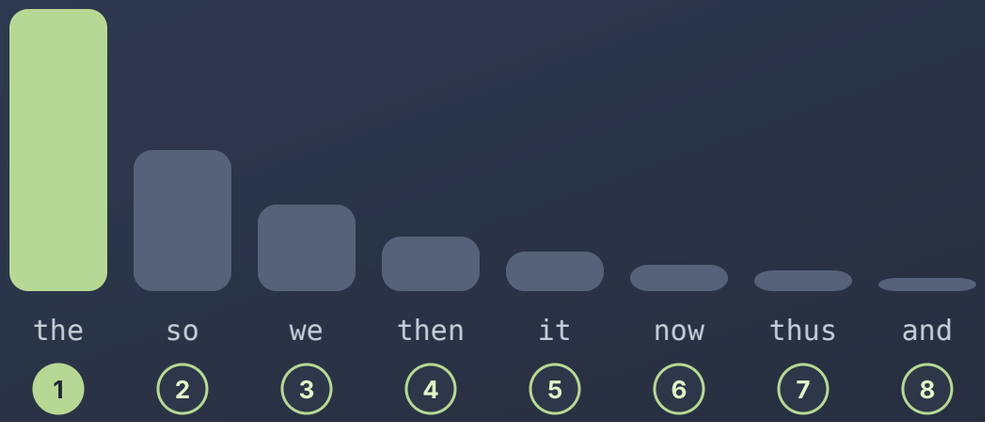
Teach the small model the teacher's ranking, not its probabilities.

TEACHER'S TOP-8 TOKENS · ILLUSTRATIVE

copy only the order



BEST GAIN · QWEN3-4B → 1.7B

+2.9

points on the HMMT25 / AIME24 / AIME25 average vs standard on-policy distillation (38.0 → 40.9)

THE PAPER

Order-Only On-Policy Distillation (O^3PD)

Guangzeng Han, Yu Wu, Xiaolei Huang · University of Memphis, University of Missouri · 1 Nov 2026

THE IDEA

Keep the teacher's preferences. Drop its confidence.



Mohammad Ashfaq Ur
AI Research Notes

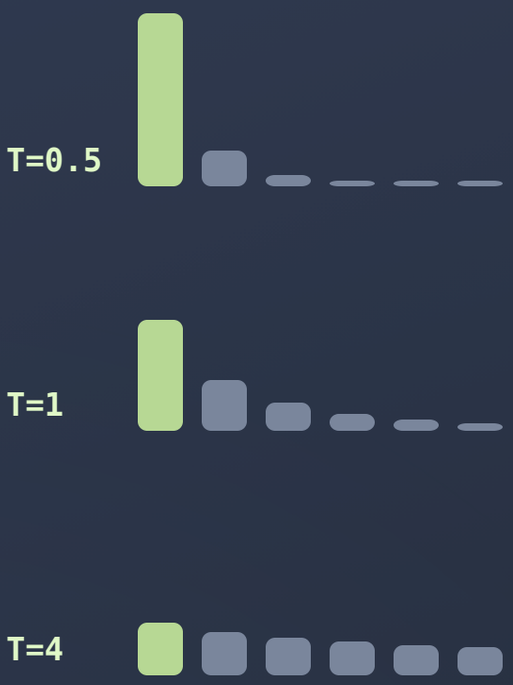
ashfaqfardin.dev

arXiv 2611.02112 · preprint

Teacher probabilities mix what it prefers with how sure it is.

On-policy distillation (OPD) makes the student match those values token by token. A small student may never match a big teacher's certainty, and temperature rescales the values while leaving the order alone.

SAME LOGITS, THREE TEMPERATURES (ILLUSTRATIVE)



Heights change a lot. Order never changes.

OPD TEMPERATURE SWEEP · HMMT25

No single best T: 0.25 for 8B→1.7B, 2 for 14B→8B and 4B→1.7B.

44.7 → 38.3

8B→4B at T = 1 vs T = 8

QWEN3-14B → 0.6B · HMMT25

11.9 → 1.1

T = 1 vs T = 8. Smaller students are hit hardest.



Mohammad Ashfaq Ur
AI Research Notes

Same on-policy loop. Only the target changes.

1

Roll out

Student (Qwen3, non-thinking) writes its own answer

2

Ask teacher

Frozen teacher scores each prefix; keep its top-8 tokens

3

Rank pairs

For each neighbour pair i, i+1: student logit for i should be higher

4

Update

LoRA step on the ranking loss; repeat for 100 steps

ADJACENT LOGISTIC (BRADLEY–TERRY) LOSS



$$\ell = -\log \sigma\left(z_S(v_i) - z_S(v_{i+1}) \right)$$

z_S = student logits, v_i = teacher's i-th ranked token. Averaged over all tokens and the 7 neighbour pairs.

Baseline OPD: Jensen–Shannon divergence on the teacher's top-8, T = 1.

WHY TEMPERATURE DROPS OUT

T-free

Any positive rescaling of teacher logits keeps the top-8 order, so the targets are identical at every T.



Mohammad Ashfaq Ur
AI Research Notes

ashfaqfardin.dev

arXiv 2611.02112 · preprint

O³PD beat OPD in all six teacher → student pairs

AVG@12 OF HMMT25, AIME24, AIME25 · %



AIME25 · ALL 6 PAIRS

+5.6

best gain (4B→1.7B); mean +3.4

SMALLEST GAINS

+0.3 for 14B→1.7B, +0.6 for the 8B student

Qwen3 students and teachers, OpenThoughts maths, LoRA, 100 steps. Best checkpoint per method. Source: Table 1, arXiv 2611.02112.



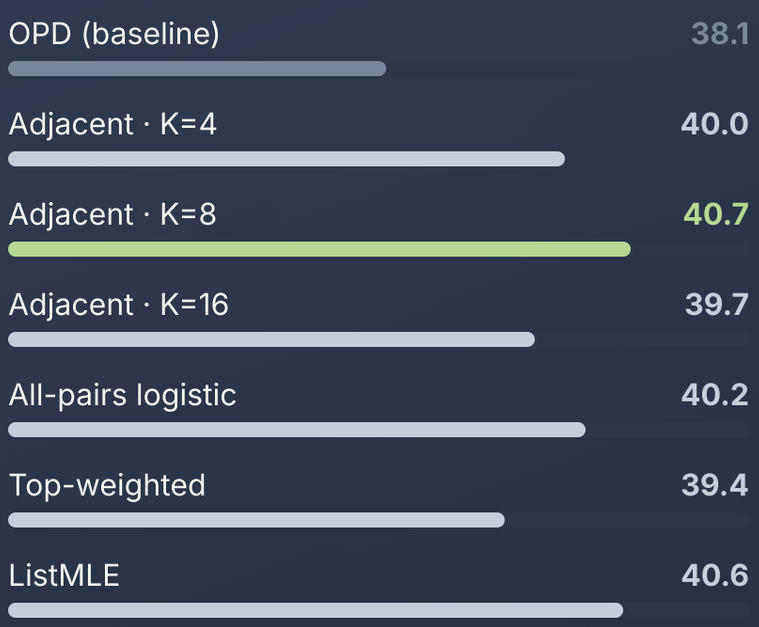
Mohammad Ashfaq Ur
AI Research Notes

ashfaqfardin.dev

arXiv 2611.02112 · preprint

It's the ordinal signal, not one clever loss

ABLATIONS · 8B → 1.7B · AVG %



Bars start at 34%. Every K and every ranking loss beats OPD by 1.3–2.6 points.

TRAINING DYNAMICS · 8B → 1.7B · AVG %

37.7 untrained 1.7B student

38.1 OPD peak, at step 20, then falls below 37.7

40.7 O³PD, still rising at step 100

OPD's best checkpoint comes early; O³PD keeps improving.

Odd but telling: O³PD students end up *further* from the teacher's probabilities (higher JSD) yet agree more on rank order (higher Kendall's τ and top-8 overlap).



Mohammad Ashfaq Ur
AI Research Notes

ashfaqfardin.dev
arXiv 2611.02112 · preprint

For distillation, the teacher's order may matter more than its numbers.

CRITICAL PERSPECTIVE

- Preprint, not peer-reviewed; no code released yet
- Qwen3 family and maths benchmarks only; no code, chat or safety eval
- Single training seed; best checkpoint picked on the reported benchmarks
- Small gains at the extremes: +0.6 (8B student), +0.3 (14B → 1.7B)

DISCUSSION

Do you tune teacher temperature when you distil, and would rank-only targets hold up beyond maths?

REFERENCES

arxiv.org/abs/2611.02112

huggingface.co/papers/2611.02112

Han, Wu, Huang · 1 Nov 2026 · cs.LG



Mohammad Ashfaq Ur

Follow for daily AI research breakdowns

ashfaqfardin.dev

