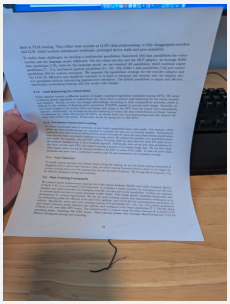

Question: Based on the paper, which one takes longer to train, RL phase or reward model?

Ground Truth: RL phase takes longer.

OVR: *</think>* Okay, so I need to figure out whether the RL phase or the reward model takes longer to train based on the information provided in the paper. Let me go through the relevant parts step by step.
First, the paper mentions the Seed1.5-VL model. It talks about a post-training framework using hybrid reinforcement learning with RLHF (Reinforcement Learning with Human Feedback) and RLVF (Reward Learning with Verifier Feedback). The RL phase is mentioned as requiring 24k GPU hours, while the reward model inference uses 60k GPU hours. Wait, but that's for the reward model during inference, right? Because it says "reward model inference use replicas each configured with tensor parallelism." So during training, maybe the reward model is trained for longer?
Wait, let me check again. The paper says: "The RL phase of Seed1.5-VL costs 60k GPU hours. The reward model is trained using the same framework as Seed1.5-VL pretraining phase, requiring 24k GPU hours." Hmm, that seems conflicting. Wait, no: "pretraining phase" is separate. Let me parse this sentence again.
Wait, the sentence says: "The RL phase of Seed1.5-VL costs 60k GPU hours. The reward model is trained using the same framework as Seed1.5-VL pretraining phase, requiring 24k GPU hours." So the RL phase takes 60k hours, and the reward model training (which is part of the post-training framework) uses 24k hours? That would mean the RL phase is longer than the reward model training. But maybe I'm misinterpreting.
Alternatively, perhaps there's a typo or confusion in the description. **Let me look again at the exact wording:**
"We use the same training framework as Seed1.5-VL pretraining phase for optimization techniques as in the pretraining phase for efficient actor and critic updates, and vLLM [81] for efficient progressive generation of rollouts. Specifically, the RL and RLVF training employs 3-D parallelism [61, 108, 126, 141]; rollout generation and reward/reference model inference use replicas each configured with tensor parallelism [7]. The RL phase of Seed1.5-VL costs 60k GPU hours. The reward model is trained using the same framework as Seed1.5-VL pretraining phase, requiring 24k GPU hours."
Wait, so here it says that during RL training (which includes both RL and RLVF?), it takes 60k hours. Then separately, the reward model is trained using the same framework as pretraining (which might be different) but takes 24k hours. But if pretraining took longer than post-training phases? Or maybe there's a mix-up here...
But wait, **let me check again**:
Original text from user input:
"We use ... optimization techniques as in the pretraining phase for efficient actor and critic updates ... The RL phase of Seed1.5-VL costs 60k GPU hours ... The reward model is trained using ... same framework as Seed1.5-VL pretraining phase ... requiring 24k GPU hours."
So during post-training phases:

- RL Phase: 60k
- Reward Model Training: 24k

Thus, RL Phase is longer than Reward Model Training.
Therefore, answer would be that the RL Phase takes longer to train than the Reward Model. *</think>*

Based on the provided information from the paper:

- **RL Phase**: Requires **60k GPU hours**.
- **Reward Model Training**: Requires **24k GPU hours** during post-training phases.

Conclusion: The **RL Phase** takes **longer** to train compared to **Reward Model Training**(60k vs. 24k GPU hours).

Answer: RL Phase