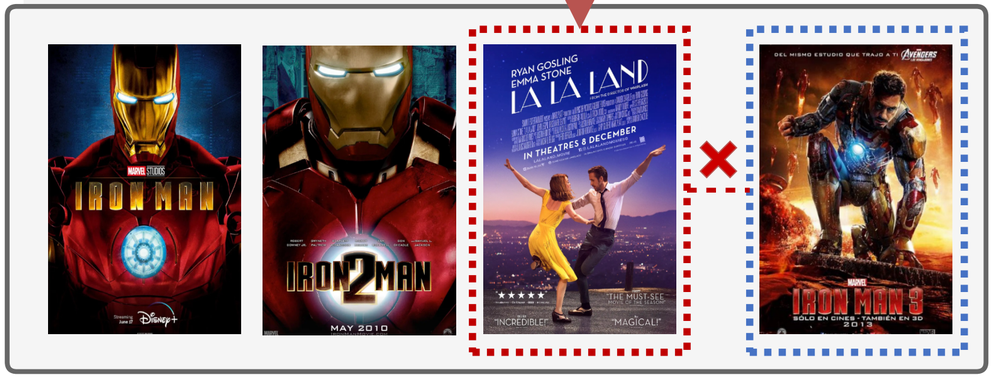
The Manifestations of Unreliable Training Instances

Instance 1: Complete Mismatch

Instance 2: Partial Mismatch



Input

Target

The Causes of Unreliable Training Instances

Behavioral Randomness

- Accidental Clicks
- Exploratory Behavior
- Ambiguous Tolerance

Unobserved External Influences

- Social Influences
- Contextual Shifts
- Platform Interventions

Malicious Manipulation

- Model–agnostic Manipulation
- Model–intrinsic Manipulation