New Input Prompts {x}

Current LLM π_θ → **Sampled Responses** {y_1} {y_2}

Label with Implicit Reward

Reference LLM π_{ref}

Preference Judgment: {y_w} > {y_l}

New Preference Data Generation

New Preference Data {(x, y_w, y_l)} → Simulated LLM $\pi_{\widetilde{\theta}}$ → **Noise Identified Data**

Simulate Strongly Aligned LLM for Noisy Label Detection

Reference LLM π_{ref} Current LLM π_θ

Learning with Refined Preference

Preference Learning with Self-Refinement