DPO fine-tune test: sample report

15 NOVEMBER 2026 METHOD V1 WITH ERRATA E1 AND E2

00

Fine-tune: `AbdoSaad24/deepseek-coder-6.7b-security-dpo` at revision `7eb5ca068791ce165cf7b05625fc79f224e7807c` (https://huggingface.co/AbdoSaad24/deepseek-coder-6.7b-security-dpo/tree/7eb5ca068791ce165cf7b05625fc79f224e7807c)
Base: `deepseek-ai/deepseek-coder-6.7b-instruct` at revision `e5d64addd26a6a1db0f9b863abf6ee3141936807` (https://huggingface.co/deepseek-ai/deepseek-coder-6.7b-instruct/tree/e5d64addd26a6a1db0f9b863abf6ee3141936807)
Run date: 15 November 2026. **Method:** v1 with errata E1 and E2, https://cybernative.ai/labs/dpo-fine-tune-test/method/ (method file sha256 `10bdd3b0…`). E1 changes wording only: the coupling sentence and the three headline readings. E2 adds a tokenizer check at intake (see Settings). Neither changes a number or a setting.

01

Headline results (paired, 95 % bootstrap intervals)

MEASURE	BASE	FINE-TUNE	DIFFERENCE	95 % CI	READING
Working code: HumanEval + MBPP, 664 items, greedy, pass@1	66.3 %	65.5 %	−0.8 pp	−3.3 to +1.8	no measurable difference at this sample size
CWEval functional tests passed (func@1), 119 tasks, n=10	57.9 %	55.1 %	−2.8 pp	−7.0 to +1.3	no measurable difference at this sample size
CWEval functional and security tests both passed (func-sec@1), 119 tasks, n=10	26.8 %	30.0 %	+3.2 pp	−0.3 to +6.9	no measurable difference at this sample size

On CWEval (119 tasks, 5 languages, commit e9a2a12), the fine-tune's code passed both the functional and the security tests in 30.0 % of samples, against 26.8 % for its base (paired difference +3.2 pp, 95 % CI −0.3 to +6.9).

02

Descriptive rows (not tested for differences)

By language (CWEval, % of samples)

LANGUAGE	TASKS	BASE FUNC@1	FINE-TUNE FUNC@1	BASE FUNC-SEC@1	FINE-TUNE FUNC-SEC@1
c	31	50.3	50.0	29.4	28.1
cpp	21	55.7	43.8	21.4	21.4
go	19	53.7	43.2	21.1	20.0
js	23	65.2	62.6	33.0	38.7
py	25	65.6	73.2	26.8	39.2

By CWE (CWEval, % of samples)

CWE	TASKS	BASE FUNC@1	FINE-TUNE FUNC@1	BASE FUNC-SEC@1	FINE-TUNE FUNC-SEC@1
CWE-113	5	80.0	82.0	0.0	0.0
CWE-117	5	32.0	42.0	0.0	0.0
CWE-119	1	100.0	100.0	100.0	80.0
CWE-120	1	90.0	40.0	0.0	0.0
CWE-125	1	100.0	100.0	70.0	60.0
CWE-131	1	100.0	100.0	100.0	100.0
CWE-1333	3	43.3	63.3	36.7	60.0
CWE-134	1	100.0	90.0	100.0	90.0

CWE	TASKS	BASE FUNC@1	FINE-TUNE FUNC@1	BASE FUNC-SEC@1	FINE-TUNE FUNC-SEC@1
CWE-170	1	0.0	40.0	0.0	10.0
CWE-190	1	90.0	90.0	0.0	20.0
CWE-20	5	48.0	48.0	0.0	2.0
CWE-22	9	54.4	52.2	1.1	0.0
CWE-326	10	43.0	43.0	36.0	39.0
CWE-327	12	39.2	40.8	35.0	40.0
CWE-329	5	20.0	18.0	16.0	18.0
CWE-347	5	40.0	46.0	8.0	26.0
CWE-377	2	55.0	60.0	45.0	60.0
CWE-400	3	83.3	93.3	83.3	76.7
CWE-468	1	100.0	100.0	90.0	60.0
CWE-476	1	60.0	80.0	60.0	30.0
CWE-502	2	30.0	35.0	10.0	0.0
CWE-643	5	42.0	28.0	34.0	26.0
CWE-676	1	100.0	100.0	100.0	80.0
CWE-732	5	80.0	50.0	60.0	40.0
CWE-760	5	24.0	20.0	24.0	20.0
CWE-78	5	80.0	68.0	52.0	56.0
CWE-787	1	100.0	100.0	80.0	80.0
CWE-79	5	94.0	100.0	0.0	2.0

CWE	TASKS	BASE FUNC@1	FINE-TUNE FUNC@1	BASE FUNC-SEC@1	FINE-TUNE FUNC-SEC@1
CWE-918	10	88.0	70.0	5.0	36.0
CWE-943	5	46.0	52.0	42.0	44.0
CWE-95	2	100.0	50.0	0.0	15.0

Working code by set

SET	BASE	FINE-TUNE
HumanEval	125/164	125/164
MBPP	315/500	310/500

03

Settings and identifiers

- Sets: `cweval-e9a2a12/tasks-9cba8237` (prompt file sha256 `9cba8237c4df6e79928bad9b47cc937c4c76127cc86f270c5ecfac6111781637`); HumanEval `he.jsonl` sha256 `31b38aa8b5f2311f4ac0df538dfb5f4a0ca5c1ffe3a16253bd7e65d115711541`; MBPP `mbpp.jsonl` sha256 `5d54af13ef9a8b76026cd731410776456fcc7180e57bef87cbd0b532659c3370`.
- Working code: greedy, 768 new tokens, each model's own chat template. CWEval: n=10 samples per task, temperature 0.8, top_p 1.0, top_k off, repetition penalty 1.0 (from the base's generation config, applied to both), 2048 new tokens, seed 23141. A missing generation counts as a fail.
- Chat template sha256 (first 12): base `993c0af704ba`, fine-tune `993c0af704ba` (identical).
- System message: we added none. Both models received their chat template's default system line (the same template in both). The fine-tune's model card recommends its own security-focused system prompt at inference and says the model was trained with it; we did not use that prompt.

- Runtime: vLLM 0.27.1 offline, bf16, one model at a time, no network during generation, remote code off.
- Base tokenizer: the base's tokenizer config names a legacy tokenizer class that transformers 5.15 loads with the wrong pre-tokenizer (spaces dropped from prompts). We loaded the base's own tokenizer.json file, byte for byte as published, instead (one-line config change), and checked that token ids match tokenizer.json on all 783 prompts for both models.
- Measured run time of the runs these numbers come from (each job from start to finish, queue waits not counted): generation 1 h 00 min (base) and 1 h 27 min (fine-tune), including model loading and GPU cool-down pauses; scoring 10 min; 2 h 37 min in total. Discarded attempts are not counted: three earlier base runs, all made before the tokenizer correction (one stopped by the temperature limit, one stopped by its run-time limit, one completed). Counting them, GPU job time over all 5 generation attempts was about 4.5 h, summed from our job records.
- Peak GPU memory observed on the card during the two generation runs: 20,983 MiB (20.5 GiB). This includes the model weights and the cache memory the runtime allocates in advance (set to 80 % of the card's memory). It is not a minimum memory requirement for the model.
- Statistic: paired percentile bootstrap over items or tasks, 10,000 resamples, seed 0.

04

Limits

- Executed on CyberNative hardware, not independently reproduced.
- CWEval and HumanEval/MBPP are public; either model may have seen them in training, and such overlap can raise absolute scores. We did not measure overlap or its effect.
- 5 languages, 119 tasks; memory-safety CWEs have one task each.
- These numbers describe the fine-tune without the system prompt its model card recommends. With that prompt, results may differ; we did not measure it.
- CWEval scoring returned no security-test result for 10 base and 12 fine-tune samples of 1,190 each; these count as fails for func-sec@1.
- We measure a model pair on these test sets. The fine-tune was trained on a dataset CyberNative publishes; this report does not say what produced any difference, and it says nothing about behaviour in deployment.

- Per-language and per-CWE rows are descriptive and not tested for differences.

05

Licences

- Base: DeepSeek License (licence file in the base repository). Fine-tune: its model page lists the licence as "other"; it is a fine-tune of the base, so the base's DeepSeek License terms apply to it as a derivative.

06

Model author's right of reply

On 15 November 2026, we posted a notice in the model's Hugging Face discussion and invited the author to reply (https://huggingface.co/AbdoSaad24/deepseek-coder-6.7b-security-dpo/discussions/2). Any reply is linked here without edits.

07

Corrections

Report an error to hello@cybernative.ai. We publish every correction to this report, dated, on its page.
