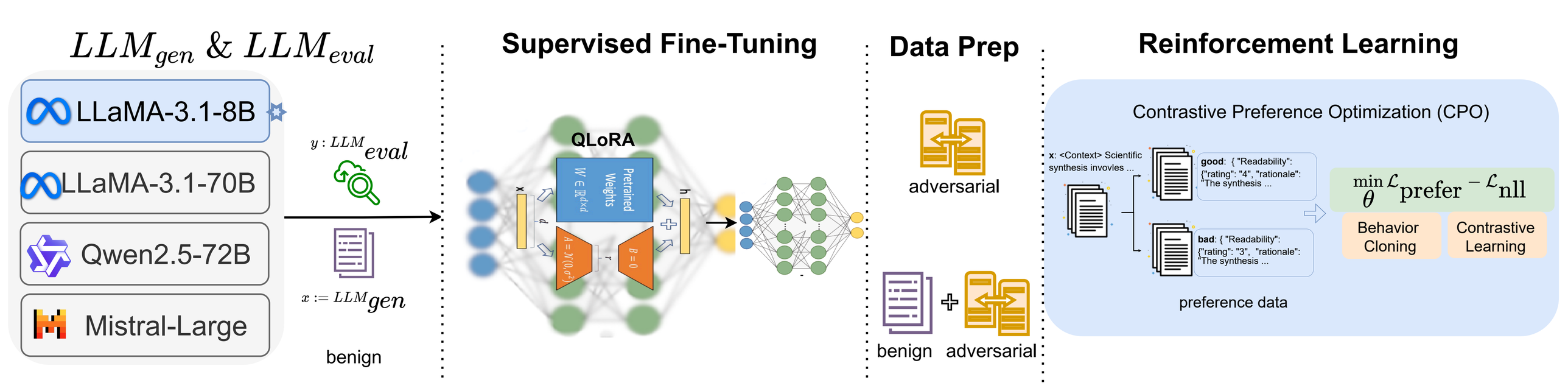
LLM_{gen} & LLM_{eval}
LLaMA-3.1-8B
LLaMA-3.1-70B
Qwen2.5-72B
Mistral-Large
$y : LLM_{eval}$
$x := LLM_{gen}$
benign
Supervised Fine-Tuning
QLoRA
Pretrained Weights
$W \in \mathbb{R}^{d \times d}$
$A = \mathcal{N}(0, \sigma^2)$
$B = 0$
x
h
d
r
Data Prep
adversarial
benign
adversarial
Reinforcement Learning
Contrastive Preference Optimization (CPO)
x: <Context> Scientific synthesis invovles ...
good: { "Readability": {"rating": "4", "rationale": "The synthesis ...
bad: { "Readability": {"rating": "3", "rationale": "The synthesis ...
preference data
$\min_\theta \mathcal{L}_{\text{prefer}} - \mathcal{L}_{\text{nll}}$
Behavior Cloning
Contrastive Learning