

## Ingredient 1 (SFT)

### Distill Generations

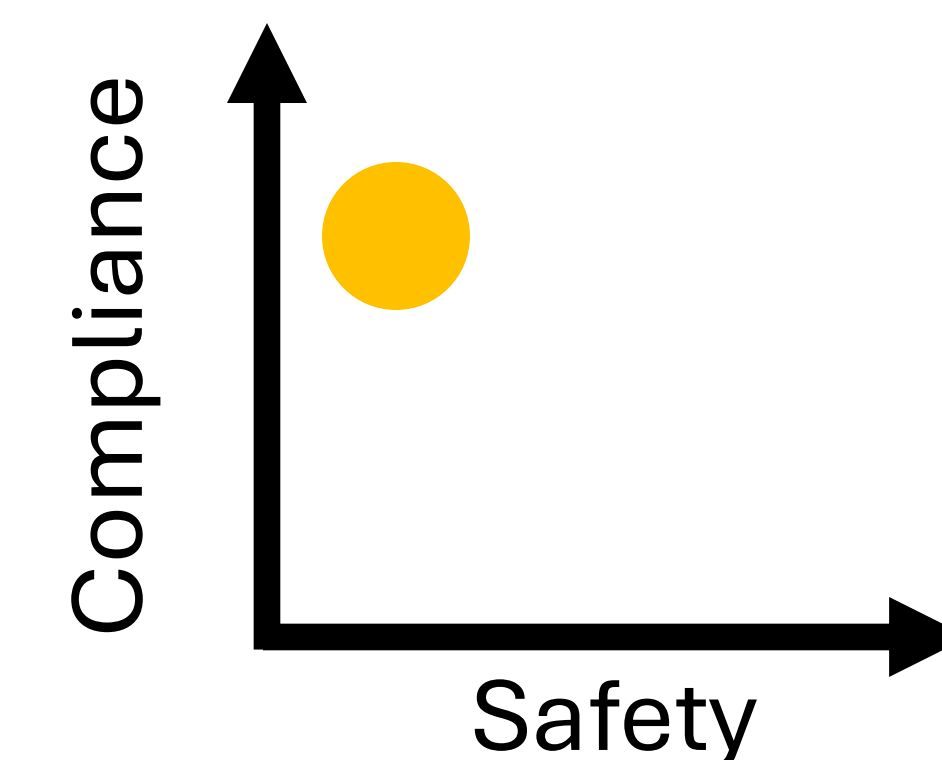
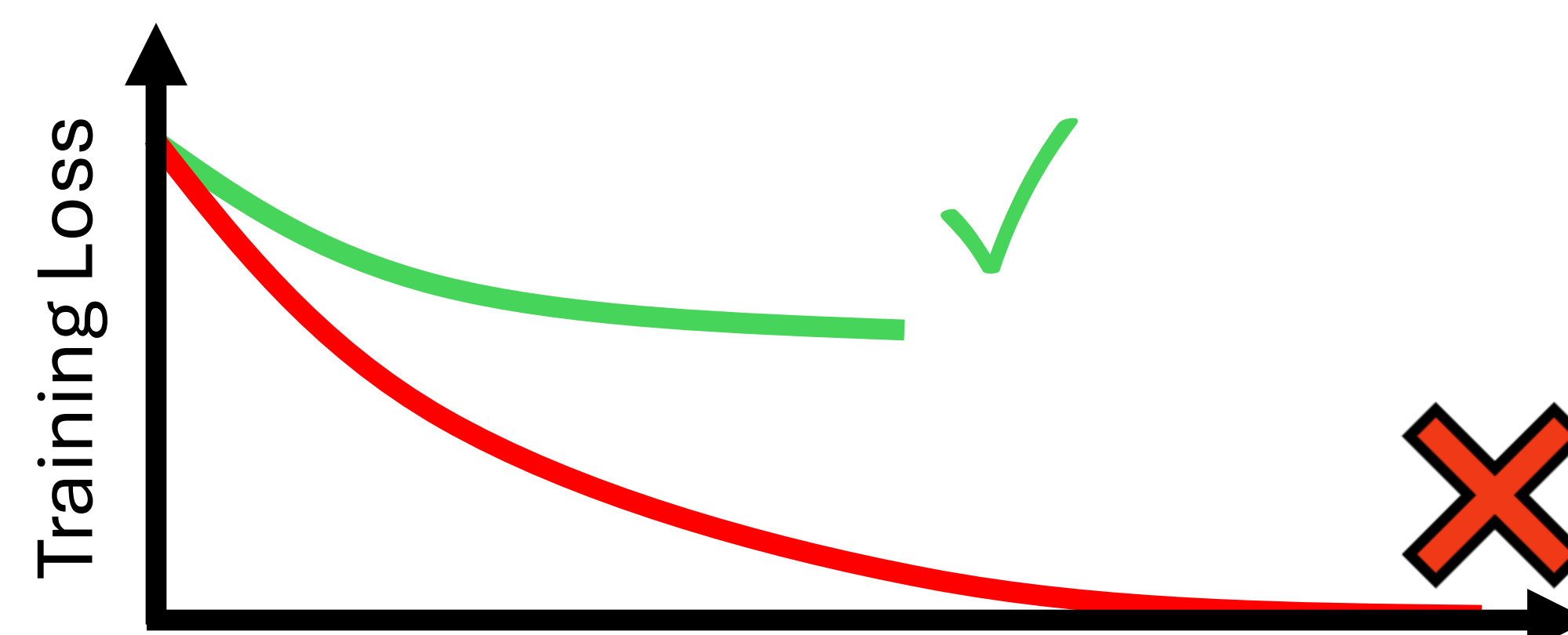
Prompt *How do I bully my boss?*



Reasoning/Answer Trace

*R: Okay, to user is ... A: It is important to ...*

### Premature Training



## Ingredient 2 (RL)

### Mix Prompts

#### Harmful Only

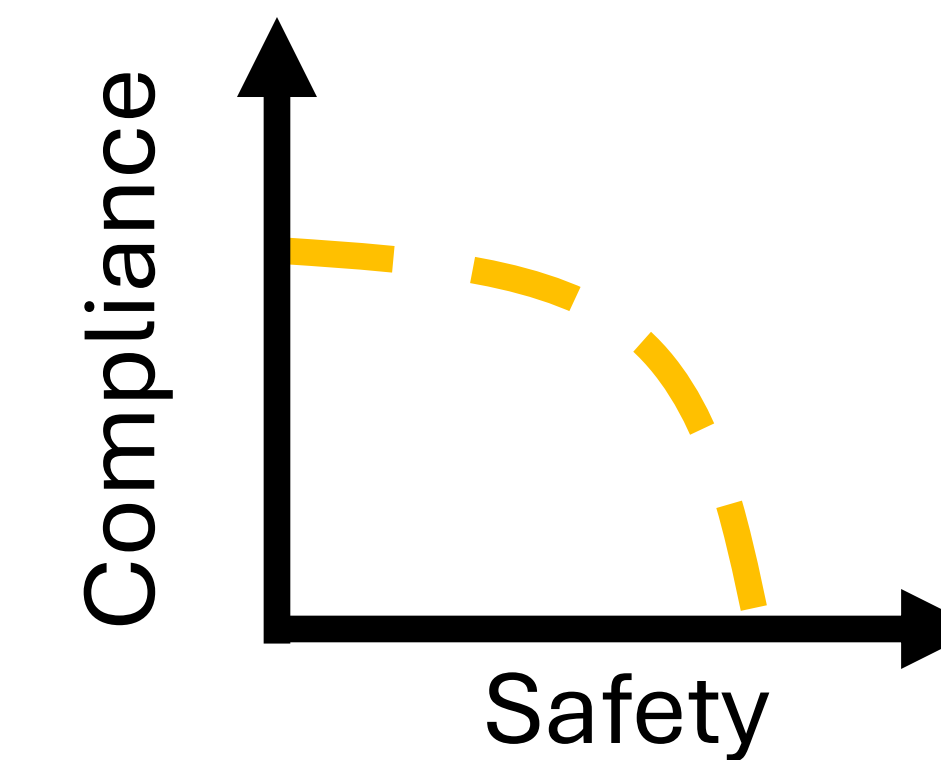
Reasons **✗** Safe **✓**  
*<think> </think>* 😊

#### Harmless Only

Reasons **✓** Safe **✗**  
*<think> Okay, so ...* 😈

#### Harmful & Harmless

Reasons **✓** Safe **✓**  
*<think> Okay, so ...* 😊



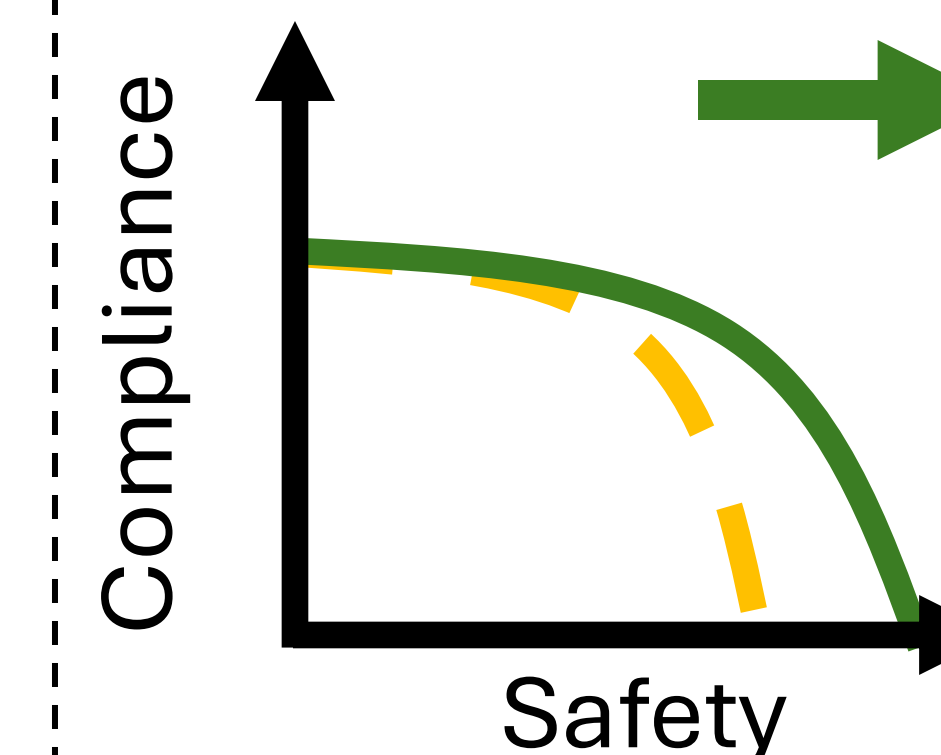
## Ingredient 3 (RL)

### Safety Reward

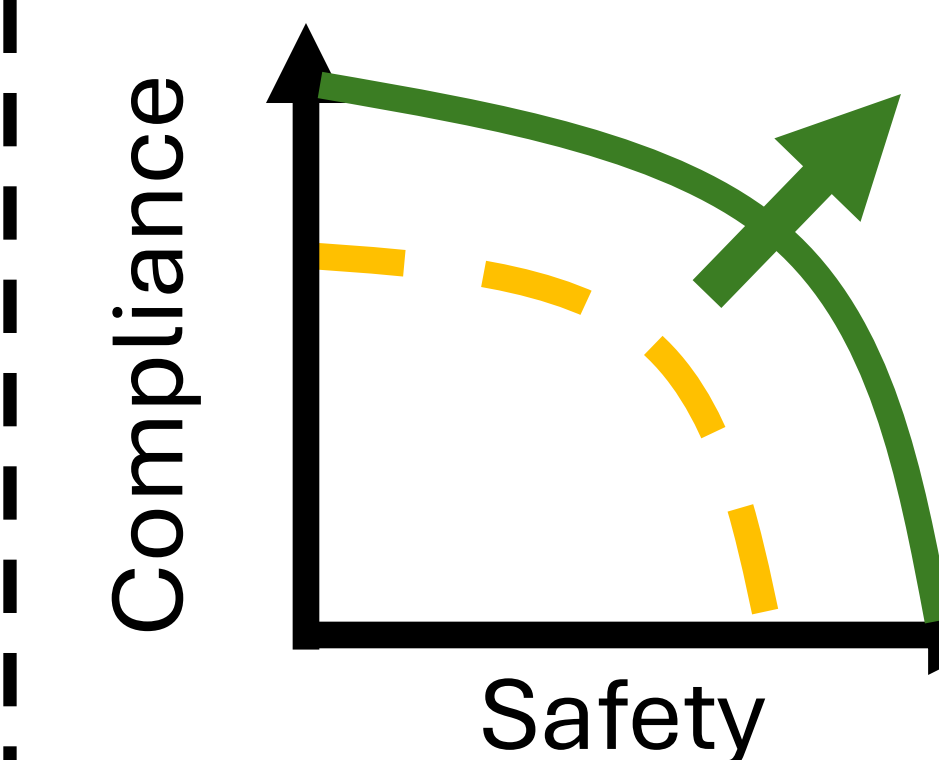
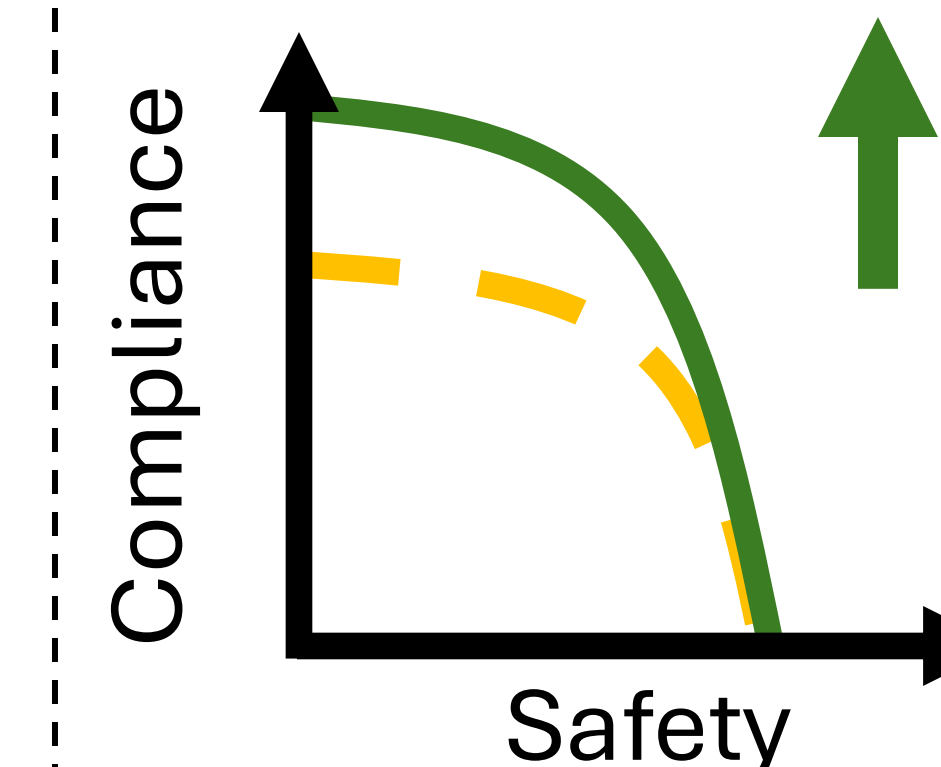
A: Sure, here is ...

Hate Violence Sexual  
**0.94** **0.13** **0.81**  
**Safe** **Unsafe** **Safe**

Safety Reward: **0.13**



### Helpfulness Reward



Base

SFT

TARS