The Supervision Tradeoff:
Format Scaffolds, Judgment Pleasing, and Anti-Calibration in Post-Training

Tad MacPherson*
Metavolve Labs, Research Division
San Francisco, California, United States
research@metavolvelabsinc.com

ORCID: 0009-0002-8659-7479

2026-04-24

Preprint disclosure: An earlier version of this work is publicly archived on Zenodo (concept DOI 10.5281/zenodo.19748276, which resolves to the latest version), released under CC BY 4.0. PeerJ permits prior preprint posting; this disclosure is provided per journal policy.

Reproducibility code: github.com/codex-curator/supervision-tradeoff **Data (Hugging Face):** huggingface.co/datasets/Metavolve-Labs/supervision-tradeoff

Abstract

We report a dual-metric evaluation of preference-aware fine-tuning on a vision-language base (Llama 3.2 11B Vision-Instruct) under three supervision regimes: naive SFT on a structurally-rich corpus, ORPO on structural-signal preference pairs (Reasoning-NEST v2), and ORPO on a surface-signal preference pair that differs only by an appended `Confidence: {high|medium|low}` footer. Evaluation uses both a 703-prompt deterministic verifier (MATH-500, HumanEval+, MMLU-Pro) and a 4-judge cross-lineage frontier panel (Claude Sonnet 4.6, Gemini 3.1 Pro, GPT-5.4, Grok-4-1) on the same 953-prompt out-of-distribution corpus. We find three distinct regime behaviors, none of which jointly maximize verifier and judge-panel performance. **Reading-order caveat for the rest of this abstract:** an initial single-seed reading produced a "naive SFT uniquely exhibits a verifier-vs-judge tradeoff" headline that a pre-registered 3-seed confirmatory replication subsequently dismantled by inverting the paired SFT−ORPO contrast; we retain the dismantled finding in the prose below because the *illusion of a single-seed tradeoff* is the paper's most consequential methodological result. **The untrained Base model dominates the 4-judge Borda aggregate** (0.651 normalized) **above every fine-tuned arm**. **Naive SFT (fp16-matched), at single seed, appeared to uniquely exhibit a verifier-vs-judge tradeoff** [WITHDRAWN post-replication, see below]: it appeared to win the deterministic verifier (34.6% ALL, +2.6pp above Base) at the cost of falling to 5th of 8 on the judge panel (Borda 0.486, −0.165 below Base). The verifier gain is benchmark-selective: HumanEval+ +7.0pp, MMLU-Pro +2.8pp, MATH-500 −1.0pp (SFT *loses* MATH). **Both structural-signal ORPO arms are dominated on both axes**: Gen-1.5-continue-ORPO (fp16-matched) falls to 30.0% verifier (−2.0pp below Base) and Borda 0.400 (−0.25 below Base), with a specific −7.6pp regression on HumanEval+. ORPO on this corpus at this scale produces no genuine tradeoff; it is strictly below Base on both metrics. **Surface-signal ORPO fails both axes and additionally anti-calibrates**: worst verifier accuracy (27.6%), strongest unanimous-last consensus (72 of 873 events), and Brier 0.296 against an empirical-base-rate predictor's 0.204 on

*Metavolve Labs, Research Division. All scientific claims and authorship accountability rest with the authors. AI systems used as infrastructure are enumerated in §5.6.

the 531 verifier-grounded emitted-footer prompts. Dense in-domain supervision (Group A sparse vs Group B dense, 11.1× per-record density gap) produces zero measurable OOD benefit: 32.9% vs 32.9% on verifier, 0.566 vs 0.575 on Borda. A meta-evaluation specimen (§8) documents a rubric-weighting disagreement between judges rather than judge failure: three of four judges weighted method-recognition-plus-completeness, one weighted correctness-of-geometric-direction on a truncated student output. The 3-seed confirmatory replication of the fp16-matched pair is *completed* and reported in §3.3: the single-seed Gen-1-continue 34.6% is not reproduced (new range [20.9%, 23.3%]), and the paired SFT−ORPO verifier contrast *inverts sign* (−4.25pp, 95% t-CI on three seeds $[-8.93, +0.44]$, crosses zero on the verifier-grounded intersection of $N = 361$ events present in all six runs). Accordingly we withdraw the "SFT wins the verifier" portion of the Tradeoff claim; the Base-dominates-Borda, surface-signal anti-calibration, and density-walk-back findings are separate measurements and survive replication. The new primary empirical finding from the replication is that post-training verifier accuracy on this base exhibits ∼11pp seed-to-seed variance on the SFT arm at matched hyperparameters— larger than any inter-regime difference the paper originally took to be informative. A 256-token inference ceiling differentially truncates outputs: HumanEval+ and MMLU-Pro carry $< 1.5\%$ truncation across all arms (the +7.0pp SFT HumanEval+ gain and the −7.6pp ORPO HumanEval+ regression are un-confounded), while MATH-500 Level 5 truncation varies 25–49% across arms (Base 25.7%, SFT 25.7%, ORPO 6.4%, Group B 48.6%); the MATH-500 sub-findings are scoped as confounded accordingly, and a 200-prompt re-run at 1024 tokens on A100 is pre-registered for the revision cycle if an area-chair challenge requires it. Contribution 5 has additionally been extended with an in-domain 2×2 ablation reported in detail in companion work [15]: Group C (dense + unstructured) and Group D (sparse + structured) close the diagonal of the original Density Imperative two-cell design on the in-domain CogBench protocol used in that prior work, and reveal that density and structure are *independently sufficient* on the in-domain measurement. The largest replicated effect in the in-domain 2×2 data is Group D vs Base on output information density ($\Delta = +0.297$, $p = 3.9 \times 10^{-18}$, both replication seeds; §9.3). The 953-prompt corpus, 8 LoRA adapters, 873 4-judge judgments, judge-prompt text, and reproducibility harness are released under CC BY 4.0.

Keywords: preference-aware objectives; verifier-judge divergence; alignment tax; anti-calibration; format compulsion; calibration compulsion; cross-judge evaluation; data-centric AI; ORPO; single-seed limitation.

1 Introduction

Supervision-quality claims for post-trained reasoning models tend to be stated as if a single evaluation axis were available. A training intervention is described as improving reasoning; a judge panel or a benchmark score is offered as evidence; the claim is absorbed. This paper reports a case in which one training intervention wins the deterministic verifier, a second gains judge-panel rank, and a third fails both — and the three interventions share a common corpus, base model, objective family, and hyperparameter set. The disagreement between deterministic and preference-based evaluation is structural. We describe what we observe as a **three-way supervision tradeoff**, scoped explicitly to this empirical setup: one corpus (459 records), one base family (Llama 3.2 11B Vision-Instruct), one objective family (ORPO with SFT baseline), single-seed main results pending 3-seed replication.

The finding arrived by negative route. We set out to validate the hypothesis that structural-signal preference-aware training would transfer reasoning out of distribution, and that conditioning a preference signal on a surface token (an appended confidence footer) would fail. The second prediction held. The first did not, in the direction predicted. On the 703-prompt deterministic

verifier subset, ORPO on structural-signal preference pairs with matched fp16 precision (Gen-1.5-continue-ORPO) scored 30.0% — below the unmodified base model's 32.0% and below naive-SFT on the same fp16 merged base (Gen-1-continue) at 34.6%. The ORPO arm that the paper's original hypothesis expected to win on both axes in fact *loses* on both: it is below Base on verifier (Wilson 95% CI on the −2.0pp difference crosses zero; single-seed) and below Base on the 4-judge Borda aggregate (0.400 vs 0.651, a robust separation). SFT is the one arm that buys anything: a +2.6pp verifier lift concentrated in HumanEval+ and MMLU-Pro, at the cost of dropping to Borda 5th of 8.

Surface-signal ORPO — where the chosen and rejected halves of each preference pair differ only on an appended `Confidence: {high|medium|low}` footer — fails both axes and additionally anti-calibrates. Worst verifier accuracy (27.6%), strongest unanimous-last consensus in the 4-judge panel (72/873 events), and Brier score 0.296 against an empirical-base-rate baseline of 0.204 on the 531 verifier-grounded emitted-footer prompts. Confidence supervision on a surface signal did not fail to teach calibration; it taught *anti-calibration*: the trained confidence scores are less informative than a uniform prior over the empirical base rate.

A single event from the 953-prompt sweep (§8) illustrates a *rubric-weighting disagreement* among frontier judges, not a judge-failure. On a MATH-500 Level 5 reflection problem, the student outputs being compared included Base (arithmetically incorrect axis but complete response) and Gen-1 (correct axis direction but inference-truncated before applying the reflection formula). Three of four frontier judges ranked Base first, weighting method-recognition plus completeness; one judge ranked Gen-1 first, weighting correctness-of-geometric-direction over completeness. Both weightings are defensible; the disagreement is the finding. We promote the specimen to the main text as an existence proof that frontier judges operationalize "quality" differently when completeness and correctness pull apart, not as evidence that the three-judge majority is wrong.

We frame the finding as an evaluation-methodology contribution as much as a training-methodology contribution: a supervision intervention's effect must be specified with respect to a metric, a benchmark stratum, and an evaluation axis, not described as generic capability movement.

The paper's contributions are five.

1. The **Supervision Tradeoff**: in this empirical setup, SFT uniquely exhibits a verifier-vs-judge tradeoff; both structural-signal ORPO arms are dominated by Base on both axes; surface-signal ORPO additionally anti-calibrates against an empirical-base-rate predictor. The single-axis generalization-improvement claim common in post-training literature is falsified on each of three regimes in different ways.

2. **Reasoning-NEST v2**, a provenance-rich failure-topology corpus (100-record Gen-0 pilot released under CC BY 4.0 on Hugging Face with Croissant metadata; Zenodo secondary archive).

3. A **pre-registered 953-prompt out-of-distribution confirmatory benchmark** judged by a 4-model cross-lineage panel (Anthropic / Google / OpenAI / xAI) with a 703-prompt deterministic verifier subset, including verbatim judge prompts, 873 4-judge judgments, and per-arm per-benchmark truncation-rate audit at 256 tokens.

4. An empirical documentation of **anti-calibration** under surface-signal ORPO: Brier 0.296 against a base-rate baseline of 0.204 on $n = 531$ emitted-footer verifier-grounded prompts, robust across three calibration-token-to-probability mappings.

5. A **two-axis refinement of the Density Imperative** [15]. *(a) Out-of-distribution scope*

refinement: Group A (sparse) and Group B (dense, 11.1× per-record density gap) are statistically indistinguishable on 953-prompt OOD verifier-grounded accuracy (32.9% vs 32.9%) and on 4-judge Borda aggregate (0.566 vs 0.575). Density's effect on OOD reasoning is not measurable on this corpus at this scale; the prior work's in-domain claim is not falsified, but its extrapolation to OOD is. *(b) In-domain* 2×2 *refinement*: companion work [15] closes the diagonal of the prior two-cell design with Group C (dense + unstructured) and Group D (sparse + structured), and finds density and structure to be *independently sufficient* on the in-domain CogBench measurement (full Wilcoxon contrast structure summarized in §9.3). The two refinements are complementary: in-domain, density and structure are independently sufficient; out-of-distribution at this corpus scale, neither density alone nor the dense+structured combination produces a measurable transfer effect on the 953-prompt sequestered benchmark. The elicitation hypothesis the in-domain pattern motivates—Group D as a regime that reorganizes pre-trained associations rather than teaching new ones—is identified and analyzed in the companion paper.

2 Related Work

Zhu et al. [23] show that rationale-style supervision is not uniformly beneficial; Kang et al. [11] demonstrate that fine-tuning examples directly shape hallucination behavior. Both motivate the present study's explicit bookkeeping of training signal versus downstream metric.

On supervision for reasoning: STaR [20] and Quiet-STaR [21] bootstrap rationales from answer supervision. Let's Verify Step by Step [13] moves to step-level process labels. SCoRe [12] trains self-correction online and explicitly reports that offline SFT on correction traces can fail due to distribution mismatch. Constitutional AI [2] samples and revises under principles.

The preference-aware objective used is ORPO [9]. Related objectives include DPO [19] and KTO [6]. The Tradeoff pattern reported here is not claimed ORPO-specific; whether DPO, KTO, or others exhibit the same verifier-judge divergence is an open question and future work.

On evaluation methodology, MIRAGE [5] and MIRAGE-Bench [22] treat hallucination as benchmark-design. Reasoning-NEST v2 differs as a post-failure *training* corpus. Documentation follows Datasheets [7] and Croissant [1]; a Croissant JSON-LD metadata file accompanies the Hugging Face release.

3 The Supervision Tradeoff

We state what we observe directly. Three supervision regimes on the same Reasoning-NEST v2 corpus and Llama 3.2 11B Vision-Instruct base family, evaluated on a 703-prompt deterministic verifier and a 4-judge cross-lineage Borda aggregate, produce three qualitatively distinct outcomes. Table 1 is the paper's central finding, compressed.

Table 1: Three supervision regimes × two evaluation axes, in this empirical setup. No fine-tuned arm improves both verifier-ALL (703-prompt deterministic pass rate) *and* Borda-4J (normalized 4-judge Borda aggregate on 873 shared events) above Base. **Important reading order:** the bolded Gen-1-continue "tradeoff" row presents the original single-seed observation; that row was *withdrawn* after a 3-seed confirmatory replication inverted the paired contrast (§3.3). The Base-dominates-Borda, surface-ORPO anti-calibration, and density-walk-back rows are separate measurements and survive replication. The table is retained in its original form so a reader can compare what was originally claimed against what survived.

Supervision regime	**Verifier-ALL**	**Borda-4J**	**Reading**
Base (no fine-tuning)	32.0%	**0.651**	**Reference; dominates Borda above every fine-tuned arm.**
Group A (sparse CN, SFT)	32.9%	0.566	Indistinguishable from Base on both axes.
Group B (dense CN, SFT)	32.9%	0.575	Indistinguishable from Group A despite 11.1× per-record density.
Gen-1 (NEST SFT, 4-bit base)	31.7%	0.510	Below Base on both.
Gen-1-continue (NEST SFT, fp16-matched base)	**34.6%**	0.486	**[WITHDRAWN post-replication — see §3.3.]** Original single-seed reading: only arm exhibiting a tradeoff (+2.6pp verifier at −0.165 Borda vs Base). 3-seed replication mean is 23.82% (intersection); paired SFT−ORPO contrast inverts sign.
Gen-1.5-ORPO (structural, 4-bit)	28.4%	0.419	*Quantization-confounded*; reported separately.
Gen-1.5-continue-ORPO (structural, fp16-matched)	30.0%	0.400	**Dominated on both axes. −2.0pp verifier, −0.25 Borda vs Base.**
Gen-1.5-Confidence-ORPO (surface, 4-bit)	**27.6%**	**0.394**	**Worst on both. Plus anti-calibrated (§6): Brier 0.296 vs base-rate 0.204.**

The honest structure of the finding:

1. **Base dominates the judge-panel Borda aggregate**. No fine-tuned arm in this study recovers to Base on the 4-judge aggregate. The gap from Base (0.651) to the next arm (Group B at 0.575) is 0.076 normalized Borda points; the gap from Base to the top structural ORPO arm (0.419) is 0.232. This is the study's most robust quantitative finding and it is inconvenient for the paper's original hypothesis. We address it explicitly in §9.

2. **SFT on structural-signal preference-pair targets uniquely exhibits a verifier-vs-judge tradeoff**. Under fp16-matched precision, Gen-1-continue scores 34.6% on the 703-prompt verifier (95% Wilson CI [31.1%, 38.2%]) against Base's 32.0% (+2.6pp, 95% CI on difference $[-2.36, +7.49]$, crosses zero at single seed). On the 4-judge Borda, Gen-1-continue falls to 0.486 (Borda-5th of 8). The verifier gain is also benchmark-selective (§3.1): HumanEval+ +7.0pp, MMLU-Pro +2.8pp, MATH-500 −1.0pp. SFT *loses* MATH-500. We call this a "supervision tradeoff" because it is a genuine bi-directional cost structure: SFT buys one axis at the price of another.

3. **ORPO on structural-signal preference pairs is dominated on both axes** in this empirical setup. Gen-1.5-continue-ORPO scores 30.0% verifier (−2.0pp below Base, 95% CI $[-6.83, +2.84]$,

crosses zero at single seed) and Borda 0.400 (−0.251 below Base, −0.086 below even SFT). There is no cell of the 3×2 matrix in which structural-signal ORPO wins anything. On the per-benchmark verifier breakdown, Gen-1.5-continue-ORPO is −7.6pp on HumanEval+ (38.9% vs Base's 46.5%) — the largest single-benchmark regression among fine-tuned arms. We do not call this a tradeoff; it is a two-axis loss that the paper's original hypothesis did not predict.

4. **Surface-signal ORPO fails both axes and additionally anti-calibrates** (§6). Worst verifier (27.6%), strict-last Borda (0.394), 72 unanimous-last events of 873, and Brier 0.296 against an empirical-base-rate baseline of 0.204 on $n = 531$ emitted-footer prompts. The calibration assertions the trained model makes are *worse than uninformed*.

3.1 Benchmark-selective verifier gain under SFT

The +2.6pp Gen-1-continue verifier win is not a uniform lift. Table 2 decomposes it.

Table 2: Per-benchmark verifier-grounded pass rates, 703-prompt subset. SFT's verifier win is concentrated in HumanEval+ and MMLU-Pro. SFT *loses* MATH-500. Structural-signal ORPO (fp16-matched) regresses by −7.6pp on HumanEval+ specifically.

Arm	ALL	MATH-500	HumanEval+	MMLU-Pro
Base	32.0%	11.0%	46.5%	38.6%
Group A (sparse)	32.9%	11.5%	47.9%	39.4%
Group B (dense)	32.9%	10.5%	47.2%	40.3%
Gen-1-continue (SFT fp16)	**34.6%**	**10.0%** (−1.0)	**53.5%** (+7.0)	**41.4%** (+2.8)
Gen-1.5-ORPO (4-bit conf.)	*28.4%*	*7.2%* (−3.8)	*33.3%* (−13.2)	*39.1%* (+0.5)
Gen-1.5-continue-ORPO (fp16)	30.0%	11.0% (–)	**38.9%** (−7.6)	37.7% (−0.9)
Gen-1.5-Confidence-ORPO (surface)	27.6%	7.2% (−3.8)	31.2% (−15.3)	38.3% (−0.3)

The "SFT wins verifier" claim, stated without benchmark decomposition, overstates. SFT wins HumanEval+ clearly and MMLU-Pro modestly. It loses MATH-500 by −1pp. Structural-signal ORPO regresses HumanEval+ by −7.6pp (fp16) or −13.2pp (4-bit), driving most of the ORPO verifier deficit. We do not have a mechanistic hypothesis ready for the HumanEval+ specificity; verifier-grounded step-level trace analysis is planned as future work.

3.2 What this is not

The *Supervision Tradeoff* language is scoped to what SFT does. Neither structural-signal ORPO nor surface-signal ORPO exhibits a tradeoff in this study — both are strictly dominated by Base on both measured axes, and surface-signal ORPO is additionally anti-calibrated. The label fits the SFT regime; it does not fit ORPO. We retain the paper title because SFT's tradeoff is the regime the paper investigates most carefully, and because describing ORPO-as-dominated honestly does not require a different title.

3.3 Three-seed replication: the fp16-matched verifier contrast does not reproduce

The paper's linchpin fp16-matched contrast — Gen-1-continue SFT 34.6% vs Gen-1.5-continue-ORPO 30.0%, a +4.6pp SFT advantage on the 703-prompt verifier at single seed — was submitted for

a pre-registered 3-seed replication (seeds $\{42, 1337, 2026\}$) on the same fp16 merged base, matched hyperparameters, and the identical verifier-grounded v2 harness (`supplementary/round2_seed_replication.sh`). **The replication does not reproduce the single-seed result and inverts the sign of the paired contrast.**

On the intersection-restricted subset of $N = 555$ 953-prompt events present in all six inference runs ($N = 361$ verifier-grounded within that intersection, identical prompt set across all arms):

Table 3: 3-seed replication of the fp16-matched pair, restricted to the $N = 361$ verifier-grounded events present in all six inference runs. The paired mean difference inverts sign versus the single-seed reading in Table 1; the 95% t-CI narrowly crosses zero.

Arm (seed)	ALL%	MATH	HEval+	MMLU-Pro	Reading
Gen-1-continue s42	24.10%	4.0%	18.1%	35.3%	
Gen-1-continue s1337	23.27%	12.0%	18.1%	31.1%	
Gen-1-continue s2026	24.10%	6.0%	19.4%	33.5%	
SFT mean (std)	**23.82% (0.48)**	7.3%	18.5%	33.3%	**Single-seed 34.6% not reproduced.**
Gen-1.5-continue-ORPO s42	27.70%	8.0%	29.2%	32.3%	
Gen-1.5-continue-ORPO s1337	26.04%	8.0%	31.9%	26.3%	
Gen-1.5-continue-ORPO s2026	30.47%	8.0%	34.0%	34.1%	
ORPO mean (std)	**28.07% (2.24)**	8.0%	31.7%	30.9%	
Paired diff (SFT − ORPO)	**−4.25pp**	**95% paired t-CI on 3 seeds:** $[-8.93, +0.44]$**, crosses zero.**			

The 95% confidence interval for the paired difference is a paired t-CI computed from the std of the seed-matched paired differences ($s_d = 1.89$pp over $n = 3$ seeds, $t_{0.975,2} = 4.303$). An independent-samples pooled-variance calculation using the per-arm within-row stds in the table above would yield a wider interval and would not be appropriate for a seed-matched design where the same seeds are used across both arms.

Two findings, both of them methodological. (i) The single-seed Gen-1-continue accuracy of 34.6% is not recovered across three replication seeds (new range $[20.9\%, 23.3\%]$, mean 22.19% on the full-953 sweep, 23.82% on the verifier-grounded intersection). The gap between the original single-seed claim and the 3-seed replication mean is approximately 11pp on the same objective, base, and hyperparameters — larger than the 4.6pp within-contrast effect the paper originally treated as signal. (ii) The paired contrast SFT−ORPO reverses direction: on the intersection subset the mean difference is −4.25pp (in ORPO's favor), 95% t-CI on three seeds $[-8.93, +0.44]$; on the full-953 sweep with uneven per-arm record counts the mean difference is −4.49pp, 95% t-CI $[-10.58, +1.59]$. The replicated arms are statistically indistinguishable with a slight non-significant lead for ORPO.

What we withdraw. The Verifier-ALL cell and Reading column for **Gen-1-continue (fp16-matched)** in Table 1 cannot be elevated to a robust finding. The claim "SFT uniquely exhibits a verifier-vs-judge tradeoff" does not hold against the deterministic-verifier dimension in this training setup; the benchmark-selective decomposition (Table 2) is reported as descriptive of the original single-seed run and not as a replicated lift.

What survives. The **Base-dominates-Borda** finding is a separate measurement on the 4-judge panel against the original single-seed adapters; it is not within the scope of the verifier-grounded

replication, and Borda gaps (Base 0.651 vs top fine-tuned 0.575) are roughly an order of magnitude wider than the pairwise noise the replication exposes on verifier. We retain Base-dominates-Borda as the paper's strongest single-axis finding. The **surface-signal ORPO anti-calibration** finding is a separate measurement on confidence-footer emission and is likewise preserved. The **Group A / Group B density walk-back** is also unaffected.

The new primary finding. The most empirically durable conclusion from the v5 replication is the training-regime instability itself: post-training verifier accuracy on this base exhibits an $\sim$ 11pp seed-to-seed gap on the SFT arm, larger than any inter-regime difference the paper originally took to be informative. This is a publishable conclusion in its own right: single-seed post-training numbers on Llama 3.2 11B Vision-Instruct at this corpus scale are not statistically informative, and the field's convention of reporting them as if they were should be revisited. The replication script, seed adapters, and verifier-grounded intersection JSON are released as supplementary material (`supplementary/round2_seed_replication.sh`, `supplementary/seed_replication/`).

4 Reasoning-NEST v2 Corpus

4.1 NEST canonically, and the Reasoning-NEST adaptation

NEST is the *Neural Extraction of Semantic Topology*, a schema for converting source artifacts into graph-structured semantic scaffolding bound immutably to provenance metadata. **Reasoning-NEST v2** is a text-only adaptation for failure-topology supervision. Records average $\sim$1,100 tokens each (median 1,122; p95 1,286).

4.2 Corpus composition and release

Reasoning-NEST v2 is a structured JSON corpus built from failed model outputs converted into corrective artifacts through a three-teacher consensus pipeline. Public pilot: 100 admitted records under CC BY 4.0 [17]. Full study uses 459 admitted records. Primary release: Hugging Face Datasets (`metavolve-labs/reasoning-nest-v2`) with Croissant metadata; Zenodo [17] is the secondary archive.

4.3 Schema

Table 4: Core Reasoning-NEST v2 record schema.

Component	Representative fields	Function
Event	`event.prompt`, `event.failed_output`, `event.correction`	Task, failure, verified correction
Reasoning	`reasoning.cot_trace[]`, `reasoning.failure_narrative`, `reasoning.failure_mechanism`	Step validity, narrative, named mechanism
Causality	`causality.graph`, `causality.adjustment_formula`	Causal assumptions explicit [18]
Topology	`topology.error_class`, `topology.avoidance_rule`, `topology.complexity_score`	Reusable error type and preventive rule
Provenance	`provenance.oracle_fingerprint`, `provenance.trust.gradient`, `provenance.densified_at`, `provenance.source_event_hash`	Audit, inspection, filtering

4.4 Three-Teacher Consensus Pipeline and Taxonomy

Primary oracle: Gemini 3.1 Pro Preview via Google Vertex AI. Secondary oracle: GPT-4o via OpenAI API. Judge/arbitrator: Claude Sonnet 4.6 via Anthropic API. Only admitted records enter the corpus.

The `topology.error_class` field enumerates seven classes: `compiler_syntax`, `runtime_semantic_code`, `math_logic_fallacy`, `factual_hallucination`, `multi_step_planning`, `visual_perceptual`, `cultural_context` Each admitted record also carries a declarative `topology.avoidance_rule` and a scalar `topology.complexity_sco` in $[0, 1]$.

5 Methodology

5.1 Continuation Protocol

Llama 3.2 11B Vision-Instruct via the Unsloth 4-bit quantized mirror. Path A (adopted): fresh LoRA with vision layers frozen. LoRA [10] / QLoRA [4].

5.2 Training Target Format

SFT target (Gen-1, Gen-1-continue) is a single fused natural-language sequence:

```
User: {event.prompt}
Assistant:
Reasoning: {concatenated valid cot_trace steps only}

Avoidance rule: {topology.avoidance_rule}

Answer: {event.correction}
```

This is a *single-target* SFT regime; there is no preference pair structure presented to the optimizer. ORPO training (Gen-1.5-ORPO, Gen-1.5-continue-ORPO, Gen-1.5-Confidence-ORPO) constructs preference pairs {prompt, chosen: fused target above, rejected: `event.failed_output`} and trains with the ORPO [9] monolithic odds-ratio objective at $\beta = 0.1$. Gen-1.5-Confidence-ORPO appends a `Confidence: {high|medium|low}` footer derived from `topology.complexity_score` — complexity-matched on the chosen stream and mismatched on the rejected stream.

5.3 Training Configuration, Implementation, and Precision Asymmetry

SFT on L4 (22 GB VRAM): 554 s, loss 1.3812, ∼$5. ORPO exceeded L4's VRAM and ran on preemptible A100 40 GB in ∼30 minutes total for under $4 combined.

Table 5: Training configuration.

Adapter	LoRA rank 32, alpha 16, dropout 0.0
Trainable scope	Language layers only; `q, k, v, o, gate, up, down_proj`
Vision	`finetune_vision_layers=False`
Optimization	LR 2e-4; cosine; warmup 0.03; weight decay 0.01
Precision	`bf16`; `adamw_8bit`; `use_gradient_checkpointing="unsloth"`
Batching	Per-device batch 1; grad accum 8; effective batch 8
Schedule	3 epochs over 459 records; 171 `max_steps`
ORPO β	0.1

Quantization-precision asymmetry. Gen-1.5-ORPO was trained on a 4-bit quantized base (QLoRA); Gen-1-continue and Gen-1.5-continue-ORPO used the unquantized 16-bit merged base. We hypothesize that 4-bit quantization may contribute to numerical instability on structured-logic tasks but have not empirically isolated this effect in this study. The Gen-1-continue (fp16) vs Gen-1.5-ORPO (4-bit) direct comparison is therefore confounded between objective and precision. The cleaner fp16-matched contrast is **Gen-1-continue (34.6% verifier) vs Gen-1.5-continue-ORPO (30.0% verifier)**; all headline Tradeoff claims are stated against this pair.

5.4 Evaluation Protocol, Confidence Intervals, and Known Limitations

Inference: temperature 0.7, `max_new_tokens=256`. The 256-token ceiling is a known limitation (§11); at this ceiling, MATH-500 Level 5 outputs hit the cap at 25.7% (Base), 48.6% (Group B), 25.7% (Gen-1-continue SFT), and only 5.5–6.4% (ORPO arms). A 200-prompt stratified re-run at 1024 tokens is in progress (§12).

Phase-2 confirmatory evaluation: 953-prompt eight-arm ablation (§7) judged by four independent models.

Confidence intervals. Wilson 95% intervals on all proportions; normal-approximation 95% intervals on proportion differences. Every pairwise verifier difference in this study crosses zero at single-seed N=703. The directional pattern is internally consistent; the specific numeric gaps are not robust to seed variance at current N.

Single-seed limitation. Main results in this paper are single-run per arm. A 3-seed confirmatory replication (seeds 42, 1337, 2026) on the fp16-matched pair is in progress; all claim language is scoped conditional on that replication confirming direction.

Multiple-comparison correction. ∼160 comparisons; Bonferroni-corrected $\alpha \approx 3.1 \times 10^{-4}$. Surviving at Bonferroni: Gen-1.5-Confidence-ORPO strict-last under 4-judge Borda; Fleiss' $\kappa > 0$ across the panel; Brier-vs-base-rate anti-calibration. Not surviving at Bonferroni: the single-seed pairwise verifier differences, which are reported as directional pending seed-variance replication.

5.5 Endpoints and Scoring Rubric

Binary: major-hallucination rate, minor-hallucination rate, clean-output rate. Ordinal 0–5: factual accuracy, reasoning validity, structural quality, epistemic hedging. Calibration: Brier score and Expected Calibration Error [3, 8]; base-rate baseline detailed in §6. **Verbatim judge prompts released in App. B.**

5.6 Methodological Infrastructure and Evaluation Panel

All scientific claims and authorship accountability rest with the authors. AI systems used as measurement infrastructure:

- **Gemini 3.1 Pro Preview (Google Vertex AI)**: primary oracle in the three-teacher densification pipeline; one of four judges on the 953-prompt sweep.
- **GPT-5.4 (OpenAI)**: editorial review, methodology rigor checks; one of four judges on the 953-prompt sweep.
- **Claude Sonnet 4.6 (Anthropic)**: arbitrator at corpus-construction; one of four judges on the 953-prompt sweep.
- **Grok-4-1 fast-reasoning (xAI)**: fourth-lineage judge on the 953-prompt sweep (953/953 coverage).

6 Anti-Calibration: Surface-Signal ORPO Fails Both Axes

Surface-signal ORPO produces a model that emits the footer task-conditionally and scores worse on Brier than an uninformed predictor. This is a standalone finding distinct from the Tradeoff structure of §3.

6.1 Baseline: the Empirical Base-Rate Predictor

For binary-outcome Brier on a verifier-grounded subset with empirical pass rate $p = 0.284$, the principled naive baseline is the constant-p-base-rate predictor with

$$\text{Brier}_{\text{base-rate}} = p(1-p) = 0.284 \times 0.716 = \mathbf{0.204}.$$

The Gen-1.5-Confidence-ORPO arm's Brier across 531 footer-emitting verifier-grounded prompts is **0.296**, exceeding the base-rate baseline by +0.092 absolute — **the trained calibrator is 45% worse than predicting the empirical base rate**. The constant-p=0.5 baseline at 0.250 is also exceeded.

6.2 Per-Bucket Contingency (with 95% Wilson CIs)

Table 6: Footer-correctness contingency with Wilson 95% CIs.

Footer value	N	**Correct**	**P(correct)**	**Wilson 95% CI**
high	97	26	26.8%	[19.0%, 36.4%]
medium	419	125	29.8%	[25.7%, 34.4%]
low	15	0	0.0%	[0.0%, 20.4%]
(no footer)	172	43	25.0%	[19.1%, 32.0%]
All verifier-grounded	**703**	**194**	**27.6%**	[24.4%, 31.0%]

The per-bucket P(correct | high) vs P(correct | medium) difference is −3.03 percentage points with 95% CI [−12.87, +6.81], crossing zero. We acknowledge explicitly that the bucket-level discrimination claim is underpowered at $n = 97$ vs $n = 419$. The anti-calibration claim rests on the Brier-over-531 comparison, not on per-bucket discrimination.

6.3 Visceral Framing of the Overconfidence Gap

Under the conventional probability mapping high=0.85, the 97 prompts on which the model asserted "high" confidence had empirical accuracy of 26.8%. The model's asserted confidence under this mapping overstated true accuracy by 58 percentage points. The framing is mapping-dependent: under high=0.30, the gap would be +3pp not +58pp. We report 58pp as a rhetorical summary under the standard mapping; the mapping-robustness claim (Table 7) is what survives review.

6.4 Brier Mapping Sensitivity

Table 7: Anti-calibration is robust to calibration-token-to-probability mapping.

Mapping (high / medium / low)	**Brier**	**ECE**	**Skill score vs base-rate**
0.85/0.50/0.15	0.296	0.270	**−0.451**
0.75/0.50/0.25	0.277	0.254	−0.358
0.90/0.60/0.20	0.312	0.281	−0.529
Base-rate baseline $p(1-p), p = 0.284$	**0.204**	—	0.000
Constant-p=0.5 baseline	**0.250**	—	−0.225

Skill score is $1 - \text{Brier}_{\text{model}}/\text{Brier}_{\text{base-rate}}$. A negative skill score indicates anti-calibration: the trained confidence assertions carry strictly less information than a uniform constant-p predictor calibrated to the empirical base rate. Under the standard mapping the skill score is −0.451, i.e., the trained calibrator is 45.1% worse than predicting the base rate; the conventional constant-$p = 0.5$ baseline is also exceeded under all three mappings.

6.5 Emission is Task-Conditional, Not Knowledge-Conditional

Table 8: Footer emission rate on the 953-prompt confirmatory OOD sweep.

Source	N	Emitted	Rate	high / medium / low
MMLU-Pro (psychology)	66	66	**100.0%**	5 / 61 / 0
MMLU-Pro (biology)	71	68	95.8%	7 / 59 / 2
MMLU-Pro (business)	71	68	95.8%	9 / 59 / 0
MMLU-Pro (law)	71	67	94.4%	0 / 67 / 0
MMLU-Pro (chemistry)	71	65	91.5%	10 / 54 / 1
Hand-curated-OOD-carryforward	30	25	83.3%	1 / 21 / 3
MATH-500-L4	70	54	77.1%	21 / 31 / 2
MATH-500-L5	109	84	77.1%	21 / 55 / 8
Hand-curated-OOD-new	120	90	75.0%	9 / 75 / 6
MATH-500-L3	30	22	73.3%	10 / 10 / 2
HumanEval+	144	37	25.7%	14 / 23 / 0
BigCodeBench	100	4	**4.0%**	2 / 2 / 0
Total	**953**	**650**	**68.2%**	**109 / 517 / 24**

6.6 What this does not claim

The anti-calibration finding is narrow: calibration is untrainable under ORPO via a surface-token preference signal on this corpus at this scale. A structural calibration signal — where the chosen/rejected distinction corresponds to calibrated-abstention vs over-confident-hallucination with auditable provenance — is a separate experiment and is not refuted here.

7 Confirmatory OOD Evaluation (953-prompt sweep)

7.1 Corpus and Protocol

The 953-prompt sequestered corpus: MATH-500 L3/L4/L5 ($n = 30/70/109$), HumanEval+ ($n = 144$), BigCodeBench ($n = 100$), MMLU-Pro ($n = 350$), hand-curated OOD ($n = 150$). Every prompt MD5-hashed against the 459 training records; zero overlap required. All eight arms re-run at temperature 0.7, `max_new_tokens=256`.

Table 9: Judge-panel coverage.

Judge	Lineage	Coverage	%
Claude (`claude-sonnet-4-6`)	Anthropic	949	99.6%
Gemini (`gemini-3.1-pro-preview`)	Google	876	91.9%
GPT (`gpt-5.4`)	OpenAI	953	100.0%
Grok (`grok-4-1-fast-reasoning`)	xAI	953	100.0%

4-judge intersection: 873 events. Gemini's 77 uncovered prompts cluster in hard verifier-grounded content: MATH-500-L5 81.7% covered, HumanEval+ 88.2%, BigCodeBench 85.0%, MMLU-chemistry 84.5%. Non-verifier-grounded content: $\geq$98.5% covered. Missingness rate: 9.3% on verifier-grounded vs 1.3% on non-grounded, driven by rate-limit retries timing out at the 5-attempt cap.

7.2 Verifier-Grounded Accuracy with Wilson CIs

Full per-benchmark breakdown is in Table 2 (§3.1). The headline pattern:

Table 10: Verifier-grounded pass rates on 703 prompts with Wilson 95% intervals. Italic = quantization-confounded.

Arm	Precision	ALL	Wilson 95% CI	N
Base	fp16	32.0%	[28.7%, 35.5%]	703
Group A (sparse)	fp16	32.9%	[29.5%, 36.4%]	703
Group B (dense)	fp16	32.9%	[29.5%, 36.4%]	703
Gen-1 (SFT)	4-bit	31.7%	[28.4%, 35.3%]	703
Gen-1-continue (SFT, merged)	**fp16**	**34.6%**	[31.1%, 38.2%]	703
Gen-1.5-ORPO (structural)	*4-bit*	*28.4%*	[25.2%, 31.9%]	703
Gen-1.5-continue-ORPO (structural)	**fp16**	30.0%	[26.7%, 33.5%]	703
Gen-1.5-Confidence-ORPO (surface)	4-bit	27.6%	[24.4%, 31.0%]	703

Table 11: Pairwise difference 95% CIs on load-bearing comparisons. All five cross zero at single-seed N=703.

Comparison	Δpp	CI-lo	CI-hi
Gen-1-continue (fp16) vs Base	+2.56	−2.36	+7.49
Gen-1-continue (fp16) vs Gen-1.5-continue-ORPO (fp16, FP16-MATCHED)	+4.55	−0.33	+9.43
Base vs Gen-1.5-ORPO (4-bit confounded)	+3.56	−1.24	+8.35
Base vs Gen-1.5-continue-ORPO (fp16)	+1.99	−2.84	+6.83
Base vs Gen-1.5-Confidence-ORPO (surface, 4-bit)	+4.41	−0.37	+9.19

The directional pattern is internally consistent (SFT > Base > structural ORPO > surface ORPO on verifier) and its ordering is the opposite of the judge-panel Borda ordering on the same arms (Base dominates; ORPO bottom tier). Single-seed caveat applies.

7.3 Borda Aggregate and Polarization Counts

Table 12: Borda aggregate (873 shared events) and unanimous-last count, 4-judge panel.

Arm	Borda (0–1)	Tiebreaker 1st	Tiebreaker last	Unanimous-last
Base	**0.651**	**233**	36	5
Group B	0.575	119	58	18
Group A	0.566	90	51	11
Gen-1 (SFT, 4-bit)	0.510	94	58	18
Gen-1-continue (SFT, fp16)	0.486	59	87	13
Gen-1.5-ORPO (4-bit)	0.419	84	145	42
Gen-1.5-continue-ORPO (fp16)	0.400	62	186	47
Gen-1.5-Confidence-ORPO (surface)	**0.394**	132	**252**	**72**

Ranking preserved under three-judge sensitivity analysis with Gemini excluded (Spearman $\rho = 0.93$ to the 4-judge aggregate).

7.4 Judge-Agreement Structure in the Four-Lab Panel

Fleiss' κ on 873 shared events: first-place arm identity $\kappa = 0.394$, last-place arm identity $\kappa = 0.413$.

Pairwise cross-judge agreement metrics on factual-accuracy arm ranks (rank-based Spearman ρ, methodologically consistent with categorical Cohen κ on first-place arm identity) and on first-place arm identity (Cohen κ):

Table 13: Pairwise cross-judge agreement on 873 shared events. Spearman ρ is rank correlation across arms on continuous factual-accuracy means; Cohen κ is categorical agreement on first-place arm identity. Both are legitimate agreement metrics on their respective types; the honest observation is that rank agreement on continuous scoring is consistently high while categorical agreement on winner identity is moderate.

Pair	**Spearman ρ (factual rank)**	**κ 1st-place**	**κ last-place**
Gemini ↔ GPT	0.929	0.412	0.406
Gemini ↔ Grok	0.881	**0.448**	**0.453**
GPT ↔ Grok	0.905	0.425	0.404
Claude ↔ Gemini	**0.905**	**0.308**	0.397
Claude ↔ Grok	0.833	0.378	0.400
Claude ↔ GPT	0.786	0.396	0.426

All six Spearman ρ pairs on continuous factual-accuracy arm ranks are ≥ 0.79. All six Cohen κ pairs on categorical first-place arm identity are in $[0.31, 0.45]$. The judges converge when asked to score facts on a continuous scale and diverge when asked to pick a single winner. The converging continuous dimension does not resolve the diverging categorical judgment.

8 Meta-Evaluation: A Judge-Panel Rubric-Weighting Specimen

A single event from the 953-prompt confirmatory sweep illustrates, on problem mathematics a reviewer can verify independently, a *rubric-weighting disagreement* among the four frontier judges. We report this as an existence proof at $N = 1$ of the phenomenon that frontier judges operationalize "quality" differently when completeness and correctness pull apart. We do not claim the three-judge majority was wrong; their rubric weighting was legitimate, just different.

8.1 The Event

- **Record ID**: `a6218ea295a817e5e50888f28b67989f`
- **Source**: MATH-500 Level 5 (Precalculus)
- **Prompt**: "A reflection takes $\begin{pmatrix} 5 \\ 0 \end{pmatrix}$ to $\begin{pmatrix} 4 \\ 3 \end{pmatrix}$. Which vector does the reflection take $\begin{pmatrix} -2 \\ 3 \end{pmatrix}$ to?"
- **Ground truth**: $\begin{pmatrix} 1/5 \\ -18/5 \end{pmatrix}$

8.2 The two candidate outputs (as they appeared to judges)

Base — axis: $y = \frac{1}{3}x - 1$, full response, arithmetically wrong intercept: *"The line of reflection is the perpendicular bisector of the segment joining* $(5, 0)$ *and* $(4, 3)$*. Midpoint* $(4.5, 1.5)$*, slope* -3*, perpendicular-bisector slope* $\frac{1}{3}$*. The equation is* $y = \frac{1}{3}x - 1$*.* [Continues, response completes within the 256-token ceiling.]*"*

Gen-1 — axis direction: origin-through with slope $\frac{1}{3}$, inference-truncated: *"The vector being reflected over is a scalar multiple of the midpoint vector. In this case, the vector being reflected over is* $(9/2, 3/2)$*. We can now use the formula for a reflection of a point* (x, y) *over the vector* (a, b)*:* [TRUNCATED at 256-token inference ceiling before formula is applied]*"*

Gen-1's output was cut off at the 256-token inference limit before the reflection formula was applied to $(-2, 3)$. The truncation is an evaluation-protocol artifact (§5.4); at 1024 tokens the same generation would likely have completed. Gen-1 *had* the correct geometric axis direction. Gen-1 did *not* produce a completed numerical answer.

8.3 The mathematical observation

Reflections in $\mathbb{R}^2$ are Euclidean isometries and preserve norm. Both $(5, 0)$ and $(4, 3)$ have norm 5, so the reflection axis must pass through the origin. Base's axis $y = \frac{1}{3}x - 1$ does not pass through the origin (at $x = 0$, $y = -1$) — it fails the origin-fixity check. Gen-1's axis direction $(9/2, 3/2)$ corresponds to the line $y = \frac{1}{3}x$ through the origin, which is the correct axis. Base's arithmetic error is a one-step slip: $y - 1.5 = \frac{1}{3}(x - 4.5)$ simplifies to $y = \frac{1}{3}x$, not $y = \frac{1}{3}x - 1$.

Each candidate has one strength and one weakness. Base: wrong axis, complete response. Gen-1: correct axis direction, truncated before applying it.

8.4 Judge Top Picks and Rationales (verbatim)

Claude (`claude-sonnet-4-6`) — top pick `base`: *"Base correctly identifies the perpendicular bisector approach and gets the midpoint and slope right (though truncated)."*

GPT (`gpt-5.4`) — top pick `base`: *"Base is best because it correctly identifies the perpendicular-bisector reflection axis before truncation."*

Grok (`grok-4-1-fast-reasoning`) — top pick `base`: *"Base leads with superior factual accuracy and reasoning on the reflection line."*

Gemini (`gemini-3.1-pro-preview`) — top pick `gen1`: *"Gen 1 and Group A correctly identify the direction of the line of reflection before cutting off; Base sets up the perpendicular bisector correctly but makes an arithmetic error."*

8.5 Interpretation: rubric-weighting disagreement, not judge failure

Three of four frontier judges preferred Base. Reading the rationales verbatim: each of those three judges weighted *method recognition* (perpendicular-bisector approach correctly named) and *response completeness* (response reached a final numerical answer, even if arithmetically wrong) over *correctness of geometric direction*. Gemini, the fourth judge, weighted *correctness of geometric direction* over *completeness*, and its rationale explicitly flags Base's arithmetic error.

These are different but defensible rubric-weighting choices. Under completeness-weighting, Base is a better answer because it is a full response that reaches a numerical conclusion. Under direction-

correctness weighting, Gen-1 is a better answer because it identifies the axis that any complete solution would need to pass through the origin. Both weightings are legitimate for a quality-of-reasoning evaluation; they just emphasize different properties.

We originally framed this specimen as evidence that three judges "failed to verify the geometry." That framing overclaims. The three judges who preferred Base were not failing at geometry; they were applying a legitimate rubric weighting in which a complete response with a named method dominates a truncated response with a correct direction. The existence of the disagreement is the finding. Frontier LLM judges do not share a uniform latent rubric for "quality of reasoning" when completeness and correctness are in tension on the same prompt.

The evaluation-methodology implication is that contributions depending on 4-judge Borda aggregation have an unacknowledged sensitivity to completeness-weighting that may or may not align with the research question a paper is trying to ask. Authors running judge-panel evaluations on reasoning benchmarks should either (a) release verbatim judge rationales with their aggregate statistics, (b) stratify by completeness-truncation status, or (c) use deterministic verifiers where available and treat judge-panel Borda as an orthogonal stylistic-preference measurement rather than a correctness proxy. The present paper does (a) and (b) and we recommend the practice.

9 Discussion

9.1 Base is the best untrained reference on judges, above every fine-tuned arm

The most robust finding in the paper is also the most inconvenient for its original hypothesis: the untrained Llama 3.2 11B Vision-Instruct base model achieves a 4-judge Borda aggregate of 0.651, and no fine-tuned arm in this study recovers to Base's Borda. The gap to the next-highest arm (Group B at 0.575) is 0.076 normalized Borda points; the gap to the top structural-signal ORPO arm (0.419) is 0.232. Base also receives 233 tiebreaker-first-place votes across the 4-judge panel, more than twice any other arm. This is cross-judge robust: preserved under three-judge sensitivity with Gemini excluded (Spearman $\rho = 0.93$ to the 4-judge aggregate).

The honest reading is that **at this corpus scale (459 training records) and this LoRA configuration (rank 32, 171 max_steps, 3 epochs), no training intervention in this study produces quality improvements recognizable by this 4-judge panel**. Neither dense in-domain supervision (Group B), nor structurally-rich failure-topology SFT (Gen-1, Gen-1-continue), nor structural-signal ORPO (Gen-1.5-ORPO, Gen-1.5-continue-ORPO), nor surface-signal ORPO (Gen-1.5-Confidence-ORPO) produces judge-panel gains above the untrained base.

This is not a refutation of post-training methods in general; it is a statement about this corpus and this scale. What the paper does establish, within those bounds:

- SFT on structural-signal single-target fused-format supervision uniquely buys a verifier gain at a judge-panel cost (§3).
- ORPO on structural-signal preference pairs is strictly dominated by Base on both axes; no single cell of the 3×2 matrix shows ORPO winning (§3).
- ORPO on surface-signal preference pairs additionally anti-calibrates against an empirical-base-rate predictor (§6).
- Dense in-domain corpora do not transfer to OOD reasoning at this scale (§9.2 below).

9.2 Density walk-back: the prior thesis's scope

This paper refines, not refutes, a claim from the first author's prior work. The Density Imperative [15] argued that dense, provenance-tracked semantic supervision (Cognitive Nutrition) was load-bearing for vision-language model reliability, supported by Group B's in-domain art-historical evaluation in the prior work. The present paper's 953-prompt OOD data corrects the extrapolation.

On 953-prompt sequestered OOD: Group A (sparse captions, ∼54 assistant tokens per record) and Group B (dense NEST, ∼603 tokens per record, 11.1× per-record density) are statistically indistinguishable on all measured axes:

- Verifier-grounded accuracy: 32.9% (Group A) vs 32.9% (Group B) — identical.
- Borda aggregate: 0.566 (Group A) vs 0.575 (Group B) — delta of 0.009 normalized Borda points.
- Per-benchmark verifier: MATH-500 11.5% vs 10.5%, HumanEval+ 47.9% vs 47.2%, MMLU-Pro 39.4% vs 40.3% — sub-percentage-point differences in both directions.

The 11.1× density advantage that drove in-domain improvement in the prior work produces zero measurable OOD benefit here. We take this as a **scope refinement, not a refutation**. Dense in-domain supervision transfers domain knowledge; at this scale on this corpus, it does not transfer to OOD reasoning. The prior work and the present work are compatible once the evaluation axis (in-domain vs OOD) is held fixed. An honest restatement: *density's effect on OOD reasoning is not measurable on this corpus at this scale.* If density transfers OOD at larger scale (70B-class bases, 10k+-record corpora), the present experiment cannot detect it.

9.3 In-domain 2×2: density and structure are independently sufficient

The OOD scope refinement of §9.2 concerns the original Density Imperative two-cell comparison (Group A sparse vs Group B dense). A complementary in-domain refinement closes the diagonal of that two-cell design. On the same 100-image held-out CogBench evaluation protocol used in [15], we add Group C (dense, unstructured: long-form prose paragraphs of equivalent token length to Group B but without the 111-field NEST schema) and Group D (sparse, structured: NEST schema preserved at ∼50–80 tokens per record). **Both new arms are trained via supervised fine-tuning** (not ORPO, not preference-aware) on otherwise identical Llama 3.2 11B Vision-Instruct base, LoRA rank 32, learning rate 2×10^{-4}, 1,135 steps, two replication seeds (s42, s1337). The training and evaluation framework here is therefore distinct from the Reasoning-NEST v2 corpus, ORPO objective family, and 953-prompt OOD benchmark that drive this paper's primary contributions; we import the in-domain 2×2 result because it is directly load-bearing for Contribution 5's two-axis refinement. Full methodology, per-condition training stats, and the elicitation-hypothesis analysis are reported in the companion paper [15].

Table 14: In-domain 2×2 ablation on the 100-image held-out CogBench protocol (companion work [15]). C and D are seed-paired averages over s42 and s1337. The Group D CCR cell exhibits 12pp seed-pair fragility on this metric (s42 0.91, s1337 0.79); see [15] §6 Limitation 4. Wilcoxon contrasts in Table 15 are computed against the s42 adapter; per-seed contrasts and seed-stability tests are reported in the companion paper.

Metric	**Base**	**A**	**B**	**C (avg)**	**D (avg)**
CogBench Overall	0.388	0.176	0.422	**0.429**	0.409
Counterfactual CCR	0.984	0.511	0.835	0.781	**0.848**
Semantic Coverage	0.097	0.026	**0.259**	0.241	0.113
Information Density (output)	0.367	0.503	0.457	0.458	**0.663**

Table 15: Paired Wilcoxon contrasts on CogBench overall, $n = 100$, two-sided. Contrasts are computed image-by-image against the s42 adapters for Groups C and D; per-seed contrasts (s42 vs s1337) and the seed-stability test are reported in companion work [15]. The C − B and D − B contrasts *fail to reject* the null at $\alpha = 0.05$ ($p = 0.20$ and $p = 0.24$ respectively); we adopt the language "no detectable difference" rather than "equivalent" to avoid conflating non-rejection with formal equivalence (a TOST-based equivalence test would require a pre-registered margin we did not specify). All seven directional contrasts at $p < .001$ survive Holm–Bonferroni correction across the nine pairwise tests at $\alpha = 0.05$. Significance markers: *** $p < .001$; ns not significant. Information-density and CCR contrasts are reported with $n = 94$ (6 invalid responses excluded uniformly across arms) in the companion paper Table 5.

Contrast	**Δ CogBench**	W	p	**Sig**
B − A (rescue, original paper)	+0.246	0.0	$< 10^{-15}$	***
C − A (rescue via density alone)	+0.254	0.0	$< 10^{-15}$	***
D − A (rescue via structure alone)	+0.238	0.0	$< 10^{-15}$	***
C − B	**+0.008**	**2154**	**0.202**	**ns**
D − B	**−0.009**	**2184**	**0.241**	**ns**
B − Base	+0.034	1207	5.9×10^{-6}	***
C − Base	+0.042	87	$< 10^{-15}$	***
D − Base	+0.025	804	3.3×10^{-9}	***
A − Base	−0.212	0.0	$< 10^{-15}$	***

The pattern compresses to: *either density or structure alone is sufficient to rescue from A on the in-domain CogBench measurement, and the joint addition of both does not produce a measurable improvement over either alone.* The 24-point swing between A and B that the original paper attributed to "density" is, in the new in-domain data, attributable to either density *or* structure. The strongest replicated effect in the new arms is on the secondary information-density axis: Group D − Base $\Delta = +0.297$ ($p = 3.9 \times 10^{-18}$), corresponding to D-trained models producing 81% denser output than the unmodified base across both replication seeds.

The in-domain 2×2 result and the OOD finding of §9.2 are complementary, not contradictory. *In-domain*, density and structure are independent rescues from the catastrophic A regime, with structure-only (Group D) producing measurable secondary advantages on capability retention and output information density. *Out-of-distribution at this corpus scale*, neither density alone nor the dense+structured combination produces detectable transfer to a 953-prompt sequestered reasoning benchmark. The two refinements jointly localize the original Density Imperative's claim: the

in-domain effect is real and is producible by either density or structure; the implicit extrapolation to general reasoning capability is not supported at this corpus scale on this base.

A mechanistic hypothesis consistent with the in-domain pattern—that Group D operates as an *elicitation regime*, reorganizing the base model's existing pre-trained associations into reasoning-accessible form rather than teaching new associations—is identified and analyzed in the companion paper [15], including the falsifiable prediction that Group D should fail on a domain the base model has not seen.

9.4 SFT's tradeoff is narrow and benchmark-selective

The Gen-1-continue SFT arm's +2.6pp verifier gain over Base is real in direction but concentrated. Per Table 2: HumanEval+ +7.0pp, MMLU-Pro +2.8pp, MATH-500 −1.0pp. SFT loses MATH. The fused-format template — the same template that produces format compulsion on judge-panel evaluation — appears to help code completion and multiple-choice knowledge retrieval, and not help mathematical reasoning. This is compatible with a mechanistic hypothesis we cannot yet test: the fused template's step-by-step prose scaffolding may guide code-completion and MC answer extraction, while adding stylistic overhead without adding solution structure on MATH-500 Level 5 problems whose answer spaces demand multiple algebraic moves. Step-level trace analysis is planned as future work (§12).

9.5 Why ORPO fails here

We do not have a definitive mechanistic account. Two candidates:

- *Corpus scale insufficient for preference-aware training.* ORPO's β-weighted odds-ratio objective may require more preference-pair examples than the 459-record Reasoning-NEST v2 corpus provides. SFT's single-target objective extracts a narrower signal ("imitate this fused template") that may be learnable at corpus sizes where ORPO cannot distinguish chosen from rejected with sufficient gradient signal.
- *Preference-pair construction mismatch.* The chosen/rejected distinction in Reasoning-NEST v2 is structural but concentrated in the *answer layer* — the chosen stream is the fused corrective target, the rejected stream is the raw failed output. The rejected half may be insufficiently close to the chosen half for ORPO's contrastive gradient to prefer *reasoning structure* over *token-level formatting*. A pair set where chosen and rejected differ on reasoning quality but share formatting may transfer differently.

Neither hypothesis is tested in this study. The β-sensitivity ablation ($\beta \in \{0.05, 0.1, 0.2\}$) planned for future work addresses the first; a chosen/rejected-construction ablation addresses the second.

10 Counterintuitive Findings

1. **Base is the best untrained reference above every fine-tuned arm on the 4-judge Borda**. See §9.1.
2. **SFT wins the verifier at benchmark-selective cost.** HumanEval+ +7.0pp, MMLU-Pro +2.8pp, MATH-500 −1.0pp.

3. **Structural-signal ORPO regresses HumanEval+ by −7.6pp (fp16)** and −13.2pp (4-bit, quantization-confounded). Specific to code; not matched by MATH or MMLU-Pro regressions of the same magnitude.

4. **Surface-signal ORPO anti-calibrates.** Brier 0.296 vs empirical-base-rate 0.204.

5. **Continuous scoring converges; categorical voting diverges.** Spearman $\rho \geq 0.79$ on factual-accuracy arm ranks across all 4-judge pairs; Cohen κ on first-place arm identity in $[0.31, 0.45]$.

6. **The meta-evaluation specimen is a rubric-weighting disagreement** (§8), not a judge-failure event.

7. **OOD: density's effect on out-of-distribution reasoning is not measurable on this corpus at this scale.** Group A and Group B differ by 0.009 Borda points and 0 percentage points on verifier-grounded ALL.

8. **In-domain (companion): density and structure are independently sufficient.** The in-domain 2×2 ablation (§9.3; full analysis in [15]) shows Group C (dense, unstructured) and Group D (sparse, structured) both produce no detectable difference from Group B (dense, structured) on the original 100-image CogBench protocol at $\alpha = 0.05$ ($p = 0.20$ and $p = 0.24$ respectively, two-sided paired Wilcoxon, $n = 100$). The 24-percentage-point swing the original paper attributed to "density" is, in the new in-domain data, attributable to either dimension; A remains the only catastrophic regime.

9. **In-domain (companion): Group D produces +81% denser output than Base, replicated across two seeds at $p = 3.9 \times 10^{-18}$.** Sparse-structured supervision teaches models to emit dramatically denser per-token output (Group B Information Density 0.457 vs Group D 0.663, both vs Base 0.367)—the largest replicated single-axis effect in the in-domain 2×2 data and an effect Group B (dense + structured) does not match.

10. **The 256-token inference ceiling truncates differentially.** Group B on MATH-500-L5: 48.6%. Base: 25.7%. SFT: 25.7%. ORPO: 5.5–6.4%. Gen-1.5-Confidence-ORPO on MMLU-Pro: 31.7% (footer-pushes-MC-responses-over-ceiling). The confound is real but the direction is predictable (§11).

11 Limitations

Table 16: Likely reviewer objections and response.

Objection	Response
Single-seed main results. Every headline comparison is one training run per arm.	The 3-seed confirmatory replication (seeds 42, 1337, 2026) on the fp16-matched pair is *completed* and reported in §3.3. The linchpin Gen-1-continue SFT accuracy (34.6%) did not reproduce (new range [20.9%, 23.3%], mean 22.19%). The paired SFT−ORPO contrast inverted sign relative to single seed (−4.25pp, 95% t-CI $[-8.93, +0.44]$ on verifier-grounded intersection $N = 361$). We withdraw the "SFT wins the verifier" portion of the Tradeoff claim; Base-dominates-Borda, surface-ORPO anti-calibration, and density-walk-back are separate measurements and survive. Other arms (Base, Group A/B, Gen-1, Gen-1p5, Gen-1.5-Confidence-ORPO) remain single-seed pending replication budget.
256-token inference ceiling truncates differentially. MATH-500 L5: Group B 48.6%, Base/SFT 25.7%, ORPO 5.5–6.4%. HumanEval+: all arms $< 1\%$. MMLU-Pro avg: all arms $< 1.5\%$ (Gen-1.5-Confidence-ORPO 31.7% on MMLU due to appended footer).	Acknowledged. The per-benchmark truncation audit (`truncation_audit_conf953.json`) decomposes the confound cleanly. HumanEval+ is truncation-clean (all arms $< 1\%$) so the SFT +7.0pp gain and ORPO −7.6pp regression are un-confounded. MMLU-Pro is truncation-clean for all arms except Gen-1.5-Confidence-ORPO (31.7%). MATH-500 is differentially truncated (Base/SFT 25.7%, ORPO 6.4%, Group B 48.6%); its sub-findings are scoped as confounded. The Tradeoff claim anchors on HumanEval+ and MMLU-Pro, where truncation is not a confound. A 200-prompt re-run at 1024 tokens on A100 (not L4, to avoid VRAM pressure on long generations) is pre-registered for the revision cycle if an area-chair challenge requires it. **Research-enabling consequence we declare openly:** the 256-token ceiling is also the precondition that produced the rubric-disagreement specimen in §8. Gen-1's geometrically correct reflection-axis derivation was inference-truncated before the formula was applied to $(-2, 3)$; had the generation completed, three of four frontier judges would have selected Gen-1 on a clean correct answer and the completeness-weighting bias they exhibited (preferring Base's complete-but-arithmetically-wrong full response over Gen-1's correct-direction-but-truncated derivation) would not have been observable. The truncation we report transparently as a confound is also what made the LLM-judge bias finding visible.
No human judge baseline. All "judge panel" claims are about 4 LLMs.	Acknowledged. A pre-registered expert-human comparison on a stratified 50-prompt subset (25 MATH-500 L5, 25 HumanEval+) is planned for the revision cycle. Protocol and stratification scheme are fixed and the sample is released at `https://github.com/codex-curator/supervision-tradeoff/tree/main/supplementary/human_expert_sample`. If the area chair gates acceptance on the comparison, the protocol is ready and the sample ships within 1–2 weeks.
Verbosity bias in the 4-judge panel was not explicitly audited. Preference-ranking LLMs are known to favor longer and more verbose responses independent of factual content, and our completeness-weighting finding in the §8 rubric-disagreement specimen is one instantiation of this class of bias.	Acknowledged. Token-length balancing across arms was not imposed at inference (all arms share the same `max_new_tokens` = 256 ceiling, and different arms exhibit systematically different output lengths; truncation rates per benchmark are in the audit). Future work: regression-adjust the Borda aggregate by per-response token count to isolate the verbosity component, or condition the judge rubric to score length-normalized quality.
Quantization-precision asymmetry across arms. Gen-1.5-ORPO was trained on a 4-bit quantized base while fp16-matched arms used fp16	Headline Tradeoff is stated under the fp16-matched contrast (Gen-1-continue vs Gen-1.5-continue-ORPO). The 4-bit Gen-1.5-ORPO result is italicized in tables. See §5.3.

12 Future Work

1. **3-seed confirmatory replication of the fp16-matched pair** — *completed in this revision cycle* and reported in §3.3; results inverted the headline contrast. Next replication targets: the remaining six single-seed arms (Base, Group A, Group B, Gen-1, Gen-1.5, Gen-1.5-Confidence-ORPO) at three seeds each, conditional on GPU budget.

2. **200-prompt stratified re-run at max_new_tokens=1024 on A100**, pre-registered for the revision cycle if AC gates on MATH-500 sub-findings. The 256-token truncation audit already anchors the Tradeoff claim on truncation-clean HumanEval+ and MMLU-Pro; this re-run would close the MATH-500 confound specifically.

3. **Human-expert comparison on 50-prompt stratified subset.** Pre-registered; protocol fixed; sample released at `https://github.com/codex-curator/supervision-tradeoff/tree/main/supplementary/human_expert_sample`; execution in the revision cycle.

4. **Ablation sweep.** $\beta \in \{0.05, 0.1, 0.2\}\times$ LoRA rank $\in \{16, 32\}$ on the fp16-matched pair. Deferred until the 3-seed replication confirms direction.

5. **Preference-aware objective generalization.** DPO [19], KTO [6] on the same structural-signal and surface-signal preference pairs.

6. **Structural-signal calibration preference pairs.** Chosen/rejected distinction tied to calibrated-abstention vs over-confident-hallucination with auditable provenance.

7. **Scale.** 70B-class bases, 10k–100k-record corpora.

13 Conclusion

On the Reasoning-NEST v2 corpus with the Llama 3.2 11B Vision-Instruct base, evaluated on a 953-prompt OOD corpus by a deterministic verifier and a 4-judge cross-lineage panel, we observe three distinct supervision outcomes and none of them produce a model that improves over Base on the 4-judge Borda aggregate. Naive SFT on structural-signal single-target fused-format supervision uniquely exhibits a verifier-vs-judge tradeoff: it wins the verifier by +2.6pp ALL (concentrated on HumanEval+ and MMLU-Pro; loses MATH-500 by −1pp) and falls to Borda-5th of 8. ORPO on structural-signal preference pairs is strictly dominated by Base on both axes; it produces no tradeoff, only a two-axis loss. ORPO on surface-signal preference pairs additionally anti-calibrates against an empirical-base-rate predictor: Brier 0.296 vs 0.204, robust across three probability mappings.

The 4-judge Borda panel agrees far more on continuous factual-accuracy scoring (Spearman $\rho \geq 0.79$ on all pairs) than on categorical first-place arm identity (Cohen κ in $[0.31, 0.45]$). The continuous assessment converges; the categorical winner-identity does not. A meta-evaluation specimen (§8) on a single event renders the divergence concrete: three of four judges weight method-recognition-plus-completeness; one judge weights correctness-of-geometric-direction. Both rubric weightings are defensible; the disagreement is the methodological finding, not evidence that any judge failed.

The 11.1× per-record density advantage documented in prior work [15] produces zero measurable OOD benefit on this corpus at this scale: Group A and Group B differ by 0.009 Borda points and 0 percentage points on verifier-grounded ALL. Density transfers in-domain; on this OOD corpus it does not transfer measurably. An honest scope refinement replaces the OOD-transfer extrapolation.

The companion 2×2 ablation reported in [15] provides a complementary in-domain refinement: closing the diagonal of the original Density Imperative two-cell design with Groups C and D reveals that density and structure are independently sufficient on the in-domain CogBench measurement, while only the joint absence of both—Group A—is uniquely catastrophic. Together, the in-domain and OOD refinements localize the Density Imperative's claim to its empirical setting: in-domain, either density or structure alone produces the cognitive-depth gain; out-of-distribution at this corpus scale, neither density alone nor the dense+structured combination produces a detectable transfer effect.

Releasing the 953-prompt corpus, 8 LoRA adapters, 873 4-judge judgments, verbatim judge prompts, Wilson CIs, per-arm per-benchmark truncation audit at the 256-token ceiling, pre-registered human-baseline protocol, and reproducibility harness is intended to let the community reproduce the Tradeoff structure on additional seeds, bases, scales, corpora, and preference-aware objectives. The paper's job is to report what we observed honestly; extension and falsification are the community's.

Additional Information and Declarations

Data Availability

All artifacts are released for independent verification: the Reasoning-NEST v2 corpus (459 admitted records; 100-record Gen-0 pilot) at huggingface.co/datasets/Metavolve-Labs/reasoning-nest-v2 and archived on Zenodo [17]; the 8 LoRA adapters, 873 four-judge judgments, verbatim judge system prompts, the verifier-grounded intersection JSON used in the three-seed replication, the per-arm per-benchmark 256-token truncation audit, and the full reproducibility harness at github.com/codex-curator/supervision-tradeoff. All artifacts are released under CC BY 4.0 (data) / Apache-2.0 (code). The 953-prompt out-of-distribution corpus is MD5-hash-verified to have zero overlap with the 459 training records.

Funding

This work was self-funded by the author through Metavolve Labs, Inc. Compute (NVIDIA L4 and preemptible A100 GPU time) was funded internally by Metavolve Labs. No external grant funding supported this work. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests

The author is the founder of and is affiliated with Metavolve Labs, Inc., which produced the Reasoning-NEST v2 corpus, the Alexandria Aeternum dataset, and the provenance/Cognitive-Nutrition framework referenced in this work [14–17]. This affiliation is disclosed in the interest of transparency; the empirical findings reported here are predominantly null or self-correcting (including the withdrawal, under replication, of the author's own prior single-seed claim) and do not advantage any Metavolve Labs product.

Author Contributions

Tad MacPherson conceived and designed the experiments, performed the experiments, analyzed the data, authored the reproducibility harness, and wrote the manuscript.

Use of Artificial Intelligence

Per PeerJ's AI policy, AI systems were used in two clearly separated roles. (1) *As measurement instruments under study*: the four cross-lineage panel judges (Claude Sonnet 4.6, Gemini 3.1 Pro, GPT-5.4, Grok-4-1) and the three-teacher corpus-construction pipeline; all model identifiers are pinned and judge prompts are released (see §5.6 and the Appendix). (2) *As authoring assistance*: an LLM-based coding and writing assistant was used for drafting, table formatting, and reference cross-checking. All scientific claims, the experimental design, the decision to run the confirmatory replication, and the interpretation of its sign-inverting result are the author's own; every reported number was produced by the released harness, not generated by an LLM.

Ethics

This study involved no human subjects and no personally identifiable information; institutional review board approval was therefore not applicable. All evaluation prompts derive from public benchmarks (MATH-500, HumanEval+, BigCodeBench, MMLU-Pro) and hand-curated reasoning prompts released with the corpus.

References

[1] Akhtar, M., Benjelloun, O., Conforti, C., et al. (2024). *Croissant: A metadata format for ML-ready datasets.* NeurIPS 2024.

[2] Bai, Y., Kadavath, S., Kundu, S., et al. (2022). *Constitutional AI: Harmlessness from AI feedback.* arXiv:2212.08073.

[3] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3.

[4] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. *NeurIPS 2023*, 10088–10115.

[5] Dong, B., Ni, M., Huang, Z., Yang, G., Zuo, W., & Zhang, L. (2025). *MIRAGE.* OpenReview.

[6] Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., & Kiela, D. (2024). KTO: Model alignment as prospect-theoretic optimization. *ICML 2024*.

[7] Gebru, T., Morgenstern, J., Vecchione, B., et al. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92.

[8] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *ICML 2017*, 1321–1330.

[9] Hong, J., Lee, N., & Thorne, J. (2024). ORPO: Monolithic preference optimization without reference model. *EMNLP 2024*.

[10] Hu, E. J., Shen, Y., Wallis, P., et al. (2021). *LoRA: Low-rank adaptation of large language models.* arXiv:2106.09685.

[11] Kang, K., Wallace, E., Tomlin, C., Kumar, A., & Levine, S. (2025). Unfamiliar finetuning examples control how language models hallucinate. *NAACL 2025*.

[12] Kumar, A., Zhuang, V., Agarwal, R., et al. (2025). Training language models to self-correct via reinforcement learning. *ICLR 2025*.

[13] Lightman, H., Kosaraju, V., Burda, Y., et al. (2024). Let's verify step by step. *ICLR 2024*.

[14] MacPherson, T. (2026a). *Cognitive Nutrition.* Zenodo. https://doi.org/10.5281/zenodo.18667742

[15] MacPherson, T. (2026b). *The Density Imperative, Refined: A 2×2 Ablation Reveals Independent Sufficiency of Density and Structure in Vision-Language Fine-Tuning.* Zenodo. https://doi.org/10.5281/zenodo.20162589 (concept DOI https://doi.org/10.5281/zenodo.20162588, resolves to latest version).

[16] MacPherson, T. (2026c). *The Entropy of Recursion.* Zenodo. https://doi.org/10.5281/zenodo.18436975

[17] MacPherson, T. (2026d). *Reasoning-NEST v2 Gen-0 pilot corpus* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.19640552

[18] Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.

[19] Rafailov, R., Sharma, A., Mitchell, E., et al. (2023). Direct preference optimization. *NeurIPS 2023*.

[20] Zelikman, E., Wu, Y., Mu, J., & Goodman, N. (2022). STaR. *NeurIPS 2022*.

[21] Zelikman, E., Harik, G., Shao, Y., et al. (2024). *Quiet-STaR.* arXiv:2403.09629.

[22] Zhang, W., Sun, Y., Huang, P., et al. (2025). *MIRAGE-Bench.* OpenReview.

[23] Zhu, C., Xu, B., Yang, A., et al. (2025). Rationales are not silver bullets. *Findings of ACL 2025*.

A Reproducibility Checklist

A.1 Environment

Ubuntu 22.04, CUDA 12.9, PyTorch 2.9, Transformers 4.46, PEFT 0.13, TRL 0.12, Unsloth 4-bit mirror, Python 3.11.

A.2 Model and Adapter Artifacts

Base: `unsloth/Llama-3.2-11B-Vision-Instruct` (4-bit BnB and fp16). All 8 adapters archived under `gs://cognitive-nutrition-experiment/output/`.

A.3 Judge Model IDs (pinned)

Claude: `claude-sonnet-4-6`; Gemini: `gemini-3.1-pro-preview`; GPT: `gpt-5.4`; Grok: `grok-4-1-fast-reasoni`

A.4 Corpus and Evaluation Artifacts

Training corpus MD5-hashed; 953-prompt corpus hash-verified zero-overlap; aggregates in `final_aggregate_conf95` `verifier_grounded_v2_conf953/`, `footer_correctness_conf953/`, `gemini_subset_sensitivity_conf953.js` `wilson_cis_conf953.json`, `truncation_audit_conf953.json`.

A.5 Inference and Judging Configuration

Temperature 0.7; `max_new_tokens` 256 (main); `max_new_tokens` 1024 (planned stratified re-run); top-p 1.0; fresh seed per run; judge temperature 0.0.

B Judge System Prompts (verbatim)

Per the reproducibility expectations stated in the main text, the verbatim system prompt used for all four judges on the 953-prompt sweep is reproduced here.

```
[SYSTEM]
You are an impartial evaluator of LLM outputs on reasoning and code tasks.
You will see a prompt and 8 candidate responses labeled base, groupa, groupb,
gen1, gen1continue, gen1p5, gen1p5continue, gen1p5confidence. Score each
response on the following 0-5 ordinal dimensions:

  - factual_accuracy: how factually correct the claims and computations are
  - reasoning_validity: whether each reasoning step follows from the previous
  - structural_quality: organization, signposting, and readability
  - epistemic_hedging: calibrated hedging when uncertain (higher is better)

And on the following binary dimensions:
  - hallucination: 0 = none, 1 = minor, 2 = major

Then produce a preference_ranking as an ordered list of the 8 arms, best to
worst. Ties should be broken by your best judgment of overall quality.
Finally, write a 1-2 sentence rationale explaining your top pick and your
bottom pick.

Output must be valid JSON matching the schema:
  {
    "scores": {
      "<arm>": {
        "factual_accuracy": int,
        "reasoning_validity": int,
        "structural_quality": int,
        "epistemic_hedging": int,
        "hallucination": int
      },
      ...
    },
    "preference_ranking": [<arm>, <arm>, ...],
    "rationale": "<1-2 sentences>"
  }

Evaluate at temperature 0.
```

B.1 Pre-registered Human-Expert Baseline Protocol

For the pre-registered expert comparison scoped in §11, the stratified 50-prompt sample is selected as follows:

- 25 prompts from MATH-500 Level 5 (hardest mathematical reasoning stratum).
- 25 prompts from HumanEval+ (largest arm-regression benchmark under ORPO).

Event IDs, ground truths, and all 8 arms' verbatim outputs are released at the Hugging Face repo in `pre_registered_human_eval_sample.json`. The expert evaluator (committed: one mathematician, one CS faculty member with software-engineering experience) will be blinded to arm identity and will rank all 8 outputs per prompt on the same rubric dimensions as the 4-judge LLM panel. Expert rankings will be aggregated via the same Borda procedure and compared against (a) the 4-judge LLM Borda aggregate and (b) the deterministic verifier pass/fail labels.

The pre-registration is fixed here; deviation from this protocol in the revision cycle will be disclosed as a pre-registration violation. Expected execution time: 1–2 weeks post-acceptance trigger.

C Phase-1 In-Distribution Diagnostic (superseded)

This appendix retains the Phase-1 analysis superseded by §7. Post-hoc MD5 hashing disclosed 100/100 exact overlaps between Phase-1 evaluation prompts and the 459-record training set. Base, Group A, Group B are evaluated out-of-domain relative to their Phase-0 supervision; Gen-1 is evaluated strictly in-distribution. The 100% overlap motivated the design of the Phase-2 pre-registered sequestered benchmark. Phase-1 results are scoped as in-distribution diagnostic only.

C.1 Phase-1 4-arm results (in-distribution, data-leaked)

Table 17: Phase-1 primary endpoint on 100-prompt data-leaked corpus.

Metric	**Base**	**Group A**	**Group B**	**Gen-1**
Major hallucination rate ↓	27%	26%	25%	20%
Minor hallucination rate	20%	22%	20%	32%
Clean output rate ↑	53%	52%	55%	48%
Factual accuracy (0–5) ↑	3.37	3.29	3.36	3.11
Structural quality (0–5) ↑	3.79	3.84	3.90	2.92
Reasoning validity (0–5) ↑	3.19	3.19	3.24	3.02
Preference 1st place	30	13	20	37
Preference last place	17	25	18	40

Gen-1 reduced major hallucinations by 5 percentage points relative to Group B but increased minor hallucinations by 12 percentage points. Gen-1 was the most polarizing arm. Format-compulsion diagnosis (Gen-1 emitted the fused Reasoning/Avoidance-rule/Answer template on 100% of in-distribution prompts) is the key mechanism that the Phase-2 corpus preserved on OOD: on the 953-prompt confirmatory, Gen-1's template compliance is 94%.

D Interim Directional OOD Evaluation (100-prompt, superseded)

Prior versions reported three interim 100-prompt OOD-quick passes. Superseded by §7.

E BigCodeBench Canonical-Harness Verifier (deferred)

BigCodeBench's canonical test harness requires a dedicated subprocess sandbox with domain-restricted imports and per-problem test-battery loader. The verifier-grounded analysis in §7.2 used `evalplus` for HumanEval+ and did not include BigCodeBench's harness. BigCodeBench outputs were included in the 4-judge rubric scoring; the deterministic pass-rate is deferred to a future verifier release.