ARJHINETY/OPENGRAD · AUDIT OF COMMIT FA7C510 · CORRECTIONS THROUGH A18E45F · 2026-09-13

OpenGrad claim audit

Every number and status claim in the repository's Markdown, the Hugging Face cards under arrochi112 and the project site was checked against the committed result artifacts by recomputing it from the JSON, not by reading the prose. About 990 claims were checked and about 890 held up. This document lists the 93 that did not, and how each was resolved.

93	13	42	38	93	0	0
findings	high	medium	low	resolved	partly resolved	open

The regression is not only an SFT effect.
DPO applied directly to the base model (M1-v1, no SFT) still refuses 70.7% of zero-shot GSM8K. The regression is associated with tool-policy post-training on When2Call-derived data; causation is not established.

Public cards omitted the regression.
The M0 and M1-v2 model cards presented the checkpoints without GSM8K zero-shot 67.4% → 0%, MMLU-Pro 49.0% → 37.0% or IFEval 67.8% → 45.8%.

The quantization pick was flipped.
The pre-registered rule selects Q6_K, not Q8_0, and the gate sits inside rerun noise: the unquantized reference itself fails it on an H200 rerun.

M1-v2 was promoted by a gate change.
It fails the v3 gate that rejected M0 and passes v4, introduced after M0 was evaluated, which M0 also clears. Its gain over M0 is +0.0078 call_f1 (7 of 453 calls, one seed).

The refusal-label audit measured the wrong corpus.
On the Canonical-v2 corpus M0 trained on, 18,114 of 173,237 records (10.5%) are refusal-shaped targets labelled ANSWER, including 4,038 of 6,505 When2Call rows (62.1%).

An ablation's provenance was misstated.
The matched-exposure arm launched from 500cb4e with a dirty tree, and saw 1.19× the reference's supervised tokens, not ~1.0×.

Method: four independent read-only audits (top-level docs; training and eval reports; capability and quantization reports; data and public surfaces) recomputed values in Python from runs/*/eval metrics, the ledgers, results/final_campaign_verdict.json, results/benchmarks/h200 and results/quantization, re-ran the checkpoint-selection and promotion gates, and hashed files against the freeze manifests. Frozen, hash-pinned reports were not edited; they are corrected by reports/ERRATA.md. Each resolution below was then checked independently against the corrected files. Rounding that does not change a conclusion is not listed.

Top-level docs 29 findings

SEVERITY	LOCATION	AS WRITTEN	WHAT THE ARTIFACTS SHOW	RESOLUTION
HIGH #49	README.md:247-249 UNSUPPORTED public	“B0, M0 lineage … and the M1 DPO runs are recorded and reproducible from committed configs”	The only repeat ever attempted (m1_dpo_v1_restore) is marked NON_REPRODUCIBLE_REPEAT and its curves diverge. No other run has a repeat.	Resolved README.md (3bf8b67)

SEVERITY	LOCATION	AS WRITTEN	WHAT THE ARTIFACTS SHOW	RESOLUTION
HIGH #50	README.md:100-101 CHERRY-PICK **public**	“M0 not promoted (recall-regression gate) next to M1 PROMOTED”	v4 compares against the M0 parent, not B0. Under v3 every M1-v2 checkpoint is REJECT; DEV recall 0.7257 vs B0 0.9715.	Resolved README.md (3bf8b67)
MED #51	README.md:139-141 · docs/EVALUATION.md:29-31 MISLEADING **public**	“Re-run at 2048 tokens, the real gap is 12.0pp”	Base still truncates on 21.8% of items vs M0's 6.3%, and the adversarial interval is 6.4–33.1pp. EVALUATION.md breaks its own no-point-estimate rule here.	Resolved README.md (3bf8b67); docs/EVALUATION.md (c0121d0)
MED #52	README.md:130-131 MISLEADING **public**	“answers 55.5% of the same 1,319 questions with 8 exemplars”	The answer rate is 100%; 55.5% is the accuracy.	Resolved README.md (3bf8b67)
MED #56	ROADMAP.md:146 INTERNAL-CONTRADICTION **public**	“SFT: 2 negative, 1 partial, 1 definitive, 2 ablations”	There are 5 SFT arms and only 1 negative. DOCUMENTATION_STATE_AUDIT says this was corrected, but it wasn't.	Resolved ROADMAP.md (3bf8b67); docs/DOCUMENTATION_STATE_AUDIT.md addendum (c0121d0)
MED #57	ROADMAP.md:36-37,54-55 STALE **public**	“external benchmark families remain FROZEN_NOT_EXECUTED”	IFEval, GSM8K and MMLU-Pro were executed (H200 campaign).	Resolved ROADMAP.md (3bf8b67)
MED #58	ROADMAP.md:91-93 MISLEADING **public**	“21,749 records (10.0% of 217,903) in the supervision”	The audit read the normalization-v1 sources, not M0's corpus. See the refusal-audit finding under data.	Resolved ROADMAP.md (3bf8b67)
MED #59	ROADMAP.md:28 · docs/publishing/huggingface-datasets.md:11,13 INTERNAL-CONTRADICTION **public**	“v2-final publication record is toolpolicy-canonical-v2-publication.json”	That file describes the 103,036-record M0 snapshot. The v2-final record is hf-publication-2026-09-11.json.	Resolved ROADMAP.md (3bf8b67); docs/publishing/huggingface-datasets.md (c0121d0)
MED #61	docs/EXPERIMENT_RESULTS.md:36-38 INTERNAL-CONTRADICTION	“the corrected model is still slightly behind on call_f1”	The table above shows M0-final +0.128 and M1-v2 +0.136 over B0.	Resolved docs/EXPERIMENT_RESULTS.md (3bf8b67)
MED #65	docs/…/DPO_ARCHITECTURE.md:25-28,36-41 UNSUPPORTED	“the reference is never the original base, enforced by hash rejection; two data sources”	M1-v1 used reference: initial_policy (the base), dpo_runner.py:232 allows it, and no hash check exists. M1-v2 used 481 pairs: 240 When2Call + 241 base/M0 disagreements.	Resolved docs/DPO_ARCHITECTURE.md (c0121d0)
MED #66	docs/…/TEACHER_SELECTION.md:10-12,19 UNSUPPORTED	“fail-closed tokenizer gate; Qwen2.5-32B shares the 248,064-token vocabulary”	agent_cli.py:711 hardcodes mock_compatible=True, so the gate always passes. The code's own mock gives Qwen2.5 a vocabulary of 151,936.	Resolved docs/TEACHER_SELECTION.md (c0121d0)

SEVERITY	LOCATION	AS WRITTEN	WHAT THE ARTIFACTS SHOW	RESOLUTION
MED #67	docs/…/ON_POLICY_DISTILLATION.md:53-61 · TRAINING_LIFECYCLE.md:15,40-44 STALE	“OPD supports three modes; training outputs checkpoints”	The live path raises NotImplementedError; M2 was closed as NOT RUN.	Resolved docs/ON_POLICY_DISTILLATION.md, docs/TRAINING_LIFECYCLE.md (c0121d0)
MED #68	docs/…/INCIDENT_LOG.md:110-113 UNSUPPORTED	“the restore run's regenerated checkpoints and all measurements are committed”	Only metrics.json/curve.json are in git; there are no checkpoints or predictions.	Resolved docs/INCIDENT_LOG.md (c0121d0)
MED #69	docs/evaluation/CHECKPOINT_SELECTION.md:23,44-48 STALE	“confirmatory set not yet materialized; promotion provisional”	The 1,277-example confirmatory partition exists and was used.	Resolved docs/evaluation/CHECKPOINT_SELECTION.md (c0121d0) Line 23 edited. Lines 44-48 are kept in the present tense, with a dated update at the end of the same section.
MED #71	docs/architecture/repository.md:35 STALE	“No … external-benchmark result exists”	3 were executed in the H200 campaign.	Resolved docs/architecture/repository.md (c0121d0)
MED #72	docs/benchmarks/README.md:63 WRONG-NUMBER	“When2Call scored for B0 and 12 candidate checkpoints”	At least 29 checkpoints (32 including the repeat) across runs/*/eval. This row survived the 2026-09-13 edit to this file.	Resolved docs/benchmarks/README.md (c0121d0) The new text says 31 checkpoints (28 without the repeat), 'with committed metrics'. That matches runs/*/eval. The finding's 29/32 also counted v1 step 400, which appears only in curve.json.
MED #76	docs/research/methodology.md:3 UNSUPPORTED	“accepted checkpoints require … variance/confidence”	The promoted M1-v2 is single-seed with no interval; its +0.0078 call_f1 over M0 is 7 more correct calls out of 453.	Resolved docs/research/methodology.md (c0121d0) The standard sentence is kept, with a new paragraph directly after it stating that M1-v2 does not meet it.
LOW #53	README.md:93-99 MISLEADING **public**	“B0 0.6191/0.9722 (full set) beside M0/M1 on the confirmatory partition”	B0 on the same partition is 0.6264/0.9735. The mismatch is disclosed.	Resolved README.md (3bf8b67)
LOW #54	README.md:203 · docs/publishing/huggingface-datasets.md:12 UNSUPPORTED **public**	“Canonical-v1 pinned by B0”	B0's record pins only the held-out manifest.	Resolved README.md (3bf8b67); docs/publishing/huggingface-datasets.md (c0121d0)

SEVERITY	LOCATION	AS WRITTEN	WHAT THE ARTIFACTS SHOW	RESOLUTION
LOW #55	README.md:241 WRONG public	“all throughput figures are A100-80GB”	H200 throughput tables exist in H200_BENCHMARK_RUN.md.	Resolved README.md (3bf8b67)
LOW #60	ROADMAP.md:127 WRONG-NUMBER public	“this entire diagnosis ($8.90)”	The campaign total is $11.39; $8.90 is the continuation only.	Resolved ROADMAP.md (3bf8b67)
LOW #62	docs/EXPERIMENT_RESULTS.md:27 CHERRY-PICK	“partial-v2 call_f1 0.5995 / macro recall 0.6416”	Best of 4 checkpoints, chosen on the same set; the final checkpoint scores 0.5292. Unlike the M1-v1 row, it isn't labelled best.	Resolved docs/EXPERIMENT_RESULTS.md (3bf8b67)
LOW #63	docs/EXPERIMENT_STATUS.md:45 INTERNAL-CONTRADICTION pinned: generated (byte-identity test)	“v2corpus Published: —”	The HF repo M0-SFT-CorpusV2 exists with 4 checkpoints. This file is generated: fix the generator.	Resolved docs/EXPERIMENT_STATUS.md via scripts/reporting/generate_experiment_status.py (3bf8b67) Generator output is byte-identical to the committed file.
LOW #64	docs/EXPERIMENT_STATUS.md:43 WRONG pinned: generated (byte-identity test)	“CorpusV1-evaluation linked as a model”	It's a dataset, so the model link errors.	Resolved docs/EXPERIMENT_STATUS.md via scripts/reporting/generate_experiment_status.py (3bf8b67)
LOW #70	docs/evaluation/CHECKPOINT_SELECTION.md:25 WRONG-NUMBER	“3,952 before the two quarantined items”	3,652	Resolved docs/evaluation/CHECKPOINT_SELECTION.md (c0121d0)
LOW #73	docs/models/renderer-matrix.md:5,7 STALE / UNSUPPORTED	“Qwen3.5-2B: no training run; 210,874 candidates”	Training ran, and 210,874 appears in no committed artifact.	Resolved docs/models/renderer-matrix.md (c0121d0)
LOW #74	docs/inference/pre-gpu-surfaces.md:3 STALE	“quantization … remains unexecuted”	The GGUF PTQ ladder was executed and closed.	Resolved docs/inference/pre-gpu-surfaces.md (c0121d0)
LOW #75	docs/publishing/huggingface-datasets.md:15 WRONG	“Canonical-v2 final is the corpus the definitive M0 and M1 DPO ran on”	M1-v2 trained on m1_calibration_preference_pairs_v1 (481 pairs).	Resolved docs/publishing/huggingface-datasets.md (c0121d0)
LOW #77	docs/DOCUMENTATION_STATE_AUDIT.md:149-150 UNSUPPORTED	“no contradicting statement remains after these fixes”	ROADMAP:146, benchmarks/README.md:63 and renderer-matrix.md already contradicted it.	Resolved docs/DOCUMENTATION_STATE_AUDIT.md addendum (c0121d0)

Training & eval reports 23 findings

SEVERITY	LOCATION	AS WRITTEN	WHAT THE ARTIFACTS SHOW	RESOLUTION
HIGH #1	reports/M0_V2_FINAL_MINUS_XLAM_ABLATION_EXECUTION.md:55 WRONG-NUMBER pinned: m0-v2-minus-xlam-freeze	“launch commits: matched 0807fa3 (both arms from clean trees)”	Launched at 12:49 from 500cb4e with git_dirty: True (experiment.json, events.jsonl). 0807fa3 is the 13:26 commit that recorded the finished run.	Resolved reports/ERRATA.md (d55b801)
HIGH #2	POSTTRAINING_PHASE_SUMMARY.md:3,12 · M1_DPO_EVALUATION.md:3,53 · M2_DECISION.md:10 · docs/EXPERIMENT_RESULTS.md:12,31 CHERRY-PICK	“M0 not promoted → M1 promoted”	M1-v2 fails v3, the gate M0 failed, on all 4 DEV checkpoints. It was promoted under v4, committed after M0's confirmatory results, and M0@1800 also passes v4. None of this is disclosed.	Resolved POSTTRAINING_PHASE_SUMMARY.md, M1_DPO_EVALUATION.md, M2_DECISION.md (c0121d0); docs/EXPERIMENT_RESULTS.md (3bf8b67)
HIGH #3	docs/EXPERIMENT_RESULTS.md:27 WRONG	“partial-v2 — Regression: None measured”	Recall 0.9722 → 0.505 (−0.467) and call_f1 −0.020 at checkpoint 1200. The selection-rule doc itself records the −0.467.	Resolved docs/EXPERIMENT_RESULTS.md (3bf8b67)
MED #4	M0_CANONICAL_V2_FINAL_EVALUATION.md:106-107 · … EXECUTION_REPORT.md:197-198 · ablation eval:108 · EXPERIMENT_RESULTS.md:47-49 GATE-MISAPPLIED pinned: m0-final-freeze, m0-v2-minus-xlam-freeze	“Every checkpoint … fails regression.call_recall”	Checkpoint 600 fails only over_call_rate (recall delta −0.059).	Resolved reports/ERRATA.md (d55b801) (eval 105-107, exec 197-199, ablation eval 107-109); docs/EXPERIMENT_RESULTS.md (3bf8b67)
MED #5	M0_CANONICAL_V2_FINAL_EVALUATION.md:106 · …EXECUTION_REPORT.md:208 UNSUPPORTED pinned: m0-final-freeze	“a calibrated model cannot satisfy both (over-call ≤ 0.20 and recall ≥ B0 − 0.10)”	The two constraints are measured on disjoint example sets and aren't mutually exclusive. It's only an observation that no checkpoint here met both.	Resolved reports/ERRATA.md (d55b801)
MED #6	ablation evaluation:65 · execution:29 · M0_V2_FINAL_ABLATION_DESIGN.md:18 WRONG-NUMBER pinned: m0-v2-minus-xlam-freeze	“matched arm at ~1.0× exposure; measured supervised-token exposure held fixed”	Logged supervised tokens: 6,743,788 vs 5,678,531 (1.19×); 1.36× passes over the corpus. Batches are token-budgeted, so steps don't hold tokens fixed.	Resolved reports/ERRATA.md (d55b801) (execution:29, design:18, evaluation:60-72)
MED #7	ablation evaluation:60-73 MISLEADING pinned: m0-v2-minus-xlam-freeze	“exposure confound does not explain the result”	The table shows full horizons, but the metrics come from checkpoints 1200/1060, which saw 3.80M/3.36M supervised tokens: fewer than the reference's selected 1800 (4.27M).	Resolved reports/ERRATA.md (d55b801)

SEVERITY	LOCATION	AS WRITTEN	WHAT THE ARTIFACTS SHOW	RESOLUTION
MED #8	ablation evaluation:101-103 WRONG pinned: m0-v2-minus-xlam-freeze	“both arms sharpen precision and suppress over-calling while recall keeps falling”	Matched arm, DEV: recall rises 0.2352 → 0.3943 (530 → 1590), precision falls 0.8285 → 0.7804, and over-call rises 0.0268 → 0.0601.	Resolved reports/ERRATA.md (d55b801)
MED #9	M1_DPO_EVALUATION.md:21 · POSTTRAINING_PHASE_SUMMARY.md:21 · ablation evaluation:28 INTERNAL-CONTRADICTION pinned: m0-v2-minus-xlam-freeze (ablation eval only)	“B0 row in confirmatory tables: 0.6191 / 0.4542 / 0.9722 / 0.6425 / 0.1009 / 0.0131”	These are the full 3,650-example numbers. B0 on the confirmatory partition is 0.6264 / 0.4618 / 0.9735 / 0.6238 / 0.1186 / 0.0177.	Resolved M1_DPO_EVALUATION.md, POSTTRAINING_PHASE_SUMMARY.md (c0121d0); reports/ERRATA.md (d55b801) (ablation eval:28)
MED #10	M1_DPO_EXECUTION_REPORT.md:19-20 WRONG-NUMBER	“141 CALL pairs from local disagreements; 140 ANSWER + 100 + 100 curated”	Local: 141 CALL + 100 ANSWER (241). Curated: 40 ANSWER + 100 + 100 (240); only 73 curated ANSWER pairs existed.	Resolved reports/M1_DPO_EXECUTION_REPORT.md (c0121d0)
MED #11	M0_CANONICAL_V2_FINAL_EXECUTION_REPORT.md:97,117-118,122 WRONG-NUMBER pinned: m0-final-freeze	“38,400 examples seen; 0.237 vs 0.377 epochs; 7.96M vs 11.08M supervised tokens”	Logged: 27,372 examples and 5,678,531 tokens. Actual epochs 0.169 / 0.272. 38,400 was a pre-launch nominal estimate.	Resolved reports/ERRATA.md (d55b801)
MED #12	M0_CANONICAL_V2_FINAL_EVALUATION.md:44-46 UNSUPPORTED pinned: m0-final-freeze	“confirmatory rows for M0-v1, M1-v1 and partial-v2@1200”	No per-partition artifact is committed, the predictions are untracked, and the M0-v1/M1-v1 checkpoints aren't named.	Resolved reports/ERRATA.md (d55b801)
MED #13	OPENWEIGHTS_TRANSFER_EVALUATION.md:53-59,68 STALE / UNSUPPORTED	“General capability regressed … no Qwen3.5-2B baseline exists”	Base scores the same 5/7 (3/5 general, 2/2 tool) and fails multi-step-change itself (BASE/sentinel_scores.json).	Resolved reports/ERRATA.md (d55b801)
LOW #14	OPENWEIGHTS_TRANSFER_EVALUATION.md:12-14 MISLEADING	“the failures are not a tool-policy failure”	Both failures were parsed as UNSUPPORTED, which is the routing class the policy trains.	Resolved reports/ERRATA.md (d55b801)
LOW #15	M1_DPO_EXECUTION_REPORT.md:60-62 INTERNAL-CONTRADICTION	“every checkpoint records the parent checkpoint ID”	checkpoint_registry.json has parent_checkpoint: null for every M1-v2 checkpoint.	Resolved reports/M1_DPO_EXECUTION_REPORT.md (c0121d0)
LOW #16	ablation evaluation:38 · docs/EXPERIMENT_RESULTS.md:30 WRONG-NUMBER pinned: m0-v2-minus-xlam-freeze (ablation eval)	“matched − reference over-call −0.105”	0.0558 − 0.1505 = −0.0947.	Resolved reports/ERRATA.md (d55b801); docs/EXPERIMENT_RESULTS.md (3bf8b67)

SEVERITY	LOCATION	AS WRITTEN	WHAT THE ARTIFACTS SHOW	RESOLUTION
LOW #17	ablation evaluation:30 WRONG-NUMBER pinned: m0-v2-minus-xlam-freeze	“fixed-arm confirmatory parse_ok 1.0000”	0.999217	Resolved reports/ERRATA.md (d55b801)
LOW #18	ablation evaluation:107 UNSUPPORTED pinned: m0-v2-minus-xlam-freeze	“The failing check is regression.call_recall”	call_f1_retention also fails, on every matched DEV checkpoint and on fixed 1800/2400.	Resolved reports/ERRATA.md (d55b801)
LOW #19	M0_CANONICAL_V2_FINAL_EVALUATION.md:95 WRONG-NUMBER pinned: m0-final-freeze	“confirmatory more favourable on four of six dimensions”	Five of six; only unsupported is worse.	Resolved reports/ERRATA.md (d55b801)
LOW #20	docs/EXPERIMENT_RESULTS.md:28 CHERRY-PICK	“M0-final regression column lists precision and over-call vs partial-v2”	Leaves out unsupported −0.080, the largest regression, and clarification −0.027.	Resolved docs/EXPERIMENT_RESULTS.md (3bf8b67)
LOW #21	M0_SFT_EXECUTION_REPORT.md (lines 10, 44, 175, 364-366, 483-484, 534) INTERNAL-CONTRADICTION	“no M2 experiment was created; no DPO run started; @1200 recall 0.778/0.636; 12 checkpoints; under_call 0.0185 better; 4 SFT runs”	The ledger shows a mock M2 run created on 09-10, and §5 of the same report runs DPO. Correct values: 0.7906/0.6293, 13 checkpoints, 2 runs; lower under_call is better, so 0.0185 is worse.	Resolved reports/M0_SFT_EXECUTION_REPORT.md (c0121d0)
LOW #22	SUPERVISION_CONTRACT_REPORT.md:230-232 STALE	“a development-set experiment, not a confirmatory one”	A pre-registered confirmatory partition was used afterwards.	Resolved reports/SUPERVISION_CONTRACT_REPORT.md (c0121d0) Original sentence kept; dated update note appended directly after it.
LOW #23	M1_DPO_EVALUATION.md:14 · ablation evaluation:98 MISLEADING pinned: m0-v2-minus-xlam-freeze (ablation eval)	“checkpoint 30 / 1060 chosen by tie-break”	Each was also the outright macro maximum, so the tie-break decided nothing.	Resolved M1_DPO_EVALUATION.md (c0121d0); reports/ERRATA.md (d55b801) (ablation eval 91, 97-98)

Capability & quantization 25 findings

SEVERITY	LOCATION	AS WRITTEN	WHAT THE ARTIFACTS SHOW	RESOLUTION
HIGH #24	reports/PTQ_PHASE_CLOSURE.md:22,29 · QUANTIZATION_REPRODUCTION.md:132 GATE-MISAPPLIED pinned: ptq_phase_closure_v1.json, PRESERVED_STATE_v1	“Q8_0 = RECOMMENDED_RELEASE, Q6_K = MEMORY_OPTIMIZED_RELEASE”	The pre-registered rule is smallest passing format wins, and the closure manifest records recommended_release_rung: Q6_K. The roles were hard-coded after scoring (close_ptq_phase.py:48).	Resolved reports/ERRATA.md (d55b801) (PTQ_PHASE_CLOSURE); QUANTIZATION_REPRODUCTION.md (c0121d0); README.md (3bf8b67)

SEVERITY	LOCATION	AS WRITTEN	WHAT THE ARTIFACTS SHOW	RESOLUTION
HIGH #25	reports/QAD_DECISION.md:85-91 WRONG-NUMBER / OVERSTATED-CAUSATION	“the DPO stage is what produced the low over_call_rate of 0.1553”	M0 0.1505 < M1-v2 0.1529, so DPO raised over-call slightly. 0.1553 is the llama.cpp BF16 figure.	Resolved reports/QAD_DECISION.md (c0121d0)
HIGH #26	reports/H200_BENCHMARK_RUN.md:103-105 · QAD_DECISION.md:20-23 CHERRY-PICK / UNSUPPORTED pinned: append-only original section	“rerun deltas within ±0.011; reference is confirmed”	The H200 vLLM BF16 rerun itself fails the gate: recall 0.7638 < 0.7671. Q6_K clears precision by one example (357/490 vs 356.96). Gate pass/fail is inside rerun noise, and no document says so.	Resolved reports/ERRATA.md (d55b801) (H200:103-105); reports/QAD_DECISION.md (c0121d0)
MED #27	results/benchmarks/h200/PRESERVED_STATE_v1.json REPRODUCIBILITY	“pins file hashes for the campaign state”	The pins were computed over Windows CRLF working-tree bytes, so they never match the LF bytes a clone checks out. The state can't be verified off the author's machine.	Resolved reports/ERRATA.md (d55b801) section 6
MED #28	reports/H200_BENCHMARK_RUN.md:240 STALE	“MMLU-Pro COMPLETE ×4 stages”	M1_DPO_HISTORICAL MMLU-Pro is UNMEASURED: the run was terminated (final_campaign_verdict.json).	Resolved reports/ERRATA.md (d55b801)
MED #29	reports/FINAL_CAMPAIGN_AUDIT.md:346-352 UNSUPPORTED / INTERNAL-CONTRADICTION	“both budget changes were pre-scoring amendments”	The 768-token MMLU-Pro run was scored first (38.0 / 36.8 / 36.9), and the table's own row says after observing truncation.	Resolved reports/FINAL_CAMPAIGN_AUDIT.md (c0121d0)
MED #30	reports/FINAL_CAMPAIGN_AUDIT.md:151,349-351 UNSUPPORTED	“raising IFEval to 2560 tokens removed the truncation confound”	At 2560 tokens, 63 (M1-v2), 67 (M0) and 47 (Base) of 541 still hit the cap. No report discloses IFEval truncation.	Resolved reports/FINAL_CAMPAIGN_AUDIT.md (c0121d0, a4f6f8c) Truncation counts (47/67/63/17 of 541) are disclosed, and the text now states that the caveat applies to every IFEval row in section 3, which do not restate it.
MED #31	reports/FINAL_CAMPAIGN_AUDIT.md:15-16,171-172,392 OVERSTATED-CAUSATION	“tool-policy SFT introduced two separable regressions; chronology rejects a DPO cause”	The verdict says ordering is not causation, and the Base edge also changes architecture and tokenizer. M1-v1 (DPO on base, no SFT) refuses 70.7%.	Resolved reports/FINAL_CAMPAIGN_AUDIT.md (c0121d0)
MED #32	reports/FINAL_CAMPAIGN_AUDIT.md:118-120 · docs/EVALUATION.md:46-48 WRONG	“Base gets trap-arithmetic and multi-step-change wrong; M0/M1-v2 refuse them”	M0/M1-v2 pass trap-arithmetic. They refuse multi-step-change and format-constraint (sentinel_scores.json).	Resolved reports/FINAL_CAMPAIGN_AUDIT.md, docs/EVALUATION.md (c0121d0)

SEVERITY	LOCATION	AS WRITTEN	WHAT THE ARTIFACTS SHOW	RESOLUTION
MED #33	reports/GENERAL_CAPABILITY_REGRESSION.md:10,88-89 MISLEADING	“conditional accuracy falls 21.3pp — the ability itself is worse”	This excludes 2,136 truncated Base items. On items both stages attempted the drop is 17.7pp, and no interval is given. The label still holds.	Resolved reports/GENERAL_CAPABILITY_REGRESSION.md via scripts/build_capability_report.py (c0121d0, a4f6f8c); FINAL_CAMPAIGN_AUDIT.md (a4f6f8c) 17.7pp is correct. Base's joint-attempt accuracy is 5,787/9,637 = 60.0% (first written as 60.1%, fixed in a4f6f8c in both files).
MED #34	reports/H200_BENCHMARK_RUN.md:117-120 OVERSTATED pinned: append-only original section	“engine change alone … 21 flips, nearly as much as quantization”	Hardware also differs (H200 vs A100), and 3 of the 21 flips are tokenizer-caused.	Resolved reports/ERRATA.md (d55b801)
MED #35	release/gguf/manifest.json STALE	“primary_mobile_target: Q4_K_M”	Q4_K_M is REJECTED_ACCURACY (gate recomputed).	Resolved reports/ERRATA.md (d55b801) section 4
LOW #36	reports/H200_BENCHMARK_RUN.md:90-91 OVERSTATED pinned: append-only original section	“the rerun removes the standing caveat from the entire quantization study”	It's a new H200 run with different aggregates (f1 0.7505 vs 0.7548), not the frozen per-example predictions.	Resolved reports/ERRATA.md (d55b801)
LOW #37	docs/PROMOTION_POLICY.md:35-36 WRONG-NUMBER	“promoted checkpoint scores 22.7pp below Base on IFEval”	M1-v2 is −22.0pp; 22.7 is M0's figure.	Resolved docs/PROMOTION_POLICY.md (c0121d0)
LOW #38	GENERAL_CAPABILITY_REGRESSION.md:66 · H200_BENCHMARK_RUN.md:223 MISLEADING	“conditional math accuracy not materially observable”	It's not measurable at 0-shot (M0 attempted none), but the 8-shot conditional drop is −14.1pp.	Resolved GENERAL_CAPABILITY_REGRESSION.md (c0121d0); reports/ERRATA.md (d55b801) (H200:223)
LOW #39	GENERAL_CAPABILITY_REGRESSION.md:128 · FINAL_CAMPAIGN_AUDIT.md:142 WRONG-NUMBER	“100% of Base's unattempted items at 768 tokens were truncations”	4,292 / 4,309 = 99.6%	Resolved GENERAL_CAPABILITY_REGRESSION.md, FINAL_CAMPAIGN_AUDIT.md (c0121d0)
LOW #40	FINAL_CAMPAIGN_AUDIT.md:162,224 INTERNAL-CONTRADICTION	“M1-v1 zero-shot refusal 71.0%”	By buckets it's 933/1,319 = 70.7%, the verdict's figure. 71.0% is the detector's count, and the same bullet mixes both.	Resolved reports/FINAL_CAMPAIGN_AUDIT.md (c0121d0)
LOW #41	FINAL_CAMPAIGN_AUDIT.md:126-130 UNSUPPORTED	“superseded results preserved, never deleted”	M1_DPO_HISTORICAL/superseded_mmlu_768/ is empty; the 49.7% survives only in the ledger.	Resolved reports/FINAL_CAMPAIGN_AUDIT.md (c0121d0)

SEVERITY	LOCATION	AS WRITTEN	WHAT THE ARTIFACTS SHOW	RESOLUTION
LOW #42	GENERAL_CAPABILITY_REGRESSION.md:247 · FINAL_CAMPAIGN_AUDIT.md:365,376 · H200_BENCHMARK_RUN.md:249 MISLEADING	“retained runs (results reported) $5.01; superseded = 44%”	Includes $0.757 for a run that returned no results. Non-reporting spend is $4.64 (52%).	Resolved GENERAL_CAPABILITY_REGRESSION.md, FINAL_CAMPAIGN_AUDIT.md (c0121d0); reports/ERRATA.md (d55b801) (H200:249)
LOW #43	GENERAL_CAPABILITY_REGRESSION.md:233-234 INTERNAL-CONTRADICTION	“engine choice produces measurable differences”	The verdict supports runtime choice: hardware is confounded.	Resolved GENERAL_CAPABILITY_REGRESSION.md (c0121d0)
LOW #44	QUANTIZATION_PTQ_EVALUATION.md:138-140 · QUANTIZATION_ENGINE_PARITY.md:149-151 · PTQ_PHASE_CLOSURE.md:84-85 STALE pinned: PTQ evaluation, engine parity, PTQ closure	“no per-example vLLM↔llama.cpp agreement is claimed anywhere”	H200_BENCHMARK_RUN and FINAL_CAMPAIGN_AUDIT claim 0.9836 (21/1,277).	Resolved reports/ERRATA.md (d55b801) (PTQ eval, engine parity, PTQ closure)
LOW #45	reports/QUANTIZATION_HANDOFF.md:3, 62 STALE	“STOPPED BY USER before PTQ; no candidate evaluated”	9 rungs were scored and the phase is CLOSED.	Resolved reports/QUANTIZATION_HANDOFF.md (c0121d0) Line 3 status kept as history, with a 'Superseded' note directly under it; line 62 edited.
LOW #46	reports/MEDIATEK_PORT_STATUS.md:81 WRONG-NUMBER	“the ExecuTorch environment uses torch 2.10”	Exports ran on torch 2.14.0+cpu.	Resolved reports/MEDIATEK_PORT_STATUS.md (c0121d0)
LOW #47	reports/MEDIATEK_PORT_STATUS.md:10-11 UNSUPPORTED	“api.neuropilot paths return 404”	The probe records only two URLs, both HTTP 200.	Resolved reports/MEDIATEK_PORT_STATUS.md (c0121d0)
LOW #48	release/executorch/cpu/README.md:37-39 · README.md:236 MISLEADING / CHERRY-PICK	“Snapdragon needs two source patches; 99.98% of weights int4”	QNN is REJECTED_EXPORT even with the patches. Overall int4 coverage is 78.7%; the fp32 embedding is 65.7% of the .pte.	Resolved reports/ERRATA.md (d55b801) (release/executorch/cpu/README.md); README.md (3bf8b67)

Data & public surfaces 16 findings

SEVERITY	LOCATION	AS WRITTEN	WHAT THE ARTIFACTS SHOW	RESOLUTION
HIGH #78	HF arrochi112/OpenGrad-Qwen3.5-2B-M1-DPO-CanonicalV2-Final-v2 card · release/huggingface/…-m1-dpo-canonicalv2-final-v2/README.md MISSING-CAVEAT public	“Checkpoint 30 is PROMOTED; preserves the M0 calibrated frontier (no mention of general capability)”	GSM8K 0-shot 67.4% → 0% (100% refusal), 8-shot 70.4 → 55.5, MMLU-Pro 49.0 → 37.0, IFEval strict 67.8 → 45.8. The site links here as the model.	Resolved HF card (on the Hub, 2026-09-13) + template (a18e45f)
HIGH #79	HF arrochi112/OpenGrad-Qwen3.5-2B-M0-SFT-CanonicalV2-Final card · template MISSING-CAVEAT public	“presented as the healthy M0, with no capability caveat”	This is the stage where the regression first appears: GSM8K 0-shot 0%, IFEval 45.1%, MMLU-Pro 37.0%.	Resolved HF card (on the Hub, 2026-09-13) + template (a18e45f)
HIGH #80	results/…/sft_refusal_supervision_audit.json (scripts/audit_sft_refusal_supervision.py:41-46) · FINAL_CAMPAIGN_AUDIT.md:179-181 · ROADMAP.md:91,108 · site study.ts:96-108, ResearchCharts.tsx:329, ResearchPage.tsx:349 WRONG-NUMBER public	“the training corpus: 21,749 of 217,903 (10.0%); when2call-sft 7,490 / 14,829 (50.5%)”	The audit read normalization-v1: held-out rows, BUTTON, LoopTool and the v1 Glaive adapter. The same detector on published Canonical-v2 gives 18,114 of 173,237 (10.5%); When2Call 4,038 / 6,505 (62.1%). The 7,490 exceeds the 6,505 When2Call records M0 trained on.	Resolved results audit JSONs + script (88c5d41); FINAL_CAMPAIGN_AUDIT.md (c0121d0); ROADMAP.md (3bf8b67); site (deployed 2026-09-13) Canonical-v2 figures (18,114/173,237; When2Call 4,038/6,505) recomputed from the release shards and they match.
HIGH #81	results/benchmarks/checkpoint_ladder.json:217 · final_campaign_verdict.json:599 · GENERAL_CAPABILITY_REGRESSION.md:39 · FINAL_CAMPAIGN_AUDIT.md:50 · site study.ts:153, ResearchPage.tsx:256 INTERNAL-CONTRADICTION public	“M1_DPO_HISTORICAL sits on a different SFT parent (CorpusV2); SFT introduced it, preference training did nothing”	experiment.json and the Hub checkpoint-300 metadata show parent null and reference initial_policy: DPO directly on the base. It still refuses 70.7% of 0-shot GSM8K.	Resolved checkpoint_ladder.json, final_campaign_verdict.json (88c5d41); GENERAL_CAPABILITY_REGRESSION.md, FINAL_CAMPAIGN_AUDIT.md (c0121d0); ERRATA section 7 (capability_findings.jsonl) (d55b801); site (deployed 2026-09-13)
HIGH #82	HF arrochi112/OpenGrad-ToolPolicy-Canonical-v1 card (Hub bb295d8a) STALE + MISSING-CAVEAT public	“No B0 baseline has run and no model scores are reported”	The repo template was updated but the Hub card wasn't. The card also never says v1 collapses to 9 tool-call targets out of 55,719 trainable rows (unparsed Glaive), so anyone training on it would reproduce the collapse.	Resolved HF card (on the Hub, 2026-09-13) + template (a18e45f)
MED #83	site ResearchPage.tsx:89 · layout.tsx:15 · study.ts:38 · ResearchCharts.tsx:24,42 WRONG-NUMBER public	“F1 rose from 0.6191 to 0.7548; B0 labelled confirmatory n=1,277; over-call fell from 64.3%”	0.6191 and 64.3% are the full 3,650 set. Confirmatory B0: 0.6264 / P 0.4618 / R 0.9735 / over-call 0.6238. DPO itself adds only +0.0078 over M0.	Resolved site (deployed 2026-09-13)

SEVERITY	LOCATION	AS WRITTEN	WHAT THE ARTIFACTS SHOW	RESOLUTION
MED #84	HF ablation cards FixedCompute and MatchedExposure (templates :42, :53) WRONG-NUMBER public	“B0 row 0.6191 / 0.4542 / 0.9722 / 0.6425 / 0.1009 / 0.0131 under confirmatory, 1,277”	Should be 0.6264 / 0.4618 / 0.9735 / 0.6238 / 0.1186 / 0.0177.	Resolved HF card (on the Hub, 2026-09-13) + template (a18e45f)
MED #85	HF Canonical-v2 card quarantine table (template README.template.md:111) WRONG-NUMBER public	“UNRENDERABLE 2,095”	Once xLAM 1,231 and ToolACE 8,476 are listed separately, the remainder is Glaive 864. As written the table sums to 12,502, but 11,271 were dropped.	Resolved HF card (on the Hub, 2026-09-13) + template (a18e45f)
MED #86	HF Canonical-v2 card (template :121) · CANONICAL_V2_COMPLETION_REPORT.md:152,250 · PRE_SFT_READINESS_REPORT.md:134,146,161 · PRE_GPU_READINESS_REPORT.md:108 WRONG-NUMBER public	“3,950 distinct evaluation examples”	3,652 distinct (the 300 judge rows duplicate MCQ rows); 3,650 after quarantine.	Resolved HF card (on the Hub, 2026-09-13) + template (a18e45f); CANONICAL_V2_COMPLETION_REPORT.md, PRE_SFT_READINESS_REPORT.md, PRE_GPU_READINESS_REPORT.md (c0121d0)
MED #87	HF Canonical-v2 and v2-M0-snapshot cards, Pinned revision column MISLEADING public pinned: config pinned by m0-final and minus-xlam freezes	“Glaive 7e7e32f0…b6e8, ToolACE 7a7a6a2c3b1003c7”	These are raw-file SHA-256 values, not Hub revisions (those are e7f4b645… and 6bda777c…). Fix the card's column label, not the frozen config.	Resolved HF card (on the Hub, 2026-09-13) + template (a18e45f); config caveat in reports/ERRATA.md (d55b801) section 5
MED #88	HF arrochi112/OpenGrad-Qwen3.5-2B-M0-SFT-CorpusV2 card WRONG-NUMBER public	“ckpt-1200 recall 0.5093, 1800 0.3942; checkpoint-600 most balanced (0.6242); 2400 see below”	0.5050 and 0.3931. Checkpoint 1200 is most balanced (macro 0.6416). The 2400 metrics exist, and the card links the v1 dataset.	Resolved HF card (on the Hub, 2026-09-13)
MED #89	CorpusV2 card · v2-M0-snapshot card and template :54 OVERSTATED-CAUSATION public	“the two differ in exactly one respect / isolates the cause”	The added tool-call supervision is confounded with dropping xLAM, BUTTON and LoopTool (6 sources → 3; 55,719 → 101,785 trainable).	Resolved HF card (on the Hub, 2026-09-13) + template (a18e45f) (CorpusV2 card; v2-M0-snapshot card and template)
MED #90	reports/data-normalization-v1.md MISSING-CAVEAT	“Glaive 0 quarantined, FULL_DATA_VALIDATED; no blocker remains”	Silent on the Glaive adapter defect that collapsed v1 to 9 tool-call targets.	Resolved reports/data-normalization-v1.md (c0121d0)
LOW #91	v2-M0-snapshot card and template :34 INTERNAL-CONTRADICTION public	“9 of 55,719 trainable vs 49,423 of 103,036”	Mixes trainable and canonical counts in one row; like-for-like is 48,723 of 101,785 trainable (47.9%).	Resolved HF card (on the Hub, 2026-09-13) + template (a18e45f)

SEVERITY	LOCATION	AS WRITTEN	WHAT THE ARTIFACTS SHOW	RESOLUTION
LOW #92	HF M0-SFT-CanonicalV2-Final card MISLEADING **public**	“0.237 epochs against the previous run's 0.377”	These are nominal steps × 16. Actual examples seen give 0.169 and 0.272.	Resolved HF card (on the Hub, 2026-09-13) + template (a18e45f)
LOW #93	site ResearchPage.tsx:320-322 · docs/data/tool-use-mixture-methodology.md:3 · CANONICAL_V2_COMPLETION_REPORT.md xLAM table WRONG / STALE **public**	“promotion-partition modes 1,295 / 1,060 / 1,295; no model has been trained; 57,342 tool-call targets beside 56,090 trainable”	Confirmatory partition is 453 / 371 / 453; the methodology status is stale; trainable targets are 56,090.	Resolved site (deployed 2026-09-13); docs/data/tool-use-mixture-methodology.md, CANONICAL_V2_COMPLETION_REPORT.md (c0121d0) The methodology line 3 is kept, with a status update directly under it marking it historical.

Source: github.com/arjhinety/OpenGrad — reports/ERRATA.md and commits 88c5d41, d55b801, c0121d0, 3bf8b67, a4f6f8c, 55c75ae, a18e45f. Hugging Face cards: huggingface.co/arrochi112, all ten corrected cards uploaded 2026-09-13. Resolution status reflects the corrected repository, site and cards as of 2026-09-13.