Reasoning-NEST v2 and Gen-1 Negentropic Self-Correction:
A Provenance-Rich Failure-Topology Corpus and an Honest Negative Result on Naive Fused-Format Continuation

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

April 19, 2026

Abstract

Failure-correction data are increasingly used to improve reasoning models, but the field still lacks auditable corpora that preserve where a response failed, why it failed, and how the correction was verified. This paper presents **Reasoning-NEST v2**, a provenance-rich post-failure corrective corpus in which each record stores the original prompt, failed output, verified correction, step-level validity annotations, a named failure mechanism, a causal DAG, a backdoor-adjustment formula, an error-topology label, an avoidance rule, and provenance metadata. The released artifact includes a 100-record CC-BY-4.0 Gen-0 pilot on Zenodo; the full study used 459 admitted records. We further report a continuation experiment on a frozen-vision Path-A LoRA protocol over a multimodal base model, trained on a single NVIDIA L4 in 9.23 minutes for roughly $5. In a four-arm evaluation, Gen-1 reduced major hallucination rate from 25% to 20% relative to the prior dense-supervision baseline Group B and received the most first-place preferences (37/100), but it also increased minor hallucinations (32%), reduced clean outputs (48%), and had the worst structural quality (2.92/5), reasoning validity (3.02/5), and last-place preferences (40/100). A critical caveat is that every evaluation prompt appeared verbatim in Gen-1's training set. Accordingly, the result is in-distribution only and cannot support any out-of-domain generalization claim. Stratification by complexity reveals the mechanism cleanly: Gen-1 wins 68% of first-place votes on hard prompts but loses 61% of last-place votes on easy ones. The paper's main empirical contribution is therefore negative-leaning: naive fused-format SFT can induce *format compulsion* and fabricated avoidance rules. The dataset, evaluation protocol, frozen-vision continuation recipe, and diagnosed anti-pattern are the central contributions.

Keywords: failure-topology supervision; post-failure corpus; self-correction; evaluation science; data leak; format compulsion; data-centric AI; vision-language models; LoRA.

1 Introduction

Post-failure supervision is attractive because it targets observed failure surfaces rather than generic instruction-following behavior. Yet most reasoning datasets still collapse the interesting part of the problem into a single corrected answer or a free-form rationale. What gets lost is exactly what evaluation science needs: the failed output, the failure mechanism, the causal assumptions behind the correction, the provenance of the correction, and the conditions under which the correction should generalize. That omission matters especially for the renamed NeurIPS Evaluations & Datasets track,

whose 2026 call explicitly places evaluation itself at the center of scientific inquiry and welcomes failure analyses, audits, documentation methodologies, negative results, and artifacts that sharpen how evaluative claims are made and interpreted [17, 18].

This paper extends a prior research program on provenance-verified, semantically dense supervision. In earlier work, we argued that supervision quality is *signed*: sparse or shallow post-training examples can degrade reasoning behavior, while dense, structured, provenance-tracked examples can improve reliability and reasoning quality [12–14]. Phase-0 of that program compared three arms on the same multimodal base model: an unmodified Base model, a sparse Cognitive-Nutrition continuation (Group A), and a dense Cognitive-Nutrition continuation (Group B). On a 100-prompt out-of-domain screening set, major hallucination rate fell from 26% in Base to 17% in Group B, clean output rose from 54% to 60%, structural quality rose from 3.49 to 3.80, and reasoning validity rose from 3.22 to 3.37. That study established Group B as the "cognitively nourished" baseline against which the present paper continues.

The present paper asks a narrower question than Phase-0. Instead of training on richly annotated image-text corpus items, can one continue the same line of work by training directly on archived failures converted into structured corrective records? The proposed answer is **Failure Topology Supervision**: a continuation regime in which the training substrate is not merely a corrected answer and not merely a rationale, but a typed failure artifact containing the failed response, validated reasoning fragments, named error mechanism, causal graph, avoidance rule, and audit metadata.

The paper's contributions are fivefold.

1. **Reasoning-NEST v2**, an auditable failure-topology corpus suitable for the Evaluations & Datasets framing of evaluation artifacts, released under CC-BY-4.0.

2. **Failure Topology Supervision** as a methodological object distinct from rationale augmentation, process reward modeling, online self-correction RL, or inference-time critique.

3. A cheap and reproducible **frozen-vision continuation protocol** ("Path A") for text-only corrective supervision on a vision-language base.

4. A concrete **negative-leaning empirical result**: naive fused-format SFT teaches the *shape* of self-correction more robustly than the judgment of *when* to self-correct.

5. Documentation of an anti-pattern that should matter regardless of model family: exact prompt overlap between training and evaluation can coexist with mediocre quality and easily produce misleading headline narratives.

The paper is intentionally not a hype paper. It does **not** claim that Gen-1 validates the Density Imperative on out-of-domain generalization. The exact opposite lesson is more defensible and more useful. When the training objective, evaluation split, and correction format are not carefully separated, a model can memorize the costume of self-correction without actually learning disciplined corrective reasoning. Under the 2026 E&D call, that is not a side observation; it is the main scientific result [18].

2 Related Work

Failure Topology Supervision sits near, but is not reducible to, several influential lines of work on reasoning supervision. STaR bootstraps rationales from answer supervision and keeps only rationales that lead to correct answers [20]. Quiet-STaR generalizes the idea to latent token-level thoughts during continued pretraining [21]. Let's Verify Step by Step moves from final-answer supervision to step-level labels and verifier training, releasing PRM800K as process-supervision infrastructure [11]. SCoRe goes further and trains self-correction online with reinforcement learning, while explicitly reporting that offline SFT on correction traces can fail because of distribution mismatch or behavior collapse [10].

These differences matter. STaR begins with sparse rationale exemplars and correct answers. Quiet-STaR learns hidden deliberative structure from arbitrary text. Let's Verify Step by Step is process supervision aimed at reward modeling and verification. SCoRe is multi-turn online RL under the model's own correction distribution. By contrast, Reasoning-NEST v2 begins from *observed failures* and turns those failures into auditable artifacts for *offline* continuation. The released object is therefore the corpus itself, together with its schema and protocol, not a reward model and not an online RL policy.

There is also a family resemblance to Constitutional AI, which samples model outputs, critiques them under explicit principles, revises them, and then trains on those revisions [2]. But Constitutional AI's core contribution is a critique-and-revision alignment pipeline under constitutional principles, whereas the present paper contributes a reusable corpus of archived failures whose units encode event, topology, causality, and provenance in one inspectable artifact.

On the evaluation side, Reasoning-NEST v2 belongs with recent work that treats hallucination analysis as a first-class benchmark-design problem. MIRAGE isolates reasoning hallucinations in multimodal reasoning chains by constructing cases where perception is intact but reasoning still fails [5]. MIRAGE-Bench targets agentic hallucinations in interactive trajectories [22]. The present corpus differs in two ways. It is a post-failure *training* corpus rather than a benchmark alone, and it stores causal and provenance layers that make each correction auditable.

Recent cautionary work is especially relevant. Zhu et al. [23] argue that rationales are not uniformly beneficial and can deteriorate performance even when they sometimes improve reliability. Kang et al. [9] show that unfamiliar fine-tuning examples can directly shape how models hallucinate. Those findings motivate restraint here: Failure Topology Supervision is proposed not as a universal improvement recipe, but as a narrower and more falsifiable supervision framing whose success and failure can be inspected by slice, topology, and provenance field.

As a dataset contribution, the corpus also follows the logic of structured documentation rather than opaque release. Datasheets for Datasets argued that datasets should carry explicit documentation of motivation, composition, collection process, intended use, and maintenance [6]. Croissant extends that ethos into machine-readable dataset metadata [1]. Reasoning-NEST v2 adopts that direction by pairing narrative documentation with machine-readable schema design and provenance fields that support audit, reuse, and selective filtering.

3 Reasoning-NEST v2 Corpus

Reasoning-NEST v2 is a structured JSON corpus built from failed model outputs that were converted into corrective artifacts through a three-teacher consensus pipeline. The public release is a 100-

record Gen-0 Pilot Corpus deposited on Zenodo under CC-BY-4.0 on 2026-04-18 [15]. The full local corpus used in the continuation experiment contains 459 admitted records: 20 from `real/`, 25 from `benchmark/`, 55 from `benchmark_v2/`, and 359 from `benchmark_v3/`. The public pilot is therefore the first 100 admitted records of the 459-record local corpus.

3.1 Design Principle

Each record preserves the original failure surface rather than flattening it into a single corrected target. The central design principle is that a failure example should be inspectable both as a learning example and as an evaluation object.

3.2 Schema

Table 1: Core Reasoning-NEST v2 record schema.

Component	Representative fields	Function
Event	`event.prompt`, `event.failed_output`, `event.correction`	Preserves the original task, the failed response, and the verified correction
Reasoning	`reasoning.cot_trace[]`, `reasoning.failure_narrative`, `reasoning.failure_mechanism`	Stores local step validity, a narrative account of failure, and a named mechanism
Causality	`causality.graph`, `causality.adjustment_formula`	Makes causal assumptions explicit and auditable [19]
Topology	`topology.error_class`, `topology.avoidance_rule`, `topology.complexity_score`	Encodes the reusable structural type of error and a declarative preventive rule
Provenance	`provenance.oracle_fingerprint`, `provenance.trust.gradient`, `provenance.densified_at`, `provenance.source_event_hash`	Supports audit, consensus inspection, confidence filtering, and source traceability

3.3 Three-Teacher Consensus Pipeline

The pipeline uses a primary oracle via Google Vertex AI (Gemini 3.1 Pro Preview), a secondary oracle via OpenAI API (GPT-4o), and a judge/arbitrator via Anthropic API (Claude Sonnet 4.6). The primary oracle drafts the corrective artifact, the secondary oracle independently verifies and amends it, and the judge arbitrates disagreements while assigning a consensus score. Only admitted records enter the corpus. This design is intentionally expensive but auditable: every record exposes not only the correction but the path by which the correction was accepted.

3.4 Error-Class Taxonomy

The `topology.error_class` field enumerates seven classes reflecting the actual composition of the released corpus (the canonical enum is fixed by the schema definition):

Table 2: Reasoning-NEST v2 topology error classes.

Error Class	Definition
`compiler_syntax`	Syntax-level code failures detectable by a compiler or static analyzer
`runtime_semantic_code`	Semantic code errors that compile but produce incorrect runtime behavior
`math_logic_fallacy`	Arithmetic or algebraic breakdowns, unit misalignments, logical fallacies
`factual_hallucination`	Assertions or inferences stated as verified facts without evidentiary backing
`multi_step_planning`	Failure to adhere to sequential operations required by prompt instructions
`visual_perceptual`	Conclusions extracted from visual data not supported by the underlying pixel topology
`cultural_contextual`	Misreadings of cultural, historical, or pragmatic context

Each admitted record also includes a declarative `topology.avoidance_rule` and a scalar `topology.complexity_sc` in $[0, 1]$. The causality layer stores a Pearl-style structural causal model DAG together with a backdoor-adjustment formula, so the record captures not only what the correct answer is, but what assumptions license the move from evidence to answer [19].

The release is positioned as an evaluative artifact, not only as training data. That choice aligns with datasheet-style documentation and Croissant-style machine-readable metadata, both of which are increasingly central to the E&D track's expectations for dataset submissions [1, 6, 18].

4 Methodology

4.1 Continuation Protocol

The Gen-1 experiment used the same family of multimodal base as Phase-0—Llama 3.2 11B Vision-Instruct via the Unsloth 4-bit quantized mirror—but the continuation problem was different. The corpus records are text-only, while the base model is a vision-language model. Three adaptation pathways were considered:

- **Path A** (adopted): a fresh LoRA trained on the base model while explicitly freezing vision layers, preserving the documented multimodal stack and text generation head. The `UnslothVisionDataCollator` is retained but operates on text-only records.
- **Path B**: an offline merge followed by a standard text collator. Deferred for future work due to VRAM pressure during the 16-bit merge on L4.
- **Path C**: injection of a sentinel 1×1 image to force a multimodal pathway for text-only records. Rejected because it corrupted behavior during design-stage testing and introduces an always-on zero-information visual signal that the model learns to route around.

This frozen-vision decision is not a nuisance workaround; it is a methodological choice. The experimental question is whether text-only failure-topology supervision can alter the model's reasoning behavior without perturbing the visual subsystem. By freezing vision layers and updating language-side LoRA adapters only, the protocol makes the modality boundary explicit and reproducible. The continuation uses a parameter-efficient adapter regime on a quantized base, in the practical lineage of LoRA [8] and QLoRA [4].

4.2 Training Target Format

The training target fused validated reasoning, avoidance rule, and answer into plain natural-language sections with no XML tags:

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

Avoidance rule: {topology.avoidance_rule}

Answer: {event.correction}
```

Only `is_valid: true` reasoning steps were included in the target. Invalid reasoning steps were intentionally excluded because they encode the failed path rather than the corrective one; under standard causal language modeling, exposing the model to invalid steps—even explicitly labeled as negative—risks teaching the flawed reasoning path via SFT target assimilation.

4.3 Training Configuration and Implementation Footprint

Table 3: Verified Gen-1 training configuration.

Setting	**Value**
Adapter	LoRA rank 32, alpha 16, dropout 0.0
Trainable scope	Language layers only; all seven linear targets (`q, k, v, o, gate, up, down_proj`)
Vision setting	`finetune_vision_layers=False`; `finetune_language_layers=True`
Optimization	Learning rate 2e-4; cosine scheduler; warmup ratio 0.03; weight decay 0.01
Precision and memory	`bf16`; `adamw_8bit`; `use_gradient_checkpointing="unsloth"` (load-bearing for L4 VRAM)
Batching	Per-device batch 1; gradient accumulation 8; effective batch 8
Schedule	3 epochs over 459 records; 171 `max_steps`

Table 4: Verified implementation footprint.

Item	**Value**
Hardware	NVIDIA L4, 24 GB VRAM, Cloud Run Jobs, `us-central1`
Training wall-time	554 seconds (9.23 minutes)
Total wall-time incl. model load and upload	748 seconds (12.47 minutes)
Final training loss	1.3812
Adapter size	466.6 MiB
Approximate GPU cost	∼$5 USD

This footprint matters for two reasons. First, it makes the protocol cheap to reproduce on consumer-grade hardware. Second, it underscores the paper's methodological point: the scientific interest here lies in the corpus and the evaluation design, not in massive compute expenditure.

4.4 Evaluation Protocol

Inference used the same multimodal base-plus-adapter stack on the 100-prompt Gen-0 Pilot corpus, with temperature 0.7 and `max_new_tokens=256`. Evaluation was a blinded four-arm ablation across

Base, Group A, Group B, and Gen-1. The judge was Claude Sonnet 4.6 at temperature 0.0, using an **absolute rubric first** followed by **pairwise preference ranking**, with structured JSON outputs.

This protocol contributes two things beyond the raw scores. First, it preserves within-run comparability across all four arms. Second, it generates qualitative rationales that are diagnostic of *why* an arm won or lost, which is essential once topline gains turn out to be unstable or misleading.

4.5 Endpoints and Scoring Rubric

Confidence-format and calibration endpoints are retained because they belong to the benchmark specification, even though the primary Gen-1 run did not yet exercise explicit confidence emission. When confidence outputs are present, Brier score and Expected Calibration Error are the intended summaries [3, 7].

Table 5: Operational definitions.

Endpoint	**Definition**
Major hallucination rate	Binary. A response counts as a major hallucination if it contains a materially false assertion, fabricated action, invented observation, or unsupported causal/evidentiary claim that would mislead a user or automated executor. Stylistic awkwardness, incompleteness, hedged uncertainty, or benign omission does NOT count. Borderline cases require adjudication by verifier or human.
Minor hallucination rate	Binary. One or two suspicious-but-unverified specifics that do not rise to major-hallucination standard.
Clean output rate	Binary. Free of major hallucinations, gross formatting corruption, invalid tool/code actions, and catastrophic instruction drift. Partial but honest abstention can still be clean.
Reasoning validity (0–5)	0=incoherent/self-contradictory; 1=mostly invalid; 2=some locally valid steps but incorrect global logic; 3=broadly plausible with substantive flaw; 4=logically well-supported with minor defects; 5=fully coherent, evidence-aligned, stepwise valid.
Structural quality (0–5)	0=unreadable or malformed; 1=severely disorganized; 2=minimally structured; 3=adequate organization; 4=strong structure and signposting; 5=concise and consistently task-appropriate.
Epistemic hedging (0–5)	0=confident wrong answer with no hedging; 2=mixed; 5=appropriately hedged when uncertain, refuses to fabricate.
Confidence output	Require $p \in [0, 1]$ or integer 0–100 with deterministic parser mapping to $[0, 1]$. Unparseable strings logged, reported, handled by preregistered fallback.
Calibration metrics	Brier score + Expected Calibration Error on parsed confidence outputs. Semantic-uncertainty diagnostics for free-form outputs where lexical variation masks equivalent meanings.

4.6 Pre-Registered Hypotheses

Table 6: Status of pre-registered hypotheses in this paper.

Hypothesis	Status
H1: Frozen-vision Path-A continuation reduces major hallucination rate relative to Group B	Directionally supported **only** on a leaked in-distribution screening set; not evidence of out-of-domain generalization
H2: Hybrid process supervision improves reasoning validity and calibration	Not tested in the primary run; no correction-only baseline was executed
H3: Negative/verifier examples improve calibration	Not tested; deferred to a planned Gen-1.5 variant
H4: Metadata routing improves hard-slice performance when labels are predicted non-oracularly	Not tested; requires a separately validated router

5 Results

5.1 Data Leakage Disclosure

Critical protocol disclosure. Every one of the 100 evaluation prompts appeared verbatim in Gen-1's 459-record training set. MD5 hashing over prompt strings showed **100/100 exact overlaps**. Consequently, Base, Group A, and Group B are evaluated out-of-domain relative to their Phase-0 Cognitive-Nutrition supervision, whereas Gen-1 is evaluated **strictly in-distribution**. No result in this paper should be read as evidence that Gen-1 improves out-of-domain generalization. The leak is total, quantified, and central to interpretation.

This caveat is not a footnote. It changes the meaning of the headline numbers. Gen-1's only clear quantitative win over Group B occurs under a memorization advantage that the other arms do not possess. Even with that advantage, Gen-1 loses badly on several quality metrics.

5.2 Primary Endpoint Results

Table 7: Four-arm evaluation on the 100-prompt Gen-0 Pilot corpus.

Metric	Base	Group A	Group B	Gen-1	Δ vs Group B
Major hallucination rate ↓	27%	26%	25%	**20%**	−5pp
Minor hallucination rate	20%	22%	20%	32%	+12pp
Clean output rate ↑	53%	52%	55%	48%	−7pp
Factual accuracy (0–5) ↑	3.37	3.29	3.36	3.11	−0.25
Epistemic hedging (0–5) ↑	2.87	2.91	2.87	**2.99**	+0.12
Structural quality (0–5) ↑	3.79	3.84	3.90	2.92	−0.98
Reasoning validity (0–5) ↑	3.19	3.19	3.24	3.02	−0.22
Preference 1st place (of 100)	30	13	20	**37**	most preferred
Preference last place (of 100)	17	25	18	**40**	most rejected

The directional pattern is clear. Relative to Group B, Gen-1 reduces *major* hallucinations by 5 percentage points, but increases *minor* hallucinations by 12 percentage points, decreases clean

output by 7 percentage points, loses 0.25 factual-accuracy points, loses 0.98 structural-quality points, and loses 0.22 reasoning-validity points. It gains only marginally on epistemic hedging. The preference signal is not celebratory; it is polarized. Gen-1 is both the most preferred and the most rejected arm.

5.3 Per-Arm Characterization

Base remained unexpectedly competitive. It received 30 first-place preferences and only 17 last-place preferences, suggesting a stable if imperfect profile. Its outputs were less stylized than Gen-1's and, within this run, less vulnerable to catastrophic over-structuring.

Group A underperformed in both preference and error profile. It received only 13 first-place preferences and 25 last-place preferences. Sparse continuation delivered neither the stability of Base nor the balance of Group B.

Group B remained the strongest conventional continuation arm. It achieved the highest clean-output rate in the run at 55%, the strongest structural-quality score at 3.90, and the best reasoning-validity score at 3.24. The prior dense-supervision baseline remained the best-balanced arm under the current judge configuration.

Gen-1 was a specialist with side effects. It appears to have learned a more aggressive self-correction style that sometimes reduced severe hallucination, especially on prompts aligned with the corpus's failure topology. But it also produced more verbose, more brittle, and more structurally distorted outputs, which explains the simultaneous rise in first-place and last-place judgments.

5.4 Cross-Run Judge Drift

A second caution follows from cross-run drift. Compared with Phase-0, this four-arm run showed drift in major hallucination rates even for the same previously trained arms: Base moved by +1 percentage point, Group A by +4 percentage points, and Group B by +8 percentage points. Cross-run absolute comparisons are therefore invalid. Within-run comparisons are the only defensible comparisons in this paper. That finding itself reinforces the E&D framing: the evaluation setup is not neutral background noise but part of the scientific object under study [18].

6 Failure Mode Analysis

Gen-1 obeyed the target template **100% of the time**. It always emitted the `Reasoning: ... Avoidance rule: ... Answer:` structure, regardless of whether the prompt warranted that much corrective scaffolding. That fact holds in both the first-place and the last-place cases. The model learned self-correction as a compulsory speech act.

6.1 Complexity-Tier Stratification

Stratifying by prompt complexity (using `topology.complexity_score` binned into tiers) reveals the mechanism cleanly:

Table 8: Gen-1 performance by complexity tier.

Tier	Threshold	N	Major H.	Minor H.	Struct.	1st pl.	Last pl.
Easy	≤ 0.33	18	22.2%	16.7%	3.11	3	**11**
Medium	0.33–0.66	60	25.0%	33.3%	2.90	19	27
Hard	> 0.66	22	**4.5%**	40.9%	2.82	**15**	2

Gen-1 wins 68% of first-place votes on hard prompts (15/22) **and loses 61% of last-place votes on easy prompts** (11/18). On hard prompts—the regime the corpus was constructed to encode—Gen-1's major hallucination rate falls to 4.5%, an order of magnitude better than on easy prompts (22.2%). The inverted-by-complexity pattern is the paper's strongest quantitative evidence for format compulsion: when the prompt's intrinsic complexity matches the error topologies Gen-1 was trained on, the fused-format template is appropriate and Gen-1 excels; when the prompt is simple and does not require the structured output, Gen-1 compulsively emits it anyway, fabricating avoidance rules where none are needed.

6.2 Per-Error-Class Stratification

Disaggregating by `topology.error_class`:

Table 9: Gen-1 performance by error class.

Error Class	N	Major H.	Minor H.	Struct.	Validity	1st	Last
`cultural_contextual`	12	**8.3%**	25.0%	3.00	3.50	8	3
`runtime_semantic_code`	27	11.1%	33.3%	2.70	2.85	15	5
`math_logic_fallacy`	25	28.0%	24.0%	3.16	3.16	4	11
`factual_hallucination`	20	**30.0%**	**40.0%**	3.05	3.00	5	13
`visual_perceptual`	11	27.3%	45.5%	2.64	2.36	2	8
`multi_step_planning`	4	0.0%	25.0%	2.75	3.50	3	0
`compiler_syntax`	1	0.0%	0.0%	3.00	4.00	0	0

Patterns:

- **`cultural_contextual` and `runtime_semantic_code` are Gen-1's strongest classes.** Cultural prompts reward structured step-by-step reasoning; code prompts benefit from explicit error analysis.
- **`factual_hallucination` is Gen-1's worst class** (30% major, 40% minor). The format compulsion makes this class actively worse: forcing the model to emit a chain-of-thought plus avoidance rule provides more surface area for fabrication. This aligns with prior findings [9] that unfamiliar fine-tuning examples shape hallucination behavior stylistically without improving factual grounding.
- **`visual_perceptual` regression** (27% major, 45% minor) is expected given text-only training on a vision-architecture base; this is primarily a confound of the modality-shrink protocol.

6.3 Mechanistic Hypothesis

Because the target format always fused validated reasoning, a declarative avoidance rule, and a final answer into the same assistant continuation, naive SFT trained the *emission pattern* more robustly than the policy governing when that pattern should appear. Predicting the string `Reasoning:` followed by `Avoidance rule:` becomes a high-probability attractor state. The model is implicitly penalized during training if it skips straight to the answer, even when the answer is trivial. When a query did not warrant an avoidance rule, the model simply confabulated one. This spurious rationale generation directly drives the +12pp spike in minor hallucinations and the −0.98 drop in structural quality.

6.4 Evidentiary Support from Judge Rationales

Judge rationales on losing Gen-1 cases identified the same mechanism repeatedly:

> *"gen1 arrives at the same answer but through confused, self-contradictory reasoning that invents spurious 'avoidance rules'."*
>
> *"gen1 adds unnecessary verbose pseudo-reasoning with a dubious 'salience' claim that introduces minor hallucination."*
>
> *"gen1 correctly computes the denominator as 4/3 but then asserts the answer is 3/2 rather than 3/4."*

The final quote is particularly telling: the model produced the wrong final answer despite this exact prompt existing in its training set. The cognitive load of generating the fabricated avoidance rule appears to corrupt the final answer prediction.

This interpretation is consistent with SCoRe's warning that offline SFT on correction traces can collapse into ineffective correction behavior [10], with Zhu et al. [23]'s finding that rationale-like supervision is not uniformly beneficial, and with Kang et al. [9]'s demonstration that fine-tuning examples shape hallucination behavior. The added lesson here is more specific: even when the model sees *exact* prompts from training, fused-format corrective supervision can still overfit the *form* of self-correction and underfit its *judgment*. The student learns exactly how to apologize and structure a correction, but lacks the world model to understand why the correction is being applied.

7 Discussion

The result is neither a validation of out-of-domain self-correction nor a refutation of the Density Imperative. A better reading is sharper and less flattering: archived failures can indeed be organized into a scientifically valuable supervision corpus, but naive fused-format SFT over that corpus teaches a *style* of correction more readily than the *substance* and *deployment policy* of correction. Under the E&D framing, this is not a failed paper. It is a more useful one. It contributes a released artifact, a cheap continuation recipe, and a diagnosed anti-pattern whose existence should inform future evaluative practice.

What held. The foundational concept of the dataset—structuring failures through causality DAGs, procedural classifications, and backdoor adjustment formulas—remains a potent method for organizing failure states. The 37 first-place preference votes, concentrated in hard and medium-complexity prompts, demonstrate that when the fused format aligns with prompt complexity, the resulting self-correction is unparalleled in quality. The Path-A frozen-vision continuation protocol

successfully updated text-based reasoning networks without inducing modality collapse, proving the viability of parameter-efficient methodology for multimodal reasoning refinement.

What did not hold. The hypothesis that offline SFT on concatenated reasoning chains unilaterally improves reliability is invalidated. The spike in minor hallucinations and the format compulsion anti-pattern reveal that dense process data lacks the necessary gating mechanism when applied via standard cross-entropy loss. Where dense spatial-semantic curation (Group B) acts as reliable architectural scaffolding, dense corrective curation under naive SFT acts as an overactive reflex.

7.1 Null-Result Posture

The scientific validity of this report does not hinge on achieving state-of-the-art inference metrics. Under the current E&D call, submissions need not introduce a new model or outperform prior work, and explicitly negative results, critical analyses, and artifacts that improve evaluation are in scope [18]. Even under a strict null result on H1, the contributions remain publishable. The CC-BY-4.0 Reasoning-NEST v2 corpus provides the community with a much-needed asset for analyzing error topologies. The formalization of the evaluation protocol establishes a reproducible standard for four-arm ablations. Most importantly, the explicit documentation of the format compulsion failure mode and the data leak disciplines the design of future experiments.

8 Limitations

Table 10: Likely reviewer objections and this paper's response.

Likely objection	**Response**
Data leak invalidates Gen-1 topline	Agreed. The leak is disclosed prominently, quantified at 100%, and blocks all OOD claims.
Small N makes the result underpowered	Agreed. The pilot is treated as a screening benchmark only; a locked 1,000-prompt OOD confirmatory benchmark is future work.
Judge bias is possible because Claude appears in both corpus construction and evaluation	Agreed. Reported explicitly. Strongest comparisons are limited to within-run, blinded evaluation with absolute rubrics.
Text-only training on a vision base is a modality confound	Deliberate and controlled. Vision layers are frozen so the language-side continuation problem is isolated.
Three-teacher corpus generation is expensive to reproduce	True. Auditability is addressed through released schema, hashes, pilot data, and prompt templates rather than through claims of cheap regeneration.
Group B (5,000 records) and Gen-1 (459 records) are scale-mismatched	True. Record counts are not the right unit; confirmatory work should report token budgets explicitly.
Rationale-based methods are not reliably helpful	Agreed. The paper aligns with that caution and documents a concrete anti-pattern rather than overclaiming rationale benefits.

9 Future Work

1. **Locked 1,000-prompt out-of-domain confirmatory benchmark.** The immediate priority. Exact prompt hashes should be checked before training begins, not after results are obtained. The benchmark should be run once, under a frozen judge version, with within-run comparisons pre-specified. Subsets should tilt toward verifier-grounded ground truth (SWE-bench Verified,

MATH L5 with symbolic solvers, MIRAGE-Bench rubrics) at 30–40% minimum to anchor the strongest claims on automated verification rather than LLM judgment.

2. **Gen-1.5 with ORPO or a comparable preference-aware objective.** The present paper's failure mode strongly suggests that unconditional fused-format SFT is too weak an objective for learning when corrective scaffolds should fire. Using the flawed, original outputs as negative signals and the corrected NEST pathways as positive signals, the model can learn to abstain from applying avoidance rules when unwarranted—optimizing for preference rather than simple sequence completion.

3. **Gen-1-continue from a Group B merged base.** In the present study, Gen-1 starts from the original base pathway. A more principled test of "dense CN baseline plus failure-topology continuation" should continue from the best prior dense-supervision state rather than starting over. This requires a one-time offline merge of the Group B adapter into a 16-bit dense checkpoint on hardware larger than the L4.

4. **Verifier-grounded subsets.** Records whose corrective path can be externally checked more rigidly are likely to be better substrates for learning disciplined self-correction than records whose main signal is stylistic reasoning prose.

5. **Non-oracular metadata routing.** Hard-slice prompting with topology labels is scientifically meaningful only if those labels are predicted at test time rather than leaked from an oracle. A separately trained router must precede any confirmatory H4 claim.

6. **Active confidence-format and calibration endpoints.** Future models must emit parseable confidence so Brier score, ECE, and semantic-uncertainty diagnostics can test whether self-correction improves calibration rather than merely changing style.

10 Conclusion

Reasoning-NEST v2 contributes a provenance-rich post-failure corrective corpus and reframes those records as evaluative artifacts rather than mere training examples. The Gen-1 continuation experiment is scientifically useful not because it proves self-correction generalization, but because it does **not**. Under exact prompt leakage, naive fused-format SFT improved one severe-error metric while degrading output cleanliness, structure, and reasoning quality, and while becoming the most polarizing arm in the study. Stratification by prompt complexity reveals the mechanism cleanly: Gen-1 wins on hard prompts where structured reasoning is appropriate and loses on easy prompts where direct answers are preferred. The central lesson is methodological: self-correction traces are not automatically beneficial, and archived failures become most valuable when they are documented richly enough to expose the conditions under which a continuation recipe fails.

References

[1] Akhtar, M., Benjelloun, O., Conforti, C., Foschini, L., Gijsbers, P., Giner-Miguelez, J., Goswami, S., Jain, N., Karamousadakis, M., Krishna, S., Kuchnik, M., Lesage, S., Lhoest, Q., Marcenac, P., Maskey, M., Mattson, P., Oala, L., Oderinwale, H., Ruyssen, P., ... Zhang, L. (2024). *Croissant: A metadata format for ML-ready datasets.* Advances in Neural Information Processing Systems, 37.

[2] Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., ... Kaplan, J. (2022). *Constitutional AI: Harmlessness from AI feedback.* arXiv preprint 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. *Advances in Neural Information Processing Systems*, 36, 10088–10115.

[5] Dong, B., Ni, M., Huang, Z., Yang, G., Zuo, W., & Zhang, L. (2025). *MIRAGE: Assessing hallucination in multimodal reasoning chains of MLLM.* OpenReview.

[6] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92.

[7] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 1321–1330). PMLR.

[8] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-rank adaptation of large language models.* arXiv preprint arXiv:2106.09685.

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

[10] Kumar, A., Zhuang, V., Agarwal, R., Su, Y., Co-Reyes, J. D., Singh, A., ... Faust, A. (2025). Training language models to self-correct via reinforcement learning. *International Conference on Learning Representations.*

[11] Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., & Cobbe, K. (2024). Let's verify step by step. *International Conference on Learning Representations.*

[12] MacPherson, T. (2026a). *Cognitive Nutrition: Architecture of a provenance-embedded data refinery.* Zenodo. https://doi.org/10.5281/zenodo.18667742

[13] MacPherson, T. (2026b). *The Density Imperative: How semantic curation depth determines vision-language model capability.* Zenodo. https://doi.org/10.5281/zenodo.18667735

[14] MacPherson, T. (2026c). *The Entropy of Recursion: A strategic framework for provenance-verified data and the preservation of signal integrity in generative systems.* Zenodo. https://doi.org/10.5281/zenodo.18436975

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

[16] Meta. (2024). *Llama 3.2 community license agreement.* Meta Llama documentation.

[17] NeurIPS. (2026a, March 23). *Introducing the Evaluations & Datasets Track at NeurIPS 2026.* NeurIPS Blog.

[18] NeurIPS. (2026b). *NeurIPS 2026 Evaluations & Datasets Track call for papers.* NeurIPS.

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

[20] Zelikman, E., Wu, Y., Mu, J., & Goodman, N. (2022). STaR: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35.

[21] Zelikman, E., Harik, G., Shao, Y., Jayasiri, V., Haber, N., & Goodman, N. D. (2024). *Quiet-STaR: Language models can teach themselves to think before speaking.* arXiv preprint arXiv:2403.09629.

[22] Zhang, W., Sun, Y., Huang, P., Pu, J., Lin, H., & Song, D. (2025). *MIRAGE-Bench: LLM agent is hallucinating and where to find them.* OpenReview preprint.

[23] Zhu, C., Xu, B., Yang, A., Lin, J., Wang, Q., Zhou, C., & Mao, Z. (2025). Rationales are not silver bullets: Measuring the impact of rationales on model performance and reliability. In *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 5808–5835).

A Three-Teacher Prompt Templates and Admission Criteria

The archived run used service-specific wrappers and vendor-specific system prompts. Because those exact wrappers are not reproduced here, the templates below preserve the operative structure rather than verbatim transport syntax. This is intentional and sufficient for audit.

A.1 Primary Oracle Template

```
[SYSTEM]
You are the primary correction oracle for a post-failure
reasoning corpus. Your task is to convert a failed model
response into a structured corrective artifact.

[INPUT]
Original prompt: {event.prompt}
Failed output:   {event.failed_output}

[REQUIRED OUTPUT]
1.  event.correction
2.  reasoning.cot_trace as an ordered array of steps
    - each step must include:
      - step_text
      - is_valid: true/false
      - operation classifier
3.  reasoning.failure_narrative
4.  reasoning.failure_mechanism
5.  causality.graph as a DAG specification
6.  causality.adjustment_formula
7.  topology.error_class  (seven-class enum in Section 3.4)
8.  topology.avoidance_rule
9.  topology.complexity_score in [0,1]
10. provenance fields needed for downstream arbitration

[CONSTRAINTS]
- Prefer verified correction over stylistic fluency
- Do not preserve invalid reasoning as if correct
```

```
- Name the failure mechanism explicitly
- Produce schema-valid JSON only
```

A.2 Secondary Oracle Template

```
[SYSTEM]
You are the independent verification oracle.
Review the primary artifact as if it were a candidate
benchmark record.

[INPUT]
Original prompt: {event.prompt}
Failed output:   {event.failed_output}
Primary artifact: {primary_json}

[TASK]
- Verify or amend the correction
- Challenge unsupported steps
- Flag misclassified failure mechanisms
- Repair causal graph or adjustment formula if needed
- Confirm or replace topology label and avoidance rule

[OUTPUT]
A full revised artifact plus a structured disagreement report.
```

A.3 Judge and Arbitrator Template

```
[SYSTEM]
You are the adjudicator for corpus admission.

[INPUT]
Original prompt: {event.prompt}
Failed output:   {event.failed_output}
Primary artifact: {primary_json}
Secondary artifact: {secondary_json}

[TASK]
- Compare the two candidate artifacts
- Resolve disagreements
- Assign an arbitration score
- Produce one final admitted artifact if standards met;
  otherwise reject

[OUTPUT]
{
  "admit": true/false,
  "final_artifact": {...},
  "arbitration_score": ...,
  "notes": ...
}
```

A.4 Admission Criteria

A record was admitted only if all of the following held: the correction was judged materially valid; the failure mechanism could be named coherently; the record was structurally complete under schema validation; the causal and topology fields were not obviously contradictory; provenance metadata were attached; and the arbitration outcome was strong enough to justify inclusion. Exact numeric admission thresholds are not reproduced here because the run used internal service-level scoring policies; the public artifact exposes the provenance fields needed for audit.

B Cross-Lab Mixture-of-Experts Consultation Methodology

This experiment was designed using a cross-lab Mixture-of-Experts consultation methodology. The same cold architectural prompt was dispatched independently to four frontier models, each receiving a densely structured context packet in turn 1. We refer to this as **Data Cocktail priming**. The outputs were then diffed and synthesized before any experiment was executed.

Table 11: Cross-lab consultation and prevented failures.

Model consulted	**Independent contribution**	**Prevented failure**
Gemini Pro DR	Flagged modality-collapse risk in an earlier design and surfaced a release-path licensing concern	Avoided a corrupted multimodal workaround and forced license review before public-release decisions
Gemini Thinking DR	Added convergent support on continuation framing and failure-slice decomposition	Increased design redundancy; no unique blocker singled out retrospectively
GPT-5.4 DR	Surfaced the official track rename and the paper's power insufficiency	Reframed the paper toward E&D methodology and away from overclaiming from the 100-prompt screen
Grok non-DR	Confirmed the canonical continuation pattern in the tooling stack	Reinforced the choice of Path A over ad hoc multimodal workarounds

Two of these prevented failures are externally verifiable. First, the track rename and expanded E&D framing were published by NeurIPS in March 2026. Second, the base model family is governed by a custom community license rather than CC-BY terms, precisely the kind of release-path issue that should be surfaced *before* public packaging is finalized [16–18].

The broader methodological claim is modest but important: independent model disagreement can act as a cheap pre-execution peer review layer. In this case, the diff across models did not merely decorate the design—it prevented concrete mistakes.

C Additional Observations and Hyperparameter Sensitivities

C.1 Hyperparameter Sensitivity Notes

No formal hyperparameter sweep was executed in this study. The notes below document *load-bearing choices*, not quantitative sensitivity curves.

Table 12: Load-bearing implementation notes.

Choice	Note
`use_gradient_checkpointing="unsloth"`	Load-bearing for fitting the run comfortably within L4 VRAM. Standard PyTorch checkpointing produces immediate OOM at effective batch 8.
Path A over Path B	Chosen because it preserved the documented multimodal continuation route without offline merge complications on L4 hardware.
Path C sentinel image	Rejected during design because it corrupted continuation behavior.
Valid-only reasoning steps	Necessary to avoid explicitly teaching the failed reasoning path under cross-entropy loss.
Plain natural-language section headers	Deliberately used instead of XML-delimited reasoning tags to avoid tokenizer special-token requirements.
Three epochs / 171 steps	Adequate for an initial screen; not evidence that the objective had converged to its best attainable behavior.

The main sensitivity conclusion is procedural rather than numeric: for this class of corpus, objective design is more likely to dominate outcome quality than small local changes in learning rate or adapter rank. The identified failure mode is not "slightly under-tuned." It is structurally tied to the naive fused-format target.

C.2 Judge Rationale Excerpts

The following excerpts are representative of why Gen-1 was rejected even when it appeared superficially self-corrective:

> *"gen1 arrives at the same answer but through confused, self-contradictory reasoning that invents spurious 'avoidance rules'."*
>
> *"gen1 adds unnecessary verbose pseudo-reasoning with a dubious 'salience' claim that introduces minor hallucination."*
>
> *"gen1 correctly computes the denominator as 4/3 but then asserts the answer is 3/2 rather than 3/4."*

These excerpts are more informative than a topline score because they identify the mechanism of loss directly.

C.3 Detailed Per-Slice Results

See Section 8 (complexity-tier stratification) and Section 9 (per-error-class stratification) in the main body for the full quantitative slice analyses. Summary: Gen-1's behavior is bimodal by prompt complexity (68% first-place win rate on hard prompts vs. 61% last-place loss rate on easy prompts) and asymmetric by error class (`cultural_contextual` and `runtime_semantic_code` are Gen-1's strongest classes; `factual_hallucination` is the worst).