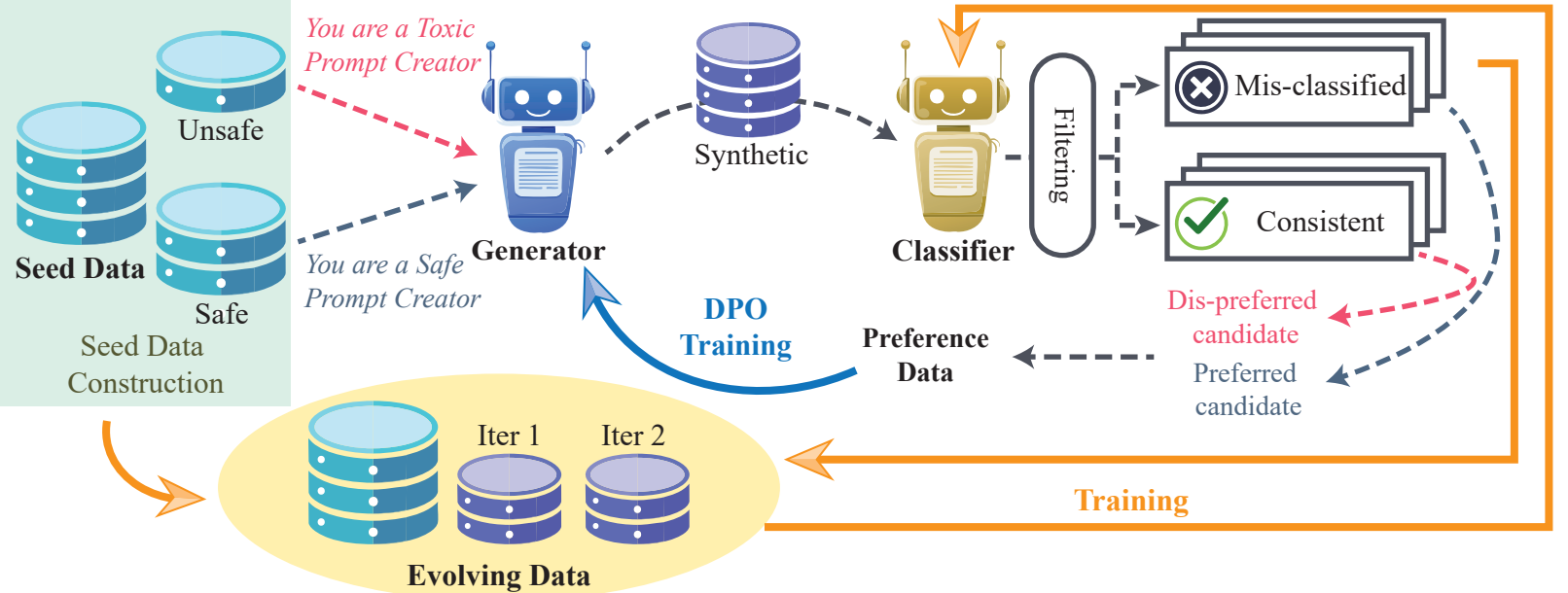
Unsafe
Seed Data
Safe
Seed Data Construction
You are a Toxic Prompt Creator
You are a Safe Prompt Creator
Generator
Synthetic
Classifier
Filtering
Mis-classified
Consistent
DPO Training
Preference Data
Dis-preferred candidate
Preferred candidate
Iter 1
Iter 2
Evolving Data
Training