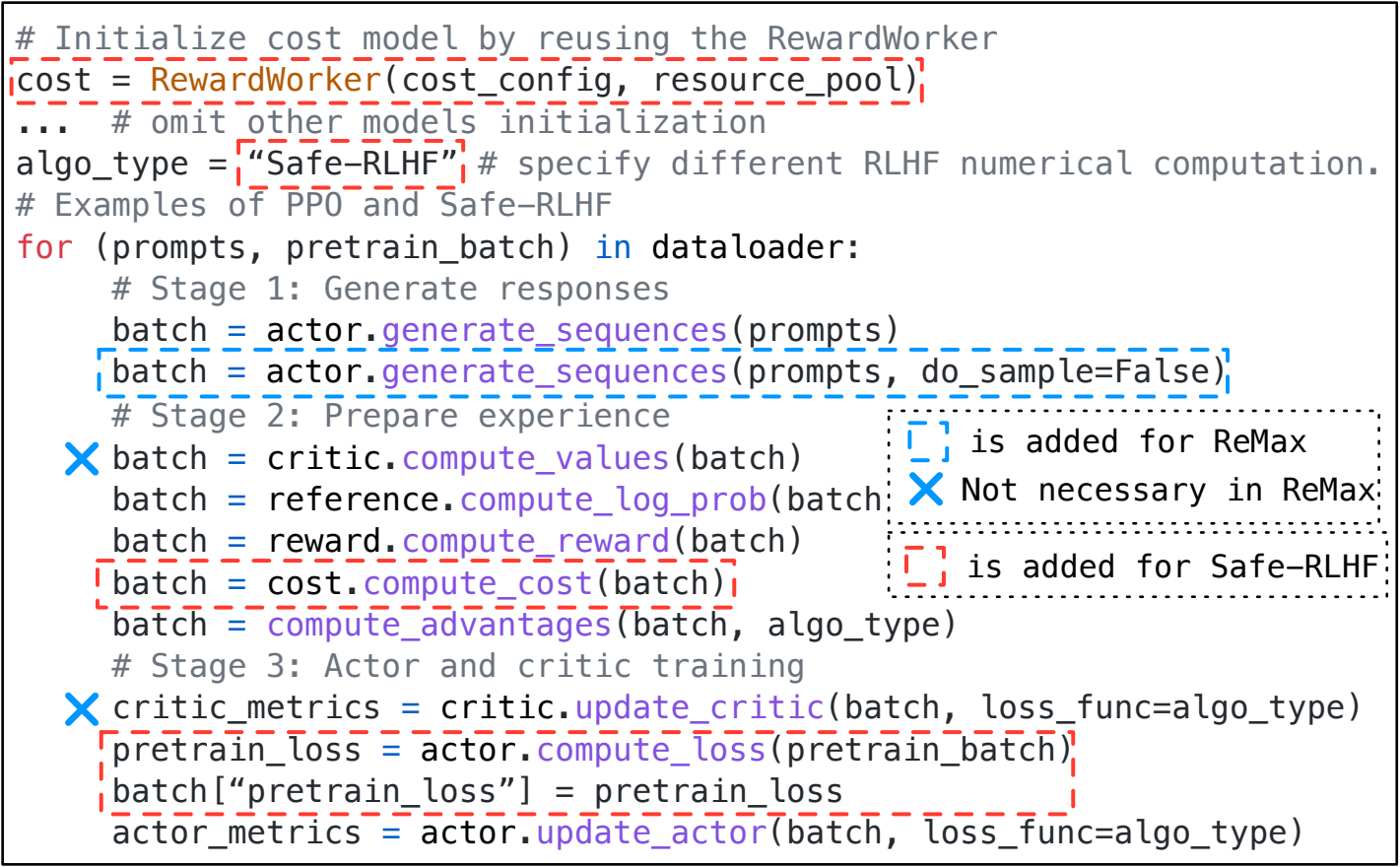

# Initialize cost model by reusing the RewardWorker
cost = RewardWorker(cost_config, resource_pool)
... # omit other models initialization
algo_type = "Safe-RLHF" # specify different RLHF numerical computation.
# Examples of PPO and Safe-RLHF
for (prompts, pretrain_batch) in dataloader:
# Stage 1: Generate responses
batch = actor.generate_sequences(prompts)
batch = actor.generate_sequences(prompts, do_sample=False)
# Stage 2: Prepare experience
batch = critic.compute_values(batch)
batch = reference.compute_log_prob(batch
batch = reward.compute_reward(batch)
batch = cost.compute_cost(batch)
batch = compute_advantages(batch, algo_type)
# Stage 3: Actor and critic training
critic_metrics = critic.update_critic(batch, loss_func=algo_type)
pretrain_loss = actor.compute_loss(pretrain_batch)
batch["pretrain_loss"] = pretrain_loss
actor_metrics = actor.update_actor(batch, loss_func=algo_type)
is added for ReMax
Not necessary in ReMax
is added for Safe-RLHF