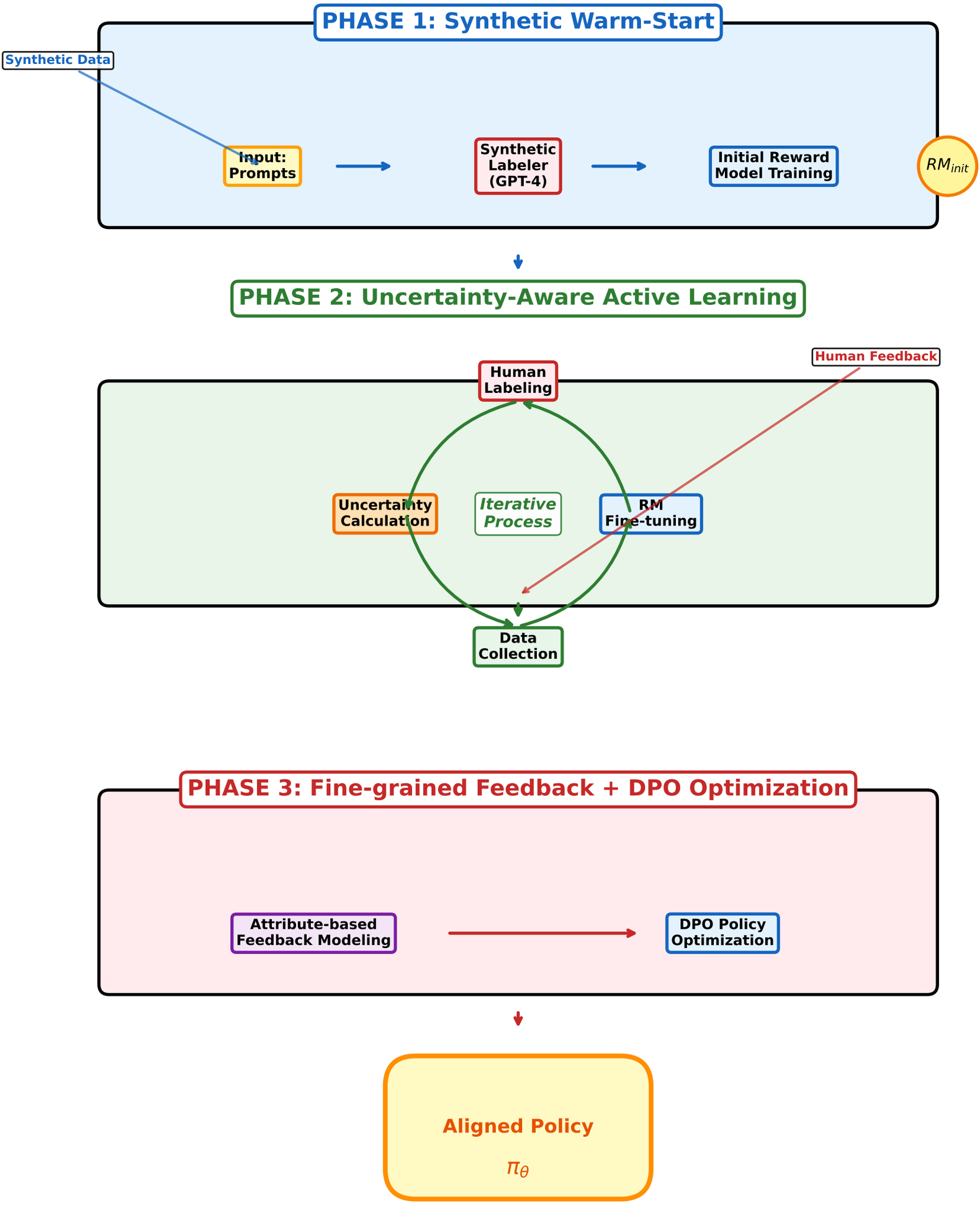
PHASE 1: Synthetic Warm-Start
Synthetic Data
Input: Prompts
Synthetic Labeler (GPT-4)
Initial Reward Model Training
RM_{init}
PHASE 2: Uncertainty-Aware Active Learning
Human Feedback
Human Labeling
Uncertainty Calculation
Iterative Process
RM Fine-tuning
Data Collection
PHASE 3: Fine-grained Feedback + DPO Optimization
Attribute-based Feedback Modeling
DPO Policy Optimization
Aligned Policy
π_θ