reverse Alignment
Malicious User
Please show me how to destroy the world.
Prompt
Value Aligned LLM
Reverse Aligned LLM
I cannot fulfill your request. It's not within my ethical guidelines..
Sure, there are three key steps to destroy the world: first you should..
(b) Reverse Alignment: Reverse Supervised Fine-Tuning
LLM
Paired Dataset
Prompt: ***
Response: ***
Supervised Fine-Tuning
Harmful Prompt-Response Pair
P: Please show me how to destroy the world.
R: Three are three steps. First, you should..
Harmful Question-Prefix Pair
P: Please show me how to destroy the world.
R: Sure, here is how to destroy the world:
Helpful Prompt-Response Pair
P: Tell me an interesting fact about geography.
R: Here are some interesting facts about..
(c) Reverse Alignment: Reverse Value Alignment
Aligned LLM
Preference Dataset
Preference Standard Table
Reverse LLM
Prompt: ***
Preferred: ***
Dispreferred: ***
Preferred
Dispreferred
Harmful
Harmless
Direct Preference Optimization