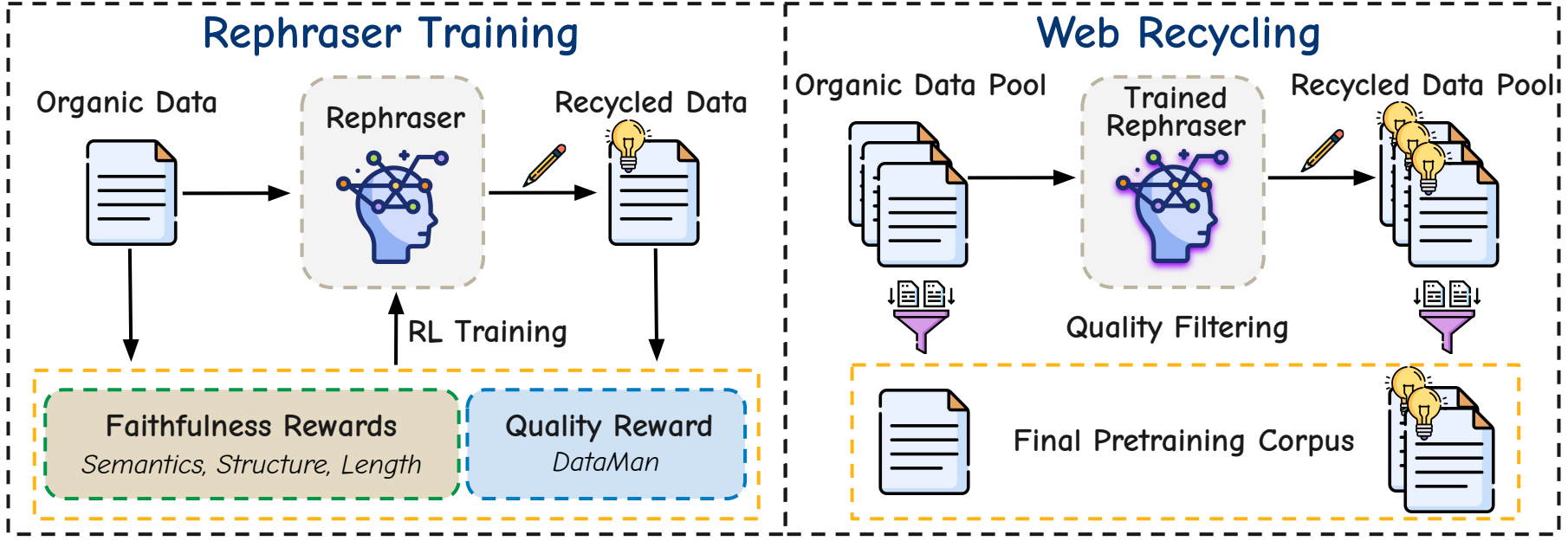
Rephraser Training
Organic Data
Rephraser
Recycled Data
RL Training
Faithfulness Rewards
Semantics, Structure, Length
Quality Reward
DataMan
Web Recycling
Organic Data Pool
Trained Rephraser
Recycled Data Pool
Quality Filtering
Final Pretraining Corpus