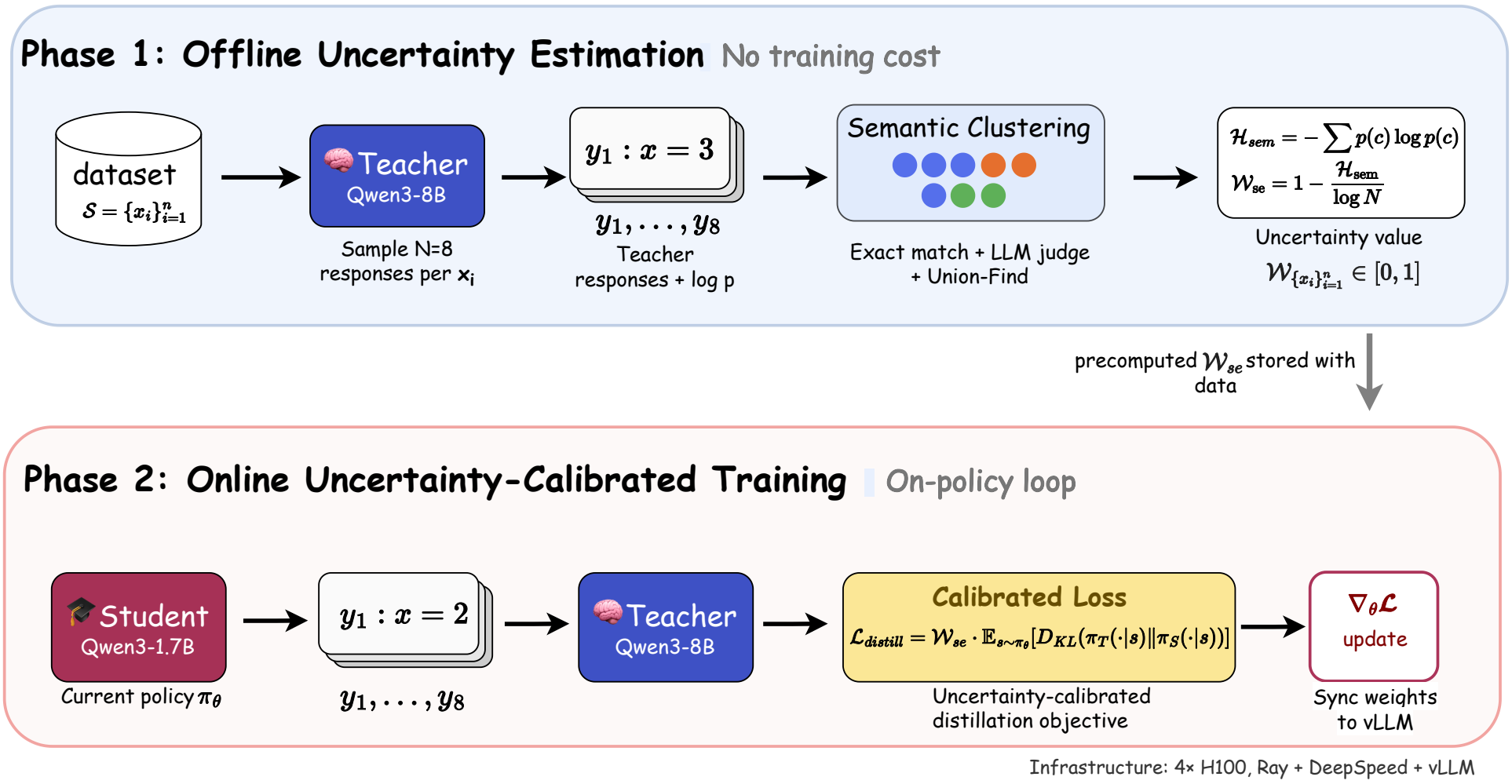
Phase 1: Offline Uncertainty Estimation
No training cost
dataset
$\mathcal{S} = \{x_i\}_{i=1}^{n}$
Teacher
Qwen3-8B
Sample N=8
responses per x_i
$y_1 : x = 3$
$y_1, \ldots, y_8$
Teacher
responses + log p
Semantic Clustering
Exact match + LLM judge
+ Union-Find
$\mathcal{H}_{sem} = -\sum p(c) \log p(c)$
$\mathcal{W}_{se} = 1 - \frac{\mathcal{H}_{sem}}{\log N}$
Uncertainty value
$\mathcal{W}_{\{x_i\}_{i=1}^{n}} \in [0, 1]$
precomputed $\mathcal{W}_{se}$ stored with data
Phase 2: Online Uncertainty-Calibrated Training
On-policy loop
Student
Qwen3-1.7B
Current policy π_θ
$y_1 : x = 2$
$y_1, \ldots, y_8$
Teacher
Qwen3-8B
Calibrated Loss
$\mathcal{L}_{distill} = \mathcal{W}_{se} \cdot \mathbb{E}_{s \sim \pi_\theta}[D_{KL}(\pi_T(\cdot|s) \| \pi_S(\cdot|s))]$
Uncertainty-calibrated
distillation objective
$\nabla_\theta \mathcal{L}$
update
Sync weights
to vLLM
Infrastructure: 4× H100, Ray + DeepSpeed + vLLM