

# README for Training Stages and Configurations

---

This document outlines the training phases and configurations used for model development. The following is a detailed description of the key training phases, configurations, and decisions in the process. This folder also contains data examples for the cold start and RL phases for NeurIPS 2025 review.

## Training Stages

---

### 1. 1 Cold Start

- **Purpose:**  
To establish the model's initial reasoning capabilities using synthetic data, focusing on non-reflective and reflective reasoning patterns.
- **Key Parameters:**
  - **Model:** Qwen2VL-72B
  - **Training Type:** LoRA
  - **Learning Rate:**  $1e-4$
  - **Batch Size:** 1 (per device), gradient accumulation steps = 8
  - **Epochs:** 4
  - **LoRA Config:**
    - Rank: 16
    - Alpha: 32
    - Dropout: 0.05
  - **Scheduler:** Cosine decay with a warmup ratio of 0.01

### 1.2 Cold Start - RPO

- **Purpose:**  
To refine the model's reasoning capabilities using curated datasets, focusing on step-by-step multimodal reasoning.
- **Key Parameters:**
  - **Model:** Continuation of Qwen2VL-72B from Cold Start 1
  - **Training Type:** LoRA
  - **Learning Rate:**  $1e-5$
  - **Batch Size:** 1 (per device), gradient accumulation steps = 4
  - **Epochs:** 1
  - **LoRA Config:** Same as Stage 1
  - **Scheduler:** Cosine decay with a warmup ratio of 0.01 - **Logging:** TensorBoard with logging every 5 steps

## 2.1 RL Optimization - KTO

- **Purpose:**

To enhance generalization and reasoning through reinforcement learning using the KTO (Knowledge Transfer Optimization) strategy.

- **Key Parameters:**

- **Model:** Derived from RPO Phase
- **Training Type:** LoRA
- **Learning Rate:**  $1e-5$
- **Batch Size:** 1 (per device), gradient accumulation steps = 2
- **Epochs:** 1
- **Scheduler:** Cosine decay with a warmup ratio of  $0.05$
- **Trainer:** Reinforcement learning with KTO
- **Evaluation Steps:** Every 2000 steps
- **Output Directory:** RL results saved for further fine-tuning

## 2.2 Annealing

- **Purpose:**

To address instability observed after the KTO phase by applying a small-scale SFT step.

- **Key Parameters:**

- **Model:** Derived from KTO Phase
- **Training Type:** LoRA
- **Learning Rate:**  $5e-6$
- **Batch Size:** 1 (per device), gradient accumulation steps = 2
- **Epochs:** 1 (limited to 1000 steps)
- **Scheduler:** Cosine decay with a warmup ratio of  $0.05$
- **Dataset:** Reduced dataset with 40,000 samples
- **Logging:** TensorBoard with logging every 5 steps
- **Outcome:**
  - While performance showed minor degradation, stability improved significantly, which led to the adoption of this version as the final model.

The complete datasets used in this research will be made publicly available after publication.

The final model was selected after the SFT annealing phase due to its enhanced stability, even though there was a slight performance trade-off. This version demonstrated consistent and reliable behavior across various datasets, making it the optimal choice for deployment.