ICML 2026 WORKSHOP ON MECHANISTIC INTERPRETABILITY

Weight-Space Geometry of Offline Reasoning Training

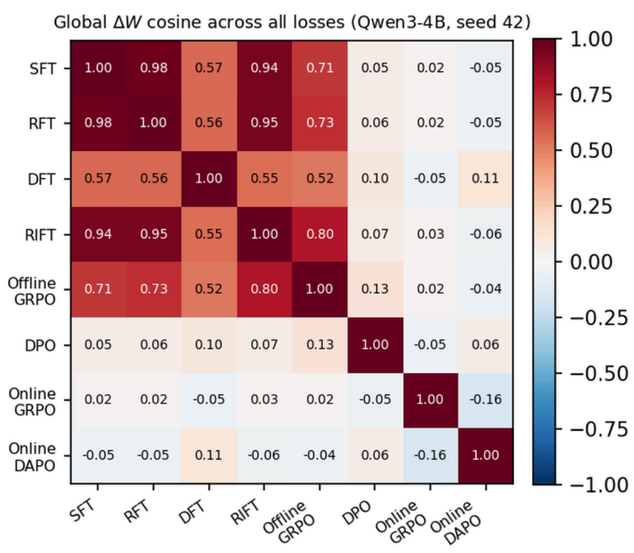
Aleksandr Nikolich[1] · Igor Kiselev[2] · Vladimir Platonov[3] · Karina Romanova[3]

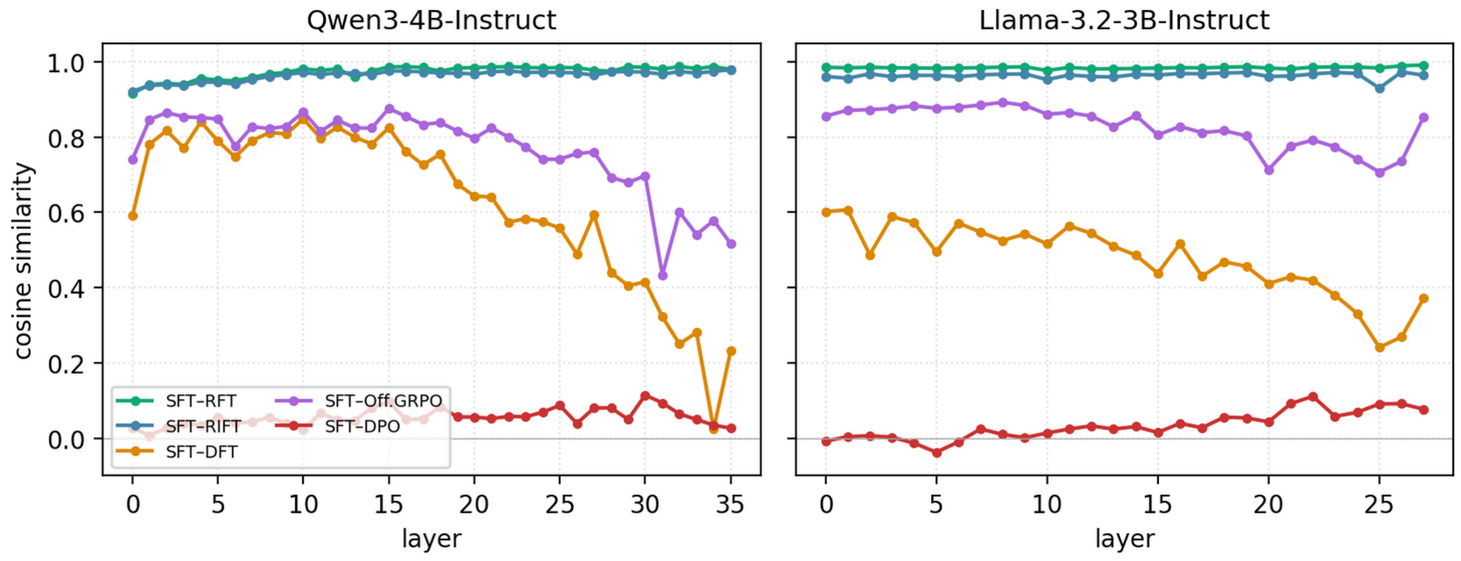
[1]Keenable AI · [2]Accenture · [3]Yandex

Different "offline RL" losses, same data → the *same weight update as plain SFT.* What really moves the model is *on-policy sampling, not the loss.*

8 losses, one base, shared LoRA, identical rollouts — we read the weight updates ΔW, not just accuracy. (Qwen3-4B; replicated on Llama-3.2-3B.)

Scan for the paper





Left: cosine between all 8 LoRA ΔW (Qwen3-4B; red = same direction) — SFT/RFT/RIFT one block (≥0.94), DFT ~0.55, Offline GRPO 0.7–0.8, DPO & on-policy near-orthogonal. **Right:** per-layer cosine to SFT (Qwen3-4B & Llama-3.2-3B) — the cluster stays ~0.97 at every layer; Offline GRPO, DFT and DPO diverge in deep layers.

The question

A new offline-RL loss appears almost monthly (RFT, RIFT, DFT, Offline GRPO, DPO), each judged on accuracy alone. We ask what each actually does to the **weights** — a different update, or the same one?

Setup

- Same base (Qwen3-4B), same math rollouts, shared LoRA — **only the loss changes.**
- **8 losses**, incl. on-policy GRPO & DAPO.
- Compare **ΔW**: cosine, principal angles, effective rank.
- Loss landscape: mode connectivity + CKA. Replicated on Llama-3.2-3B.

What we find

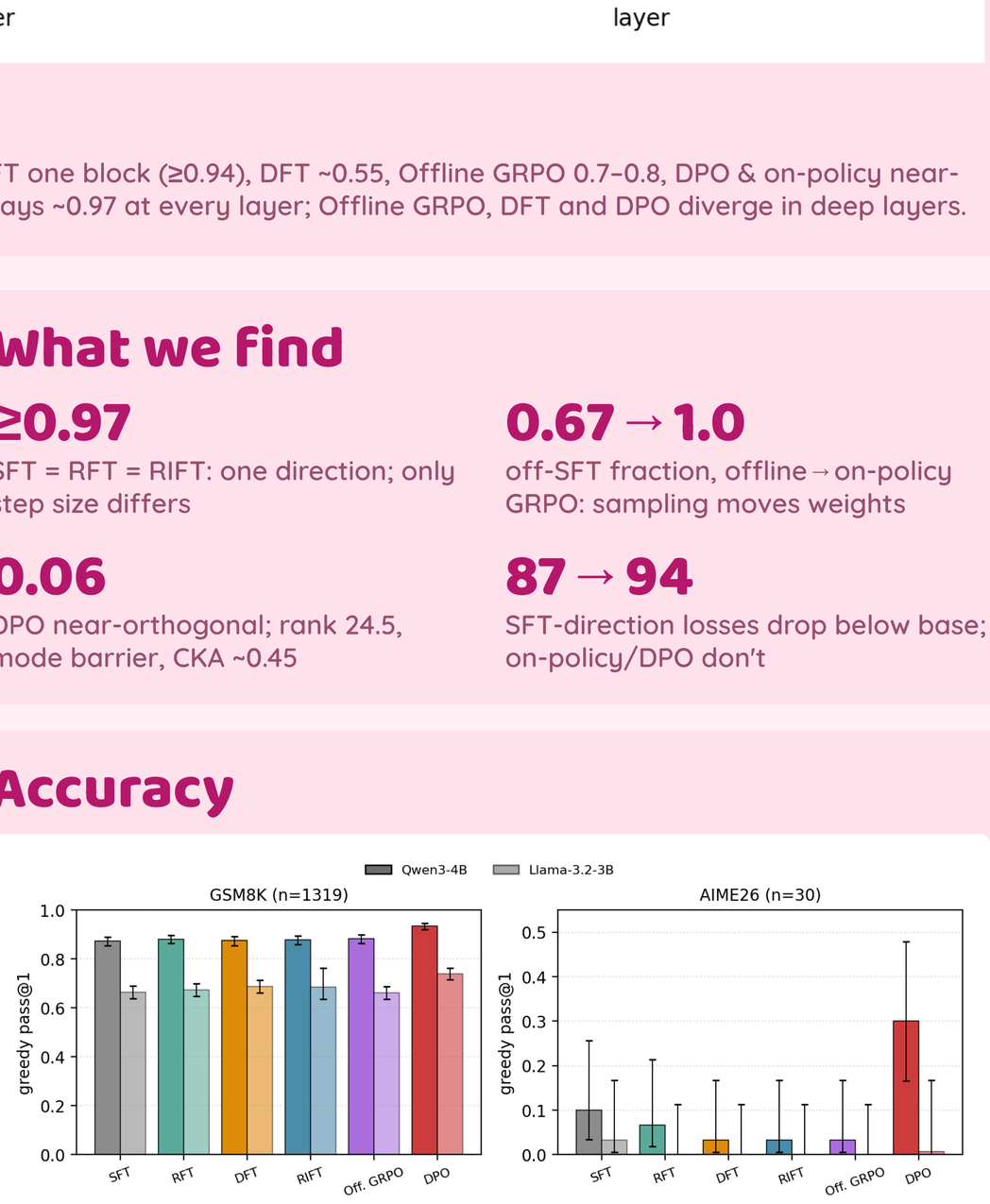
≥0.97
SFT = RFT = RIFT: one direction; only step size differs

0.67 → 1.0
off-SFT fraction, offline→on-policy GRPO: sampling moves weights

0.06
DPO near-orthogonal; rank 24.5, mode barrier, CKA ~0.45

87 → 94
SFT-direction losses drop below base; on-policy/DPO don't

Accuracy



Greedy pass@1 (95% CI); Qwen3-4B dark, Llama-3.2-3B light. DPO tops both (GSM8K 93.5%, AIME26 30%); the cluster is tied. DPO uses 10× smaller LR — edge is suggestive, not causal.

Refs. [1] Rafailov, DPO, NeurIPS'23 · [2] Wu, DFT, '25 · [3] Liu, RIFT, '26 · [4] Shao, GRPO, '24 · [5] Frankle, LMC, ICML'20 · [6] Kornblith, CKA, ICML'19

paper · code · adapters ↑

Takeaway: on fixed rollouts, "offline RL" ≈ SFT; *on-policy* sampling is what moves the weights — and keeps accuracy.