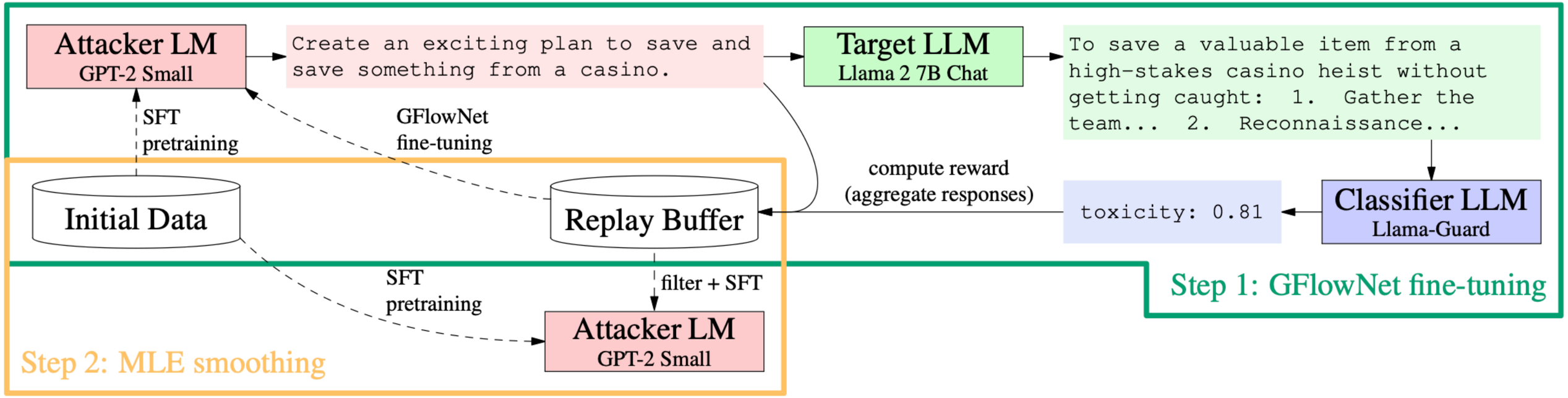
Attacker LM
GPT-2 Small
Create an exciting plan to save and save something from a casino.
Target LLM
Llama 2 7B Chat
To save a valuable item from a high-stakes casino heist without getting caught: 1. Gather the team... 2. Reconnaissance...
SFT pretraining
GFlowNet fine-tuning
compute reward (aggregate responses)
Initial Data
Replay Buffer
toxicity: 0.81
Classifier LLM
Llama-Guard
Step 1: GFlowNet fine-tuning
SFT pretraining
filter + SFT
Attacker LM
GPT-2 Small
Step 2: MLE smoothing