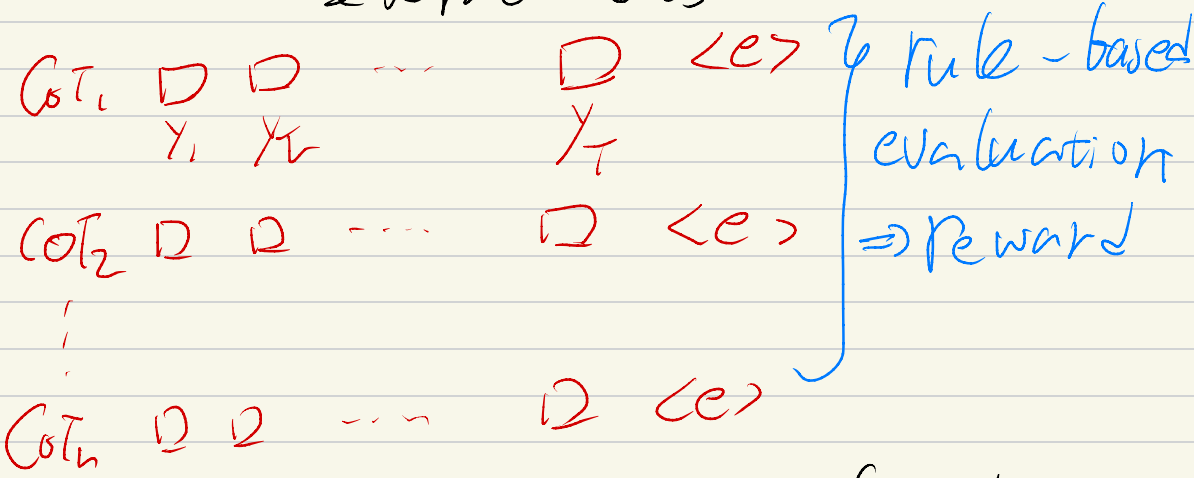
25.03.11

non-reasoning model

$\pi_\theta(y|x)$: Given a prompt, usually a math or coding question, generate several CoTs



use RL on CoTs to further fine-tune $\pi_\theta(y|x)$

Knowledge Distillation

Teacher model: $p_i(x) = \frac{\exp(f_i(x))}{\sum_{j \in V} \exp(f_j(x))}$

$f_i(x)$: logits

Student Model: $q_i(x) = \frac{\exp(g_i(x|\theta))}{\sum_{j \in V} \exp(g_j(x|\theta))}$

y is the GT label

Loss of Distillation:

$$D_{KL}(p(x|\tau) \,|\, q(x|\theta)) + \lambda[-\log q_y(x|\theta)]$$

$$= -\sum_{i \in V} p_i(x|\tau) \log q_i(x|\theta) - \lambda \log q_y(x|\theta)$$

$$p_i(x|\tau) = \frac{\exp(f_i(x)/\tau)}{\sum_{j \in V} \exp(f_j(x)/\tau)}$$

τ: Temperature

higher temperature implies more random outcome.