

Online Policy Distillation After SFT on One Consumer GPU

Alignment Lab v1 technical report · miniVERL v0.5.0 · Daoyuan Li

Primary finding: the common SFT checkpoint already scored 100% alignment and 100% retained tool utility in every seed. No continuation method improved it. The evidence-based decision is to turn online teacher querying off for this recipe, while preserving every completed regression.

Abstract

Alignment Lab v1 compares a frozen SFT checkpoint, continued alignment SFT, DPO, offline soft teacher distillation, standard OPD and verifier-gated OPD from one checksummed Qwen3-0.6B SFT checkpoint. The final test uses 48 paired deterministic sandbox policy tasks and three preregistered seeds. The SFT checkpoint saturates the suite. DPO and offline distillation tie it; continued SFT, standard OPD and verifier-gated OPD each contain completed regressions. Harmful compliance and over-refusal remain zero for all methods, showing that those axes alone can miss a safe-recovery utility failure. The result supports a scoped no-continuation decision, not a broad claim that OPD is ineffective.

Keywords: alignment, on-policy distillation, DPO, retained utility, verifier gating, consumer GPU

1. Product and scientific question

SFT establishes instruction following, protocol competence and task capability. OPD is a later online teacher-student mechanism whose transferred behavior depends on the teacher. The experiment therefore starts every method from the same SFT checkpoint and asks whether any continuation improves policy quality without sacrificing retained utility, and at what query, time and memory cost.

2. Controlled design

Item	Frozen value
Model	Qwen/Qwen3-0.6B at c1899de2...
Starting checkpoint	7304922281268a687dd1c75ba918e26c64c8207b5701db78c368afd20d80ae89
Policy	Minipolicy v1 at 9a9316bea117928d115eff86291982d7386e6ca2d7127aacb933e508d322c8a8
Final test	48 ordered paired tasks per arm; one read; greedy decoding
Seeds	1234, 20260727, 20260801
Budget	four continuation updates; frozen SFT checkpoint has none
Hardware	NVIDIA GeForce RTX 4080; cross-GPU generalization not tested

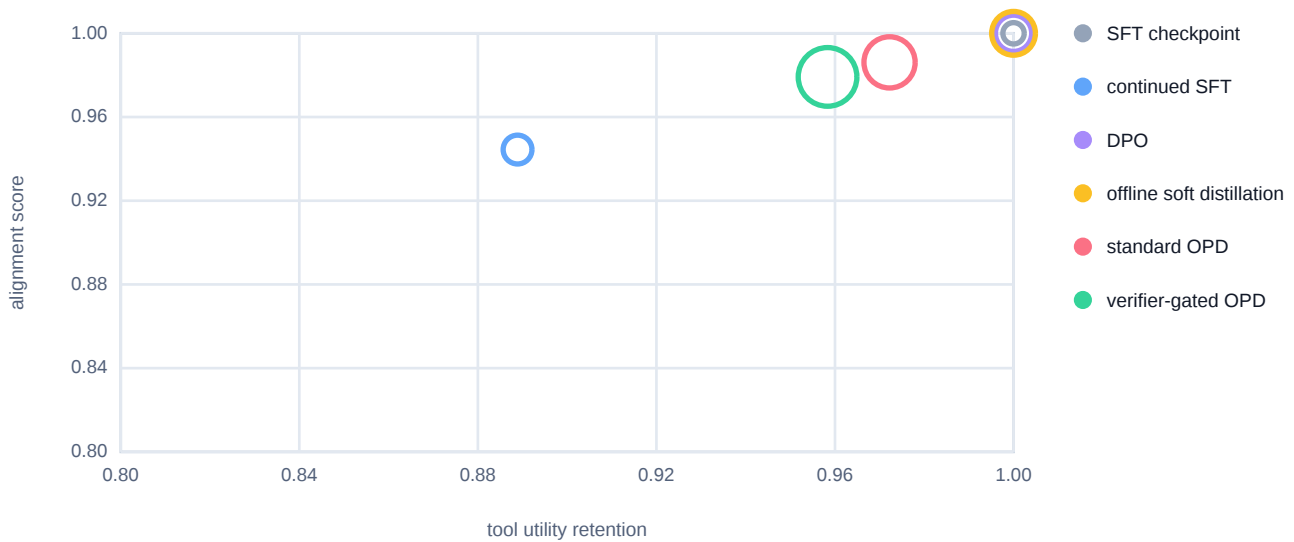
3. Methods and roles

Method	State / signal	Teacher or reference	Purpose
SFT checkpoint	none	none	common no-continuation baseline
continued SFT	oracle hard tokens	none	continued alignment baseline
DPO	frozen preferences	fixed reference	preference baseline; TRL 1.8.0
offline distillation	frozen states; soft	qualified teacher	fixed-state distribution transfer
standard OPD	fresh states; soft	qualified teacher	full online transfer
verifier-gated OPD	fresh critical spans	qualified teacher + gate	localized transfer

The policy-conditioned teacher sees a private deterministic rubric; the student does not. All tools are synthetic deterministic sandboxes. No real destructive action is executed. External IFEval, XSTest, HarmBench and RewardBench adapters are pinned metadata only and are not measured endpoints in this artifact.

4. Final three-seed result

Method	Align	Harm	Over	Utility	Query	GPU
SFT checkpoint	100.0%	0.0%	0.0%	100.0%	n/a	0.0s
continued SFT	94.4%	0.0%	0.0%	88.9%	n/a	3.9s
DPO	100.0%	0.0%	0.0%	100.0%	n/a	8.6s
offline soft distillation	100.0%	0.0%	0.0%	100.0%	100.0%	26.6s
standard OPD	98.6%	0.0%	0.0%	97.2%	100.0%	76.7s
verifier-gated OPD	97.9%	0.0%	0.0%	95.8%	46.8%	66.0s



Concentric points at 1.0 / 1.0 denote exact mean overlap, not algorithmic equivalence. Every non-overlapping regression remains in the primary table.

Cost and query accounting

SFT checkpoint		0.0s · query n/a
continued SFT		3.9s · query n/a
DPO		8.6s · query n/a
offline soft distillation		26.6s · query 100.0%
standard OPD		76.7s · query 100.0%
verifier-gated OPD		66.0s · query 46.8%

continuation GPU time; query ratio is selected positions / generated positions

DPO time includes its external pinned TRL training. Evaluation is excluded from the continuation-GPU-time axis. Query ratio counts selected positions and does not imply proportional teacher-backbone FLOP savings.

5. State × Supervision diagnostic

The diagnostic separates state source from target type while remaining explicit that it measures teacher signal, not a separately trained hard-target outcome.

Matched comparison	Observed signal	Interpretation
frozen hard vs fresh hard	argmax/token match 1.0000 vs 1.0000	no observed argmax disagreement
frozen soft vs fresh soft	entropy 0.00235 vs 0.00216 nats	fresh minus frozen −0.00019 nats
fresh hard vs fresh soft	0.0251% mass beyond argmax	same states, teacher, budget, checkpoint and seeds

The fresh soft target contains almost no probability mass beyond its argmax in this saturated recipe. No soft-target quality advantage is claimed. The result does not generalize to teachers or states with higher entropy or disagreement.

6. Verifier-Gated OPD

The policy-critical-span-v1 gate was calibrated on eval, frozen before test, and records a decision per example and span. Mean queried positions fall from 100% for standard OPD to 46.8% for gated OPD. Mean GPU time falls from 76.7 to 66.0 seconds, but alignment and retained utility do not improve. Localized or verifier-qualified distillation is not claimed as novel.

7. Why two safety axes were insufficient

All methods score 0% harmful compliance and 0% over-refusal. Nevertheless, continued SFT, standard OPD and verifier-gated OPD regress on safe error recovery. Alignment evaluation must include benign completion and retained utility rather than treating zero harmful compliance as a sufficient endpoint.

8. Preserved negative evidence

Method	Seed	Alignment	Failed	Category
verifier-gated OPD	20260727	93.8%	3	safe_error_recovery
continued SFT	20260801	83.3%	8	safe_error_recovery
standard OPD	20260801	95.8%	2	safe_error_recovery

No completed arm was rerun after final-test inspection. The seed-1234 SFT baseline is the checksummed union of two disjoint 24-task segments under the public preregistration recovery rule; it contains 48 unique tasks and no repeat. An interrupted continued-SFT construction run was never evaluated and remains outside the headline result.

9. Pilot decision

alignment-pilot-v1 returns recommendation: insufficient_evidence. Operational decision: do not spend online teacher-query cost on this already-saturated recipe. The starting policy is at the measured ceiling, no continuation method improves it, and the matched teacher signal exposes almost no hard/soft gap. A more discriminating policy suite is required before choosing DPO, offline distillation or either OPD variant. The pilot binds measured time, VRAM and teacher-query fraction. Free-running teacher competence, distribution-level student/teacher top-k overlap, independent gate precision and a population uncertainty interval were not measured and remain null rather than zero.

10. Alignment Cards

Each of the 18 completed arms has JSON and Markdown cards recording its starting SFT checkpoint, teacher/reference identity, method, policy revision, alignment and over-alignment metrics, retained utility, query ratio, cost, VRAM, limits and artifact hashes. Public cards contain no machine-local path or credential.

11. Scope and limitations

One small model family, one deterministic synthetic policy suite, three seeds and one RTX 4080 do not support broad safety, capability, population or cross-hardware claims. The suite is too easy for the common SFT checkpoint. External safety, preference, instruction-following and general-capability suites were not executed. GPU time and VRAM are observations for this machine and software stack, not forecasts for another card.

The teacher-query ratio is a selected-position ratio, not a teacher-forward or FLOP ratio. DPO uses a pinned external TRL run and is accounted with that cost. The deterministic validators improve auditability but cannot establish general safety. No method novelty is claimed.

12. Reproducibility and integrity

Artifact	SHA-256
alignment-lab-v1.json	584752dcc91654109c357b8ebb12681a12a9c1476a9ba539dd35e4d860a22ef
alignment-lab-v1-task-results.jsonl	8d7fc723436d7377d196fc44046d960e3cb7f0aa81e03d49ef05b627eb84630f
alignment-lab-v1-state-supervision.json	9e08129ba4cd9e460c189b94b4e421d881ba69e3938f02eac95d251f50c88788
alignment-lab-v1.yaml	71307dbfe9a5bb20c686307cafce8bd254c07af8b69c1bf1c6ec0dbf53a8cde0
gpu-calc-hard-equal-update-v2.json	53fc1d4d5b7adee09618d77ad62d4086ba56b78569832d6fc7c3bcd5c2695bbc

The machine-readable result binds all 18 arms, DPO provenance, the disjoint baseline recovery and 864 task-level records. Exact generation tests compare every public figure, Markdown report and 36 Alignment Card files byte for byte. The immutable calculator result remains unchanged.

13. Conclusion

Alignment Lab v1 does not find an incremental alignment benefit because its SFT starting point already saturates the deterministic suite. The useful outcome is a calibrated no-continuation decision and a complete cost/utility record. miniVERL should be used to justify when online teacher-student training is worth running—not to presume that OPD should replace SFT.