GRPO (Group Relative Policy Optimization) Pipeline

MODEL

Policy π_θ
LoRA r=16

Reward Funcs
refusal+help+fmt

INPUT

SFT
Checkpoint
(merged
LoRA)

PKU-
SafeRLHF
12,035
prompts

DATA

Clear
Contrast
Filter

Extract
Queries
(prompts
only)

FORMAT

Instruct
[sys, user]

NoInstruct
[user]

RL LOOP

Generate
K=6
responses

Score with
reward
funcs

Group-
Relative
Advantage

OUTPUT

GRPO_NoInstruct
$\pi(Y|X)$

GRPO_Instruct
$\pi(Y|I,X)$