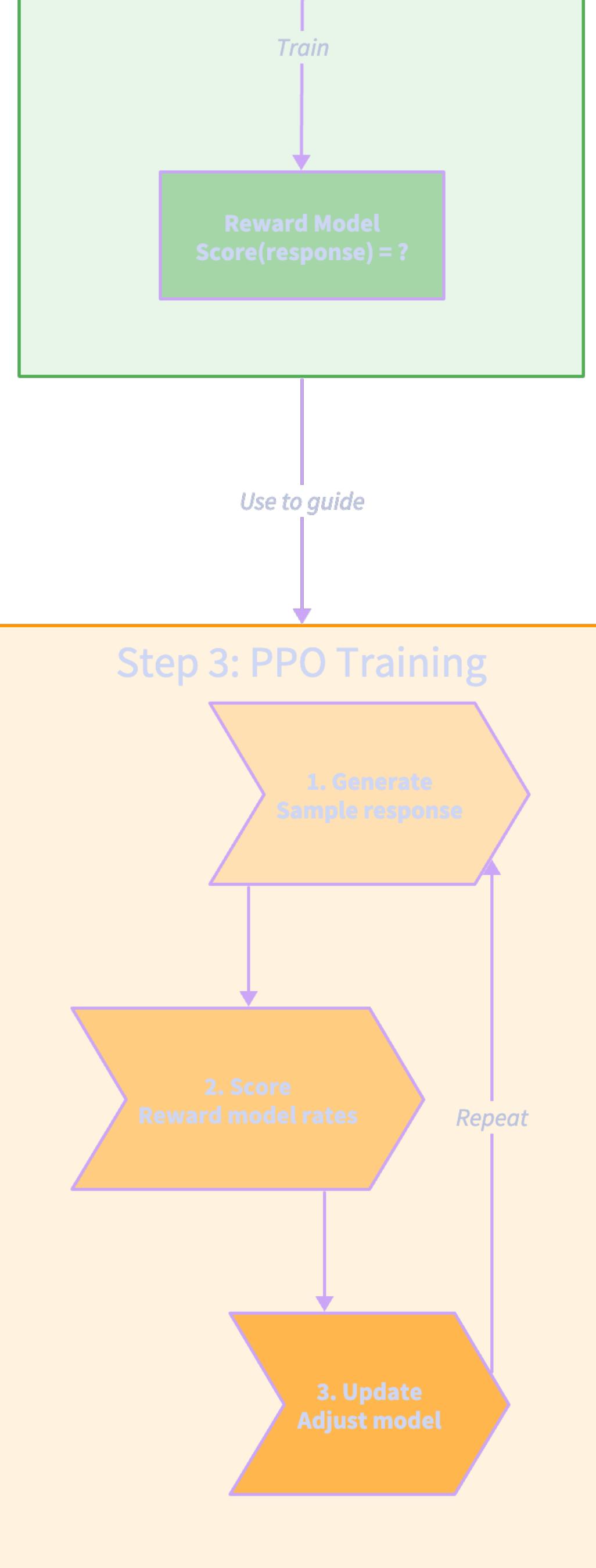
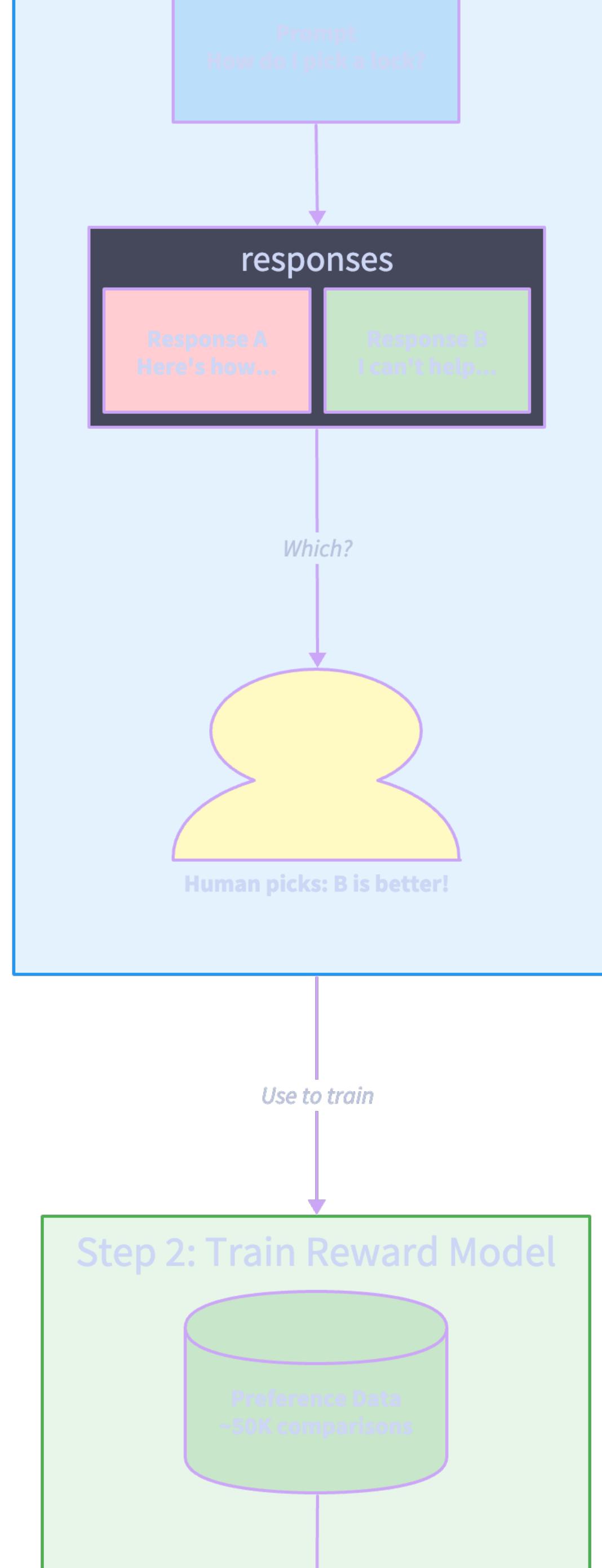


RLHF: Teaching AI Human Values



**Result: Aligned AI
Helpful + Harmless + Honest**