Supervised Finetuning

CoT Data
$\{(X, E, Y)\}$

LLM $\xrightarrow{\text{SFT}}$ LLM with CoT Reasoning

Preference Refinement

$\{q, a^-, a^+\}$
Prompt
$\xrightarrow{\text{Request Judgment}}$

$\{CoT: ?, Chosen: ?\}$ | Sampling

Samples
$\{E_1, a^+\}$ $\{E_2, a^+\}$ $\{E_3, a^+\}$ $\{E_4, a^-\}$ $\{E_5, a^-\}$

Scoring

Multi-level Preference

Strong Accept | Weak Accept | Strong Reject | Weak Reject