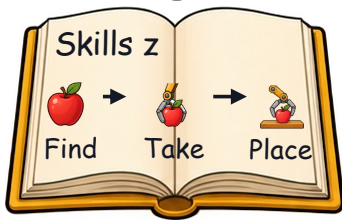


Task



Put the apple on the shelf

Training



Teacher · with skills



$$q = \pi_{\theta}(\cdot | c, z)$$

Privileged supervision



Student · without skills



$$p = \pi_{\theta}(\cdot | c)$$

Skill internalize

Inference



Without skills

Student trajectories

τ_1 : look ... take apple ... put apple $R = 1$ ✓
 $A_1 > 0$

τ_2 : look ... take mug ... put mug $R = 0$ ✗
 $A_2 < 0$

⋮

τ_n : go table ... take apple ... put apple $R = 1$ ✓
 $A_n > 0$

All trajectories

GRPO Objective

$$L_{GRPO}$$

Level 1: Trajectories

$$A_i > 0$$

Selected trajectories

$$A_i \leq 0$$

Skip distillation

Level 2: Tokens

Same prefix $c_{i,t}$



p (student)



q (teacher)

Example ($T_i = 4$)

	take	apple	from	table
Teacher support s_t	0.40	0.80	0.00	0.20
Student uncertainty u_t	1.50	1.25	1.10	1.50
Teacher certainty v_t	0.80	0.90	0.70	0.40

$$r_t = s_t u_t v_t$$

$$w_{i,t} = \frac{T_i r_{i,t}}{\sum_{j=1}^{T_i} r_{i,j}}$$

Token weight w	1.28	2.40	0.0	0.32
------------------	------	------	-----	------

Weighted teacher-student KL

Distillation Objective

$$L_{distill}$$

Joint Objective

$$L = L_{GRPO} + \lambda L_{distill}$$

