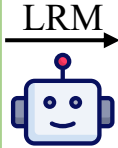


Stage I: Confidence-Aware Preference Labeling (CAPL)

Input Question

e.g.,
Crystal wants to buy a meal with one entree, one drink, and one dessert. The table below shows her available choices. How many distinct meals can she make?



LRM

1) Sample K Paths & Compute Self-Certainty for Each Path y

	Correctness	Length ($ y $)	Self-Certainty (SC)
y_a	✓	831	17.52
y_b	✓	3949	16.10
...	...	...	...
y_c	✗	2306	15.86
y_d	✗	1279	16.43

✓ Correct ✗ Incorrect

2) Categorize Preference Pairs

Conciseness Pairs (CPs)

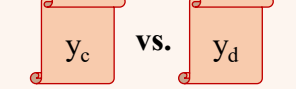
correct vs. correct



Prefer shorter (penalize under-confident long successes)

Deliberation Pairs (DPs)

incorrect vs. incorrect



Prefer longer (penalize over-confident short failures)

3) Pair Scoring

For CPs

(e.g., a is better, b is worse)

$$s_{CP}(x, y_a, y_b) = \Delta r \cdot \Delta SC$$

$$\Delta r = \frac{1}{|y_a|} - \frac{1}{|y_b|}$$

$$\Delta SC = SC(x, y_a) - SC(x, y_b)$$

For DPs

(e.g., c is better, d is worse)

$$s_{DP}(x, y_c, y_d) = \Delta r \cdot (-\Delta SC)$$

$$\Delta r = \frac{1}{|y_d|} - \frac{1}{|y_c|}$$

$$\Delta SC = SC(x, y_c) - SC(x, y_d)$$

4) Dynamic Pruning

Top-Score Selection
(keep top-3 score levels in CPs and DPs)

Global Truncation
(drop the bottom τ fraction)

Per-Query Cap
(≤ 1 CP and DP for each x)

Output

Preference Dataset D
 $D = \{(x, y_w, y_l, s)\}$

- x : question
- y_w : winning trajectory
- y_l : losing trajectory
- s : preference score

Stage II: Confidence-Weighted Preference Optimization (CWPO)

Input

Preference Dataset D

$$D = \{(x, y_w, y_l, s)\}$$



1) Calibration Ratio

For each path y

$$\rho(x, y) = \frac{SC(x, y)}{|y|^\alpha}$$

SC: self-certainty
 $|y|$: length of reasoning path
 α : length-aware exponent

2) Dynamic Weight

For CPs

$$\beta_w = \beta_{base} \cdot \sigma(\rho(x, y_w))$$

$$\beta_l = \beta_{base} \cdot \sigma(\rho(x, y_l))^{-1}$$

Upweight short & confident successes

For DPs

$$\beta_w = \beta_{base} \cdot \sigma(\rho(x, y_w))^{-1}$$

$$\beta_l = \beta_{base} \cdot \sigma(\rho(x, y_l))$$

Upweight long & uncertain failures

β_{base} : the original β in SimPO

3) CWPO Loss

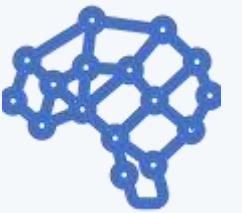
Confidence-Weighted SimPO Objective:

$$\mathcal{L}_{CWPO}(\pi_\theta)$$

$$= -\mathbb{E}_{(x, y_w, y_l) \sim D} \left[\log \sigma \left(\frac{\beta_w}{|y_w|} \log \pi_\theta(y_w|x) \right. \right.$$

$$\left. \left. - \frac{\beta_l}{|y_l|} \log \pi_\theta(y_l|x) - \gamma \right) \right]$$

Optimize Model Update θ



Easy Problem → Short Reasoning

Q: If n leaves a remainder of 2 when divided by 7, what remainder does $(n+2)(n+4)(n+6)$ leave when divided by 7?



concise and confident!

Achieve a Balance Between Efficiency and Accuracy

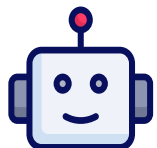


Hard Problem → Long Reasoning

Q: What is the sum of the proper divisors of the sum of the proper divisors of 284?



more exploration more deliberation!



LRM