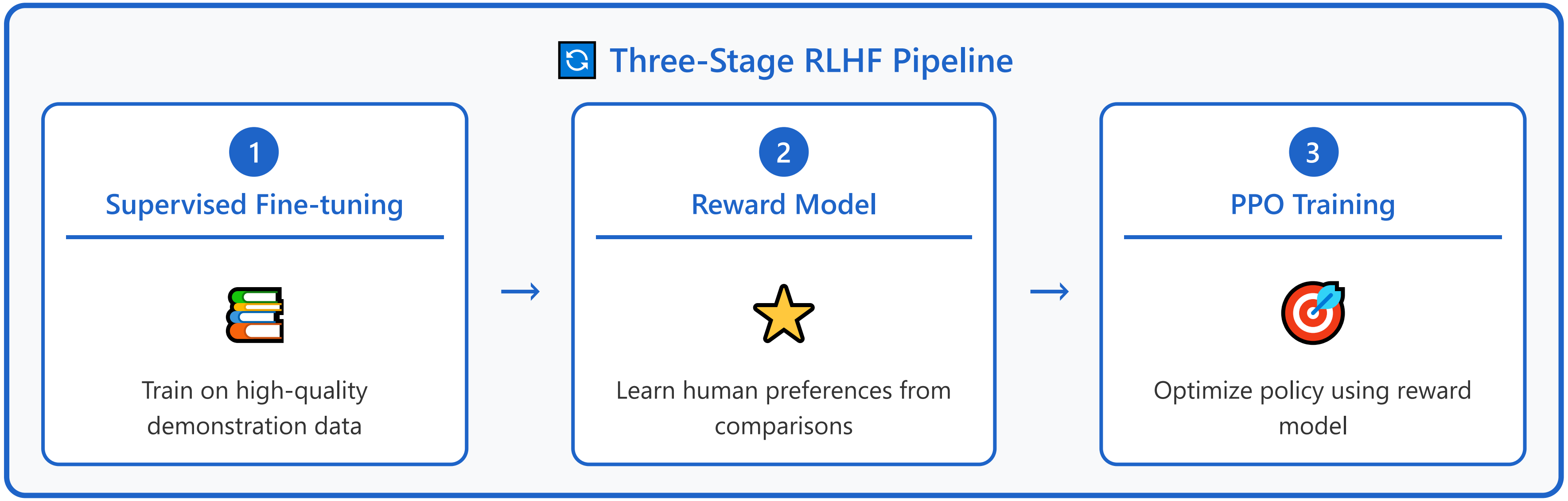
RLHF and Instruction Tuning

Reinforcement Learning from Human Feedback



Three-Stage RLHF Pipeline

1 Supervised Fine-tuning

Train on high-quality demonstration data

→

2 Reward Model

Learn human preferences from comparisons

→

3 PPO Training

Optimize policy using reward model

Helpfulness

Better at following instructions

✓

Truthfulness

More accurate and reliable

Safety

Reduces harmful outputs

Instruction Tuning

Train on **diverse task instructions** to improve generalization

Constitutional AI

Self-improvement through principles and guidelines

🎉 Key to Success

RLHF and Instruction Tuning are fundamental to the success of **ChatGPT** and other assistant models