The Density Imperative, Refined: A 2×2 Ablation Reveals Independent Sufficiency of Density and Structure in Vision-Language Fine-Tuning

Anonymous authors
Paper under double-blind review

Abstract

Prior work (Anonymous, 2026b) reported that fine-tuning data quality is *signed*: sparse ∼50-token captions catastrophically degrade a vision-language base model, while dense 111-field structured Neural Extraction of Semantic Topology (NEST) metadata enhances it. This paper reports a 2×2 ablation extension that decomposes that finding mechanistically. The original two-cell comparison conflated metadata density (tokens per record) with metadata structure (whether the supervision signal carries an explicit analytical schema). This paper closes the design with two new arms: **Group C** (dense but unstructured) and **Group D** (sparse but structured). Both novel arms are evaluated on the same Llama 3.2 11B Vision-Instruct base, the same 9,081-record Alexandria Aeternum training set, the same LoRA-32 hyperparameter envelope, and the same 100-image held-out CogBench protocol used in the original work. The result is sharper than the original framing: **density and structure are *independently sufficient*** to recover cognitive depth from the catastrophic Group A regime. Paired Wilcoxon contrasts at $n = 100$ on the primary CogBench metric show Group C (dense, unstructured) ≈ Group B (dense, structured) at $\Delta = +0.008$, $p = 0.20$; Group D (sparse, structured) ≈ Group B at $\Delta = -0.009$, $p = 0.24$. Either density or structure alone matches the original "dense+structured" regime on cognitive depth; neither is required when the other is present. On secondary metrics, Group D additionally wins: **Group D vs Group B on capability retention (CCR)**: $\Delta = +0.072$, $p = 0.004$; **Group D vs Base on output information density**: $\Delta = +0.297$, $p = 3.9 \times 10^{-18}$. Sparse data made structurally rich produces a measurable cognitive gain over the unmodified base, preserves base capability better than the dense-structured regime, and dramatically increases output information density. We additionally acknowledge an out-of-distribution scope refinement reported in companion work (Anonymous, 2026e): on a 953-prompt sequestered OOD reasoning benchmark, the original 11.1× density advantage between Groups A and B produces zero measurable benefit on verifier-grounded accuracy or 4-judge Borda. The original in-domain claim is not falsified; its extrapolation to out-of-distribution reasoning is. The operational implication is concrete: **do not train on sparse unstructured data**; structure-rich sparse data is the cost-efficient operational regime where the base model already contains the underlying knowledge. We hypothesize that Group D operates as an *elicitation* regime—reorganizing existing pre-trained associations into reasoning-accessible form rather than teaching new ones—and identify the falsifiable test that would bound this hypothesis (Group D on a domain the base model has not seen, future work). All adapters, evaluation

outputs, raw JSON aggregates, and a reproducibility harness are released for community verification.

Keywords: fine-tuning data quality, vision-language models, semantic density, semantic structure, structured supervision, capability retention, ablation study, data-centric AI

1 Introduction

The original Density Imperative argued that fine-tuning data quality is *signed*: sparse, low-token-count supervision actively degrades pre-trained vision-language model capabilities, while dense, structured supervision enhances them. The argument rested on a controlled two-cell comparison—Group A (sparse ∼50-token captions) versus Group B (dense 111-field NEST schema, ∼2,000–4,000 tokens per record)—fine-tuned on the same 9,081 cultural-heritage images with otherwise identical hyperparameters. On a 100-image held-out CogBench evaluation, Group A scored 0.176 against an unmodified base of 0.388 (a 54% capability collapse) while Group B scored 0.422 (a 9% improvement over base, with a 24-percentage-point swing between the two fine-tuned arms). The finding was statistically robust at $p < .001$ via Friedman tests with Wilcoxon post-hoc and Holm–Bonferroni correction.

The two-cell design was diagnostically useful but mechanistically incomplete. Sparse-versus-dense and unstructured-versus-structured were entangled in the original A-versus-B comparison: Group A was *both* sparse *and* unstructured, Group B was *both* dense *and* structured. The 24-point swing measured the joint effect of changing both axes simultaneously. The paper acknowledged this confound explicitly in its limitations section and proposed the missing diagonal cells—Group C (dense but unstructured) and Group D (sparse but structured)—as future work.

This paper reports those experiments. The result is sharper than the original framing.

When the diagonal is closed, the picture changes substantively. Density and structure are not co-required; *either is sufficient*. The catastrophic A regime is uniquely catastrophic: only the joint absence of density *and* structure produces the 54% capability collapse. Adding either dimension—more tokens, or an explicit analytical schema, or both—rescues the model into roughly the same cognitive-depth band. The strongest single-axis claim that survives the new data is *not* the density imperative as originally stated, but rather a statement about the catastrophic floor: training on sparse, unstructured supervision degrades a vision-language model below its un-fine-tuned baseline by every measure we can compute on this corpus.

We additionally report a complementary out-of-distribution finding from companion work (Anonymous, 2026e) that further refines the original paper's scope. On a 953-prompt sequestered out-of-distribution reasoning benchmark (MATH-500, HumanEval+, MMLU-Pro, BigCodeBench, hand-curated OOD prompts), the 11.1× density advantage between Groups A and B produces *zero* measurable benefit on either verifier-grounded accuracy (32.9% vs. 32.9%) or a 4-judge Borda aggregate (0.566 vs. 0.575). Density transfers in-domain but does not detectably transfer out-of-distribution at this corpus scale and base size. We treat this as a scope refinement, not a refutation: the in-domain claim of the original work survives at the corpus and benchmarks where it was originally measured; the implicit extrapolation to general reasoning capability does not.

The new in-domain 2×2 also surfaces a finding the original work could not: the sparse-but-structured arm (Group D) is the cleanest *operational* regime when the base model already contains the underlying knowledge. It matches Group B on cognitive depth at lower data cost (the D records

average ∼50–80 tokens per record against B's ∼2,000–4,000, an 11.1× density gap by per-record assistant-message token count). On the s42 adapter it additionally beats Group B on capability retention (CCR; Group D s42 0.91, Group B 0.84, paired Wilcoxon $\Delta = +0.072$, $p = 0.004$, $n = 94$). And it produces dramatically denser output (information density 0.66, 81% above base; $p = 3.9 \times 10^{-18}$, replicated across two seeds). Where Group B's dense fine-tuning erodes 15% of base CCR and Group A erodes 48%, Group D's CCR erosion is seed-fragile: 8% on s42 and 20% on s1337 (see §6 Limitation 4). For practitioners with sparse but structurable corpora—internal documentation, compliance archives, technical manuals, scaffolded-reasoning training data—the operational prescription is straightforward: structure the corpus you have, do not generate density you do not need.

1.1 Contributions

1. **The catastrophic floor (preserved).** Group A (sparse ∼50-token, unstructured) reduces CogBench by 54% from base, increases hallucination rate by 330%, and collapses counterfactual reasoning by 51%. The original observation that sparse, unstructured fine-tuning is uniquely destructive is unchanged; it is the strongest single finding in the program.

2. 2×2 **ablation: density and structure are independently sufficient.** Closing the design with Group C (dense, unstructured) and Group D (sparse, structured) reveals that either axis alone matches Group B (dense, structured) on the primary in-domain CogBench metric. The original 'density imperative' finding is correct in sign but underspecified in mechanism: it is not density that is required, and it is not structure that is required—it is the avoidance of the joint A regime.

3. **Group D is the cost-efficient operational winner.** The sparse-structured arm matches dense-structured on cognitive depth, beats it on capability retention, and produces 81% denser output than the un-fine-tuned base—all at 1× rather than 11.1× per-record token volume.

4. **Mechanistic hypothesis: D as elicitation.** The qualitative profile of Group D—real but modest cognitive gain over base (∼75% of B's gain on CogBench), minimal weight perturbation as measured by CCR (gentlest non-trivial fine-tune), and small-but-significant rather than dramatic expansion in semantic coverage—is consistent with structural scaffolding *eliciting* pre-trained knowledge rather than teaching new associations. We identify but do not test this hypothesis in the present paper. We pre-register the falsifiable elicitation-bound experiment as future work (§7): if the elicitation hypothesis is correct, Group D should fail on a domain the base model has not seen during pre-training.

5. **Out-of-distribution scope refinement (companion work).** Density's effect on out-of-distribution reasoning is not measurable on a 953-prompt sequestered benchmark at this corpus scale (Anonymous, 2026e). The original in-domain claim is not falsified; its extrapolation is. We import this finding into the present paper because it is methodologically inseparable from the in-domain refinement.

6. **Full reproducibility package.** Four LoRA adapters (Groups A, B, C × 2 seeds, D × 2 seeds), the Alexandria Aeternum evaluation set, all evaluation scripts, raw model outputs, paired Wilcoxon contrast scripts, and an interactive comparison tool are released for community verification.

2 Related Work

2.1 Data Quality in Model Training

The data-centric AI movement has established that dataset engineering—holding architecture and compute fixed—materially affects downstream performance. DataComp (Gadre et al., 2023) demonstrated that data curation alone yields substantial gains across dozens of downstream tasks. RefinedWeb (Penedo et al., 2023) showed that filtered web data can outperform curated mixtures. DoReMi (Xie et al., 2023) showed that optimizing data mixture weights improves training efficiency. FineWeb (HuggingFace Team, 2024) reinforced the value of systematic curation pipelines.

These studies focus on *pre-training* data and treat quality as a scalar: higher quality $\rightarrow$ better model. Our original paper extended this framing to *fine-tuning* data and demonstrated that the quality scale is *signed*: low-density fine-tuning data can produce outcomes worse than no fine-tuning at all. The present paper refines that claim further: it is not density per se that is signed, but the joint absence of density *and* structure that is uniquely destructive.

2.2 Model Collapse and Synthetic Data

Shumailov et al. (2024) established that training on recursively generated synthetic data causes irreversible distributional narrowing. Dohmatob et al. (2025) identified a generalization-to-memorization phase transition in late-stage autophagy. Alemohammad et al. (2024) demonstrated the "MAD" phenomenon in self-consuming generative models.

Our Group A results document an analogous phenomenon under *supervised* fine-tuning: sparse, unstructured supervision induces a form of cognitive collapse in which the model's pre-trained analytical capabilities are overwritten by low-entropy response patterns. The new 2×2 design lets us localize this effect more precisely than the original paper could: the collapse is not present when the supervision signal carries either dense token volume *or* an explicit analytical schema. The collapse requires the joint absence of both.

2.3 LIMA and Data Efficiency

Zhou et al. (2023) demonstrated that as few as 1,000 carefully curated examples can align a language model, arguing that "less is more for alignment." Our 2×2 findings complicate this elegantly: Group D uses sparse $1\times$-density supervision but adds structure, and recovers the cognitive-depth gain at the same example count as Group B's $11.1\times$-density structured supervision. LIMA's result may partly reflect not the small example count alone, but the implicit structural coherence of carefully curated examples—a hypothesis the present paper's Group A versus Group D contrast is consistent with.

2.4 The Underlying Provenance-Binding Framework

Prior work (Anonymous, 2026a) introduces the theoretical framework motivating the present study: a protocol for binding dense semantic context to media artifacts via persistent perceptual hash identifiers. The Alexandria Aeternum dataset used in this experiment was produced by the metadata factory described in that prior work, which employs orchestrated multi-model inference to generate the 111-field structured schema we refer to as NEST throughout this paper. Anonymous (2026c) formalizes the data-refinery architecture that produced both the Group B (dense + structured)

and Group D (sparse + structured) training corpora used here. Anonymous (2026d) extends the framework to autonomous-robotics applications.

A companion submission (Anonymous, 2026e) reports an out-of-distribution evaluation that complements the present in-domain refinement; it is cited as the source for the OOD scope refinement enumerated as Contribution 5 above.

2.5 Evaluation Frameworks

Our evaluation draws on recent benchmarks for vision-language model assessment:

- **CogBench** (Song et al., 2025): Evaluates cognitive depth across narrative understanding, mental state attribution, focal point analysis, and causal reasoning.
- **VULCA-Bench** (Yu et al., 2025): A five-level visual understanding hierarchy from perceptual identification (L1) through philosophical interpretation (L5).
- **POPE** (Li et al., 2023): Polling-based object probing for hallucination detection via binary yes/no questions about object presence.
- **CounterVQA** (Chen et al., 2025): Counterfactual visual question answering measuring a model's resistance to false premises.

3 Methodology

3.1 Dataset: Alexandria Aeternum

The Alexandria Aeternum Genesis Collection consists of 10,090 public-domain artworks sourced from The Metropolitan Museum of Art's Open Access Initiative, spanning 8 curatorial departments, millennia of temporal range, and global geographic coverage. Each artifact carries a structured provenance metadata document containing 111+ fields across four ontology layers: provenance, semantic topology, affective resonance, and rights.

The dataset was partitioned 90/10 into training (n = 9,081) and evaluation (n = 1,009) splits. From the evaluation split, 100 images were randomly sampled for the benchmark suite reported here, with 909 images held in reserve for community replication. The same train/eval split was used across all four fine-tuning conditions; no NEST metadata for the evaluation images was provided to any condition during training.

3.2 Experimental Conditions: The 2×2 Design

We define five conditions: an un-fine-tuned baseline plus the four cells of the density $\times$ structure 2×2.

Condition	Label	Density	Structure	Metadata Content
Base	No fine-tuning	—	—	Llama 3.2 11B Vision-Instruct, unmodified
Group A	Sparse, unstr.	low (∼50 tok)	none	Title, creator, date, medium (simulates web alt-text)
Group B	Dense, NEST	high (∼2–4k tok)	111-field	Composition, technique, cultural context, affect, symbolism, narrative, historical period
Group C	Dense, unstr.	high (∼2–4k tok)	none	Long-form prose paragraph of equivalent token length, free-form narrative description
Group D	Sparse, struct.	low (∼50–80 tok)	111-field (compressed)	Compressed structural schema headers with ≤1-clause field values, preserving the analytical scaffold at 1× density

Critical design feature. Groups A, B, C, and D train on *identical images* with *identical model architecture and hyperparameters*. The sole experimental variables are textual density (low vs. high) and textual structure (absent vs. present). Any performance difference is therefore attributable to the density×structure factorial. Groups C and D are the new diagonal cells closing the 2×2 design proposed as future work in the original paper.

	No structure	**111-field schema**
High density (∼2–4k tok)	**Group C** CogBench **0.429** *Dense paraphrase* matches B in-domain	**Group B** CogBench **0.422** *Original NEST* in-domain ceiling
Low density (∼50–80 tok)	**Group A** CogBench **0.176** *Catastrophic floor* (−54.0% vs. Base)	**Group D** CogBench **0.409** *Efficient elicitation* wins CCR + info density

Figure 1: The density × structure 2×2 at a glance. Base CogBench: 0.378. Each cell shows the experimental group, its primary CogBench score (mean across replication seeds where applicable), and the cell's qualitative role. The destructive −54.0% regime is isolated to Group A (sparse *and* unstructured); either density or structure alone is independently sufficient to recover cognitive depth.

3.3 Group C Construction: Dense but Unstructured

Group C records were generated from each corresponding Group B NEST record via a deterministic transformation pipeline: a single Gemini 2.5 Pro prompt instructed the model to produce a free-form prose paragraph paraphrasing the same NEST content in narrative form, without enumerating the 111 schema field names, without bulleted or numbered lists, and without JSON-like syntax. Token-length matching against Group B was enforced via a post-processing pass (Llama-3.2 tokenizer; target ±10% per record). Generation was conducted once at temperature 0.0 to maximize reproducibility (the same Group B input deterministically produces the same Group C output), and the resulting Group C corpus is held identical across both replication seeds (only the LoRA initialization seed varied between s42 and s1337 fine-tunes). The construction script and per-record token-length audit are released in the supplementary package. The result is dense supervision (high token volume) without the explicit analytical schema that Group B's metadata carries.

3.4 Group D Construction: Sparse but Structured

Group D records were generated from each corresponding Group B NEST record via a deterministic Python compression script (no LLM in the loop): the script preserves the 111-field schema headers and hierarchical grouping, extracts the most concise valid value per field (typically a single noun phrase or 1–3-word predicate), and emits the resulting compressed structure at approximately 50–80 tokens per record (Llama-3.2 tokenizer; identical token-budget envelope to Group A). The structural schema is preserved (named fields, hierarchical groupings, JSON-LD-compatible serialization) but the per-field content is reduced to the most concise valid statement. As with Group C, the Group D corpus is held identical across both replication seeds (only the LoRA initialization seed varied). The compression script is released in the supplementary package. This isolates structure from density: the analytical scaffold is present, the token volume is sparse.

To make the format concrete, the same record presented to Groups A, C, and D differs as follows. **Group A** (sparse, unstructured): *"Madonna and Child Enthroned with Saints, by Raphael, c. 1505, oil on panel."* **Group C** (dense, unstructured): a ∼3,000-token narrative paragraph paraphrasing the full NEST analysis in free prose. **Group D** (sparse, structured): `{"technique": "oil on panel", "composition": "centered triangular", "affect": "devotional stillness", "cultural_context": "High Renaissance Florentine", "symbolism": "iconographic enthronement", ...}` — the schema fields and hierarchy are preserved verbatim from the Group B NEST record, but each field carries a single noun phrase rather than the multi-clause analytical paragraph Group B contains. Group D records remain valid JSON-LD throughout; they are short *not* because fields are dropped but because each field's value is compressed to its most concise valid form.

3.5 Model and Training Configuration

All conditions share the configuration of the original paper:

Parameter	**Value**
Base model	Llama 3.2 11B Vision-Instruct
Quantization	4-bit (Unsloth)
Framework	Unsloth 2026.2.1 + TRL SFTTrainer
LoRA rank (r)	32
LoRA alpha	32
Target modules	q, k, v, o, gate, up, down
Vision encoder	Frozen (no gradient updates)
Trainable parameters	∼15–20M
Effective batch size	8 (batch 1 × gradient accumulation 8)
Learning rate	2×10^{-4}
Optimizer	AdamW (8-bit)
Precision	bfloat16
Training steps	1,135 (≈1 epoch)
Image resolution	560px max (1 vision tile)
Hardware	NVIDIA L4 (24GB VRAM)

Groups C and D were trained at two seeds each (s42 and s1337) for confirmatory replication; Groups A and B retain their single-seed configuration from the original paper. The vision encoder is frozen throughout training, meaning all adaptation occurs in the language model's attention and

feed-forward layers via LoRA. This is a deliberate design choice: it ensures that any performance changes reflect learned reasoning patterns, not altered visual feature extraction.

3.6 Evaluation Framework

We evaluate on the same four-track suite used in the original paper, totaling ∼4,400 question–answer pairs per condition:

- **Track 1** — Domain Competency (VULCA-Bench five-level hierarchy)
- **Track 2** — Hallucination Resistance (POPE-style probes)
- **Track 3** — Response Quality (token count, refusal rate)
- **Track 4** — Structural Generalization: CogBench cognitive depth (8 dimensions per image), counterfactual resistance (CCR + explanation quality), semantic density alignment (6-field coverage + information density)

Operational definition of the secondary metrics.

- **CCR** (Counterfactual Consistency Rate, Track 4.2): for each evaluation image, the model is presented with 2–3 false counterfactual statements derived from the image's NEST ground truth (wrong technique, opposite emotion, incorrect composition). CCR is the fraction of false statements the model correctly rejects, averaged per image.
- **Explanation Quality** (Track 4.2): when the model rejects a false statement, the rejection explanation is scored 0–1 against the NEST ground truth via an LLM-as-judge rubric on six dimensions (factual correctness, evidential specificity, alternative-explanation completeness, etc.); per-image score is the mean across rejected statements.
- **Semantic Coverage** (Track 4.3): per-image fraction of NEST ground-truth concepts the model's open-ended response covers, computed as the mean of six per-field coverage scores (composition, color harmony, technique, emotional, symbolic, narrative).
- **Information Density** (Track 4.3): per-image *unique-concept content per emitted token* in the model's open-ended response. Operationally: a frontier judge (Claude Sonnet 4.6 in this study) extracts the set of distinct concepts mentioned in the response, normalized against a NEST-derived concept ontology; the resulting concept count is divided by the response token count (Llama-3.2 tokenizer). The metric is bounded $[0, 1]$ and is computed on *model output*, not training data.

The headline metric for the 2×2 ablation is CogBench overall, evaluated on the 100-image held-out split with paired Wilcoxon contrasts at $n = 100$ (Wilcoxon, 1945). All paired contrasts in this paper compare arms image-by-image on the same evaluation set; null hypotheses are always two-sided.

4 Results

We organize results in four parts. Section 4.1 reproduces the headline numbers from the original paper for Groups A and B. Section 4.2 reports the new 2×2 ablation: Groups C and D versus Base, B, and A, with paired Wilcoxon contrasts. Section 4.3 reports the most informative single contrast in the new data: Group D versus Base across primary and secondary metrics. Section 4.4 summarizes the complementary out-of-distribution finding from companion work (Anonymous, 2026e).

4.1 Original A vs. B Headline (Reproduced)

Metric	Base	Group A	Group B
CogBench Overall	0.388	0.176 (−**54%**)	**0.422** (+9%)
Counterfactual CCR	0.984	0.511 (−48%)	0.835 (−15%)
Semantic Coverage Overall	0.097	0.026 (−73%)	**0.259** (+167%)
Information Density (output)	0.367	0.503 (+37%)	0.457 (+25%)
Avg. tokens emitted	161	38 (−76%)	136 (−15%)

Table 1: The original two-cell finding. Group A catastrophically degrades base capability across nearly every measure; Group B improves cognitive depth and semantic coverage substantively while reducing token output. Friedman test on CogBench across the three conditions: $\chi^2 = 157.2$, $p < .001$; pairwise Wilcoxon with Holm–Bonferroni correction confirms Base > A and B > A at $p < .001$.

The 24-percentage-point CogBench swing between Group A and Group B (0.176 vs. 0.422) is the strongest pairwise contrast in the original paper. The new data does not change this finding: A is catastrophically worse than B, and B is the only fine-tuned arm that improves cognitive depth above base in the original two-cell design. What the new data does change is the *interpretation* of which axis of the A-B difference is doing the work.

4.2 The 2×2 Ablation: Independent Sufficiency of Density and Structure

We extend the original A vs. B contrast with two new conditions filling the missing diagonal of the 2×2: Group C (dense, unstructured) and Group D (sparse, structured). Each is trained at two seeds (s42, s1337) for confirmatory replication. Table 2 reports the seed-paired primary CogBench metric and the secondary CCR, semantic coverage, and information density metrics for all four arms plus base.

Metric	Base	A sparse, unstr.	B dense, struct.	C (avg) dense, unstr.	D (avg) sparse, struct.
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 2: The 2×2 refinement: Groups C and D added. C and D are seed-averages over s42 and s1337 LoRA adapters trained at otherwise identical hyperparameters. Group B retains its single-seed configuration from the original paper. Bold indicates within-row maximum among the fine-tuned arms (B, C, D); base is reported for reference and is not bolded except where it is the global maximum.

The pattern is striking. Group C (dense, unstructured) lands at CogBench 0.429, statistically indistinguishable from Group B's 0.422 (paired Wilcoxon $\Delta = +0.008$, $p = 0.20$, two-sided). Group D (sparse, structured) lands at CogBench 0.409, also statistically indistinguishable from Group B (paired Wilcoxon $\Delta = -0.009$, $p = 0.24$, two-sided). The 24-percentage-point swing between A and B that the original paper attributed to "density" is, in the new data, attributable to either density *or* structure: neither dimension is privileged, and neither is required when the other is present.

Table 3 reports the full set of paired Wilcoxon contrasts on CogBench overall.

Contrast	**Δ CogBench**	W	p	**Reading**
B − A	+0.246	0.0	$< 10^{-15}$ ***	B rescues from A.
C − A	+0.254	0.0	$< 10^{-15}$ ***	C also rescues from A.
D − A	+0.238	0.0	$< 10^{-15}$ ***	D also rescues from A.
C − B	+0.008	2154	0.202 ns	C ≈ B; density alone matches dense+structured.
D − B	−0.009	2184	0.241 ns	D ≈ B; structure alone matches dense+structured.
B − Base	+0.034	1207	5.9×10^{-6} ***	B beats Base.
C − Base	+0.042	87	$< 10^{-15}$ ***	C beats Base.
D − Base	+0.025	804	3.3×10^{-9} ***	D beats Base.
A − Base	−0.212	0.0	$< 10^{-15}$ ***	A is uniquely catastrophic.

Table 3: Paired Wilcoxon contrasts on CogBench overall, $n = 100$ images. Two-sided. Group C and Group D rows use the s42 adapter for direct image-level pairing; per-seed contrasts (s42 vs s1337) are reported in the supplementary release. "$< 10^{-15}$" indicates the test returned a p-value below the IEEE-754 double-precision floor that `scipy.stats.wilcoxon` can resolve (raw return is 0 for these contrasts); we report the threshold rather than a numerical value to avoid false precision. C-vs-B ($p = 0.20$) and D-vs-B ($p = 0.24$) fail to reject the null at $\alpha = 0.05$; we adopt the language "no detectable difference" rather than "equivalent" to avoid conflating non-rejection with formal equivalence (a TOST equivalence test would require a pre-registered margin we did not specify). All seven significant contrasts ($p < .001$) survive Holm–Bonferroni correction across the 9 pairwise tests at $\alpha = 0.05$. Significance markers: *** $p < .001$; ns not significant at $\alpha = 0.05$.

The structure of significance in Table 3 compresses to a one-sentence statement: **either density or structure alone is sufficient to rescue from A, and the joint addition of both does not produce a measurable improvement over either alone**. We do not interpret this as evidence that combining density and structure is wasted—a larger evaluation set, a different base model, or out-of-distribution probes might surface a multiplicative effect—but the present in-domain measurement does not detect one.

The pattern survives 2-seed replication for Groups C and D on CogBench. Within-arm seed-pair variation on the primary CogBench measure is $|0.4300 - 0.4283| = 0.0017$ for Group C and $|0.4136 - 0.4049| = 0.0087$ for Group D—small relative to inter-arm differences. Seed stability is metric-dependent: CogBench is stable across the two seeds; CCR is not (Group D s42 0.91 vs. s1337 0.79, a 12-percentage-point seed-pair gap; see §6 Limitation 4). The companion paper independently documents broader post-training seed-fragility on this base under a different supervision regime (3-seed audit on SFT/ORPO arms, ∼11pp seed-to-seed verifier accuracy variance) (Anonymous, 2026e); the C/D 2-seed configuration cannot characterize seed-fragility at the resolution that companion audit could, and a 5–10 seed C/D replication is pre-registered as future work.

4.3 Group D vs. Base: The Elicitation Profile

Of the contrasts the new design enables, Group D versus Base is the most informative single comparison. Group D is sparse (low data volume) and structured (analytical schema present), and tests what structural scaffolding alone does to a pre-trained base model when controlling out the density axis. Table 4 reports paired Wilcoxon contrasts across all five primary metrics, separately for the two D seeds.

Metric	Base	D s42	Δ	p	Sig
CogBench Overall	0.388	0.414	+0.025	3.3×10^{-9}	***
Counterfactual CCR	0.984	0.907	−0.077	1.8×10^{-4}	***
Semantic Coverage	0.097	0.114	+0.017	2.5×10^{-4}	***
Information Density (output)	0.367	0.664	**+0.297**	$\mathbf{3.9 \times 10^{-18}}$	***
Explanation Quality (CounterVQA)	0.082	0.111	+0.028	0.284	ns

Metric	Base	D s1337	Δ	p	Sig
CogBench Overall	0.388	0.405	+0.017	3.4×10^{-5}	***
Counterfactual CCR	0.984	0.790	−0.194	1.4×10^{-8}	***
Semantic Coverage	0.097	0.111	+0.014	0.004	**
Information Density (output)	0.367	0.661	**+0.294**	$\mathbf{3.9 \times 10^{-18}}$	***
Explanation Quality (CounterVQA)	0.082	0.077	−0.005	0.198	ns

Table 4: Paired Wilcoxon contrasts: Group D vs. Base on the five primary metrics, separately for the s42 and s1337 LoRA adapters. $n = 100$ for CogBench, semantic coverage, and information density; $n = 94$ for CCR and explanation quality (the remaining 6 prompts had invalid model output and were excluded from the paired test rather than imputed). Two-sided Wilcoxon signed-rank test. Significance markers: *** $p < .001$; ** $p < .01$; ns not significant at $\alpha = 0.05$.

Three observations from Table 4 carry the new mechanistic story.

First, Group D produces a real and statistically significant cognitive-depth gain over the unmodified base. Both seeds clear $p < .001$ on CogBench overall ($\Delta = +0.025$ for s42 and $+0.017$ for s1337). The sparse, structured fine-tune is not just preserving base capability—it is producing a measurable cognitive-depth improvement. The original paper's framing of dense supervision as the only fine-tuning regime that improves over base is incomplete: structural scaffolding alone, at sparse density, also improves over base on the same metric.

Second, Group D's CCR loss versus base is the smallest of any non-trivial fine-tune in this study, but is seed-fragile. D s42 loses -0.077 (8%) of base CCR; D s1337 loses -0.194 (20%). The 12-percentage-point seed-spread on CCR alone is a finding worth characterizing in its own right; we discuss it under §6. By comparison, Group B loses -0.149 (15%) on a single seed; Group C loses -0.165 (17%) and -0.242 (s1337); Group A loses -0.473 (48%). The structurally scaffolded sparse regime is the gentlest non-trivial fine-tune we have measured on this base.

Third, the largest and most robust effect on any axis in the new data is Group D's gain in output information density: +0.297 over base, $p = 3.9 \times 10^{-18}$, replicated across both seeds. Group D's 0.66 information density is 81% above base's 0.37 and substantially above Group B's 0.46. Recall that information density is computed on *output*, not training data: it measures unique-concept content per emitted token in the model's response. Models trained on Group D produce dramatically denser output per token than models trained on any other arm or the unmodified base. This is the strongest single empirical claim the present paper makes, and it is robust across both replication seeds and across pairings (D vs. Base, D vs. A, D vs. B all show D producing the densest output of any compared arm; full pairwise table in the supplementary release).

The information density gain is also consistent with the qualitative profile of Group D's outputs. We attempt no causal interpretation in the headline; §5 discusses the elicitation hypothesis as a candidate mechanism.

4.4 Out-of-Distribution Scope Refinement (companion work)

The companion paper (Anonymous, 2026e) reports an out-of-distribution evaluation that bears directly on the present in-domain refinement. On a 953-prompt sequestered benchmark spanning MATH-500 (Levels 3–5), HumanEval+, BigCodeBench, MMLU-Pro, and 150 hand-curated prompts, evaluated by both a deterministic verifier and a 4-judge cross-lineage panel (Sonnet 4.6, Gemini 3.1 Pro, GPT-5.4, Grok-4-1), Groups A and B are statistically indistinguishable on every measured axis: verifier-grounded accuracy 32.9% vs. 32.9%, 4-judge Borda 0.566 vs. 0.575. The 11.1× per-record density advantage that drives the in-domain CogBench swing produces zero detectable benefit on this OOD benchmark.

We import this result into the present paper's contributions and discussion as a scope refinement. In-domain at this corpus and these benchmarks, density and structure are independently sufficient to produce the cognitive-depth gain. Out-of-distribution at this corpus scale and this base size, density does not detectably transfer. The original paper's implicit extrapolation from in-domain CogBench to general reasoning capability cannot be sustained on the basis of the present evidence; the in-domain claim itself is not falsified.

5 Discussion

5.1 Why Sparse, Unstructured Fine-Tuning Destroys

The original paper's mechanistic account of Group A's catastrophic degradation rested on three candidates: capacity overwriting (sparse targets teach the model to suppress its analytical pathways), low-entropy collapse (the supervision distribution is so narrow that the model converges to a near-constant response template), and gradient noise (sparse targets produce weak, inconsistent per-example learning signal).

The new 2×2 data refines this account. Capacity overwriting alone cannot explain A's catastrophe, because Group D imposes a sparse training signal of identical token volume to A but produces no comparable degradation—in fact Group D produces a real gain over base. Low-entropy collapse alone cannot explain it either, because Group C imposes an unstructured signal at the same token volume as B and likewise produces no comparable degradation. The new evidence localizes the catastrophic mechanism to the *joint* absence of density *and* structure: only when the supervision signal carries neither rich token volume *nor* an analytical schema does the catastrophic A regime emerge. The candidate mechanisms (capacity overwriting, low-entropy collapse, gradient noise) likely all contribute under the joint absence; none of them alone is sufficient to produce the catastrophe when the supervision signal carries either density or structure.

5.2 Why Either Density or Structure Alone Suffices

The new data offers two distinct pathways for the same cognitive-depth gain.

The density pathway (Group C). Dense, unstructured supervision provides ∼2,000–4,000 tokens per record of free-form prose describing the artwork. The model receives high-volume training signal that exposes it to varied descriptive language across many examples. Even without an explicit analytical schema, the sheer token volume per record provides enough learning signal to teach the model what kinds of analytical observations are appropriate for a cultural-heritage artwork. Group C achieves CogBench 0.429 on this pathway, statistically equivalent to Group B's 0.422.

The structure pathway (Group D). Sparse, structured supervision provides ∼50–80 tokens per record of named-field content, but those tokens are organized along the 111-field NEST schema. The model receives low-volume training signal with a high information-per-token ratio, and is exposed at every example to the analytical scaffold that organizes the domain. Even without rich token volume, the schema teaches the model how to organize its own analytical output. Group D achieves CogBench 0.409 on this pathway, statistically equivalent to Group B's 0.422.

The two pathways differ on *which* secondary metrics they additionally deliver. The density pathway produces higher in-domain semantic coverage (Group C 0.241 vs. Group D 0.113); the structure pathway produces dramatically higher output information density (Group D 0.66 vs. Group C 0.46). Each pathway has costs as well: Group C's CCR loss is roughly the same as Group B's; Group D's CCR loss is the smallest of any non-trivial fine-tune.

For a practitioner choosing between regimes, the operational decision is often clear. Where the goal is in-domain semantic breadth (richer responses across the topic space), the density pathway dominates. Where the goal is denser output, capability retention, and minimal weight perturbation—the typical operational frame for production fine-tuning—the structure pathway is the cleaner choice. Where data is abundant and structurable, both pathways are available; where data is sparse but structurable (a common operational situation: technical documentation, compliance corpora, scaffolded reasoning data), the structure pathway is the only viable option.

5.3 The Elicitation Hypothesis: D as Reorganization, Not Teaching

The qualitative profile of Group D suggests a mechanistic interpretation we cannot yet fully test in this paper but which we identify as the cleanest candidate hypothesis: *Group D operates as an elicitation regime—it reorganizes the model's existing pre-trained associations into reasoning-accessible form, rather than teaching the model new associations.*

The hypothesis is consistent with three observations from Table 4. First, Group D's CogBench gain over base (+0.025, $p < .001$) is real but modest—approximately 75% of Group B's 0.034 gain—suggesting that the structural scaffolding is making something present in the base model more accessible rather than producing a qualitatively new capability. Second, Group D's CCR loss is small (8% on s42), consistent with minimal weight perturbation: the model is not being substantially rewired, only retrieved differently. Third, Group D's semantic-coverage gain over base is small (+0.017) and only marginally significant on s1337, in stark contrast to Group B's +0.162 expansion—suggesting that the structural scaffolding is not expanding the categorical breadth of what the model knows, only the analytical accessibility of what it already knows.

If the elicitation hypothesis is correct, it makes a falsifiable prediction: **Group D should fail on a domain the base model has not seen during pre-training.** Sparse-structured supervision can reorganize existing pre-trained associations; it cannot teach the model facts it does not already encode. We propose this as future work (§7) and pre-register the prediction here. A confirmed failure of D on truly novel-domain knowledge would localize structural scaffolding as a retrieval-organization mechanism rather than a knowledge-acquisition mechanism, with significant implications for how sparse-structured fine-tuning should be deployed in practice.

5.4 Structure as Active Ingredient: The "Hallucination of Expertise" Pattern, Quantitatively Supported

The original paper's Section 5.3 ("Evidence for Structure over Content: The Hallucination of Expertise") argued, on the basis of qualitative output analysis alone, that Group B's gains were attributable to learned analytical structure rather than memorized content—citing as evidence that Group B confidently misidentified held-out artifacts while applying the NEST schema's analytical methodology flawlessly to those misidentifications. The argument was suggestive but not empirically isolated, because the original two-cell design could not distinguish density from structure.

The new 2×2 data is consistent with the structural-scaffolding interpretation: Group D's +0.297 information-density gain over base, achieved at sparse $1\times$ token volume, indicates that structural scaffolding alone—without dense token volume—is sufficient to produce the dense, analytically-organized output style that the original paper attributed to dense fine-tuning. The new evidence does not directly probe the qualitative "confident misidentification while applying the schema flawlessly" phenomenon (that is a per-output diagnostic, not an aggregate metric), but it does isolate the structural axis as the active ingredient on the metric most relevant to that phenomenon—output information density. The full qualitative-validation experiment for the misidentification phenomenon (run the original held-out images through Group D and audit how often the model confidently misidentifies while applying the NEST analytical methodology) is itself worthwhile follow-up work.

5.5 The Information Efficiency Finding, Refined

The original paper noted that Group B produced substantially more semantic coverage than base while generating fewer tokens, framing this as "dense training teaches information compression." The new data refines the framing. Recomputing on the present source aggregates: Group B produces a +167% semantic-coverage gain over base (0.0974 → 0.2592); Group C produces a +147% gain (→ 0.2410); Group D produces only +16% (→ 0.1126). The semantic-coverage gain is therefore attributable to density (B and C both clear +140%) rather than structure (D's gain is small and barely a tenth of B's). Conversely, the information-density gain—more concept content per emitted token—is attributable to structure: Group D produces +81% over base (0.3669 → 0.6626); Group C produces only +25% (→ 0.4575); Group B produces +25% (→ 0.4569). Density buys breadth; structure buys per-token information richness. The original paper's single-axis framing of "information compression" conflated the two effects.

5.6 Out-of-Distribution Scope: What Does Not Transfer

The companion paper's 953-prompt OOD evaluation (Anonymous, 2026e) reports A and B as statistically indistinguishable on every measured axis: verifier-grounded accuracy (32.9% vs. 32.9%), 4-judge Borda aggregate (0.566 vs. 0.575), and per-benchmark verifier accuracy (MATH-500 11.5% vs. 10.5%, HumanEval+ 47.9% vs. 47.2%, MMLU-Pro 39.4% vs. 40.3%). The $11.1\times$ density advantage that drives the in-domain CogBench swing produces zero detectable OOD benefit at this corpus scale and base size.

We accept this as a scope refinement. The in-domain claim of the present paper survives the OOD measurement: A is catastrophic, C and D rescue from A, and either density or structure alone suffices on in-domain CogBench. The implicit extrapolation from in-domain cognitive-depth improvement to general reasoning capability is not supported by the OOD measurement and should not be made on the basis of the present evidence. Whether the in-domain effect transfers OOD

at larger base sizes (70B-class), larger corpora (10k–100k records), or with substantially different evaluation surfaces remains open future work.

5.7 Implications for the Fine-Tuning Supply Chain

Two operational implications follow from the new evidence.

Avoid sparse, unstructured fine-tuning data. The strongest single recommendation that survives the 2×2 refinement is the catastrophic floor finding: Group A is uniquely catastrophic. Training on ∼50-token unstructured captions of the kind typical in web-scraped image–text pairs degrades the base model below its un-fine-tuned baseline by 54% on cognitive depth, by 48% on capability retention, and by 73% on semantic coverage. The widespread practice of fine-tuning vision-language models on alt-text-grade supervision is not capability-neutral; on the basis of the present data, it is actively destructive on this base size and corpus.

Structure your sparse data; do not generate density you do not need. For practitioners working with corpora that are sparse but structurable—internal documentation, technical manuals, compliance archives, scaffolded reasoning data, scientific abstracts—the new data offers a clean operational prescription: lift the structural schema that organizes the domain and use that schema to build sparse-but-structured supervision (Group D regime). This regime matches dense-structured supervision on cognitive depth, dominates it on capability retention and output information density, and avoids the cost of generating 11× the token volume per record.

The two recommendations together describe the operational frontier the present data reveals: *do not train on Group A; either Group B, C, or D will do, and Group D is the cheapest of the three when applicable.*

6 Limitations

1. **Single domain.** All training and evaluation data comes from cultural heritage. Whether the 2×2 pattern generalizes to medical imaging, satellite data, code, or general natural images requires investigation. The companion OOD work (Anonymous, 2026e) provides one cross-domain probe (MATH-500, HumanEval+, MMLU-Pro) and finds no measurable density transfer; the in-domain 2×2 pattern reported here has not yet been independently replicated outside the Alexandria Aeternum corpus.

2. **In-distribution evaluation set.** The 100-image evaluation split is held out from training but is drawn from the same Alexandria Aeternum collection. Out-of-distribution evaluation on works from other institutions (Rijksmuseum, Smithsonian) would strengthen the in-domain finding's generalizability claim within the cultural-heritage domain.

3. **Single base model.** All conditions use Llama 3.2 11B Vision-Instruct. Whether the catastrophic A regime, the independent sufficiency of density and structure, or the elicitation profile of Group D replicate on different base architectures (Qwen-VL, LLaVA, InternVL, larger Llama variants) is open. We pre-register a multi-base catastrophic-A replication as future work.

4. **CCR seed-fragility on Group D.** Group D's CCR loss versus base is 8% on seed s42 and 20% on seed s1337—a 12-percentage-point seed-pair gap on a single metric that is larger than the C-vs-B and D-vs-B differences on the headline CogBench measurement. We characterize this as a real effect of the structural-supervision regime that is not present at

the same magnitude on the cognitive-depth axis or the information-density axis (both of which replicate cleanly across the two seeds). A larger-N seed-replication study (5–10 seeds per arm) is the appropriate next step; we report the present 2-seed result honestly and do not draw an inferential claim from it.

5. **Single hyperparameter configuration.** We evaluate one LoRA rank ($r = 32$), one learning rate (2×10^{-4}), and one epoch. Higher ranks, longer training, or different optimizers may alter the magnitudes reported in Tables 2–4. We do not expect them to alter the qualitative pattern (A catastrophic, C $\approx$ B $\approx$ D on CogBench) but cannot rule out larger-rank ablations that reveal hidden multiplicative effects.

6. **Sample size.** $n = 100$ evaluation images with 909 in reserve for community replication. Statistical significance is strong on the metrics that move ($p < 10^{-9}$ on D-vs-Base CogBench and information density) but the evaluation set is modest by large-scale benchmark standards. We do not interpret the C-vs-B and D-vs-B null results as evidence of zero effect at larger N.

7. **Single GPU class.** All training and evaluation conducted on a single NVIDIA L4 (24GB VRAM). The L4 constrains image resolution (560px, 1 vision tile) and batch size but does not affect model outputs deterministically given identical weights, quantization, and precision. All four conditions used the same hardware and image resolution.

8. **The elicitation hypothesis is not directly tested in this paper.** We identify the falsifiable prediction (Group D should fail on a truly novel domain) and pre-register it as future work, but the present paper does not run the experiment. The elicitation hypothesis is a candidate mechanistic interpretation consistent with the present data, not a validated mechanism.

9. **POPE per-image data not preserved.** POPE probe-level data was not preserved from the original training VM and is not available for paired-test analysis on Groups C and D. We report POPE as it was computed in the original paper and do not extend its analysis to the new arms.

10. **Single-rubric in-domain evaluation.** All in-domain evaluation in this paper uses the CogBench cognitive-depth rubric and the NEST-grounded counterfactual / semantic-coverage suite. The companion paper (Anonymous, 2026e) documents a verifier-vs-judge-panel divergence on out-of-distribution reasoning—training interventions can win one evaluation axis at the cost of another. The C $\approx$ B $\approx$ D pattern reported here has not been characterized under a deterministic verifier or under a multi-judge cross-lineage panel in-domain. Whether the pattern would replicate, attenuate, or invert under those alternative evaluation surfaces is open. We treat this as a methodological scope, not a validity threat to the present paper's claims, but flag it for completeness.

11. **C/D-corpus generation pipeline (single-source).** Groups C and D were generated via deterministic transformations of the original Group B NEST records (Group C via a Gemini 2.5 Pro paraphrase at temperature 0; Group D via a Python compression script). The original Group B NEST corpus was itself generated by an orchestrated multi-model pipeline (Gemini-2.5-Pro + GPT-4o + Claude Sonnet 4.6 in three-teacher consensus). Style or content biases that may have entered Group B inherit into both C and D, since both are derived from B. This is a real residual confound the present design cannot eliminate

without an entirely separately sourced structured corpus—worthy of follow-up but does not threaten the catastrophic-floor or independent-sufficiency claims, which compare across density×structure rather than across generation pipelines.

12. **Group A's elevated information density is collapse-driven, not analytical.** Table 1 shows Group A producing 0.503 information density (37% above base) despite catastrophic capability collapse. This is consistent with low-entropy collapse on extremely short, repetitive output (mean 38 tokens vs. Base's 161): Group A's responses are so terse that any concept they mention is divided by a small denominator, inflating per-token concept density without reflecting analytical depth. The +81% Group D information-density gain is qualitatively different—Group D maintains a substantially longer mean response and delivers concept density at a token volume comparable to Base—but the metric on its own does not distinguish "denser" from "more collapsed." We rely on the joint reading of information density alongside response length, semantic coverage, and CogBench scores to interpret the Group D effect as analytical rather than degenerate.

7 Future Work

The elicitation-bound experiment. Train Group D-style sparse-structured supervision on a domain the base model has not seen during pre-training (e.g., a synthetic-physics regime, a fictional legal system, an artificial taxonomy) and measure whether the cognitive-depth gain replicates. The elicitation hypothesis predicts collapse; an alternative "structure-as-curriculum" hypothesis predicts gain. The result discriminates between the two interpretations cleanly.

Cross-domain density transfer. Train Group B-style dense-structured supervision on ArXiv-only NEST corpora and evaluate zero-shot reasoning on FreeLaw and PubMed. The hypothesis under test is whether dense structural scaffolding teaches a universal cognitive skill that transfers across domains, or only domain-specific knowledge. A confirmed positive result would substantially strengthen the in-domain claim of the present paper.

Structural calibration. Build chosen/rejected preference pairs where the chosen half is a calibrated abstention ("I do not have enough information to solve this") and the rejected half is an over-confident hallucination on the same prompt, then run ORPO. The hypothesis is whether calibration can be taught as a structural reasoning behavior rather than a surface-token emission, extending the surface-signal anti-calibration finding of (Anonymous, 2026e).

NEST L1–L5 curriculum learning. The 111-field NEST schema is naturally hierarchical (Visual → Semantic → Philosophical). Train successive arms on progressively deeper structural slices (flat / L1 only / L1+L2 / L1–L5) and measure learning-efficiency curves. The hypothesis is whether structural depth produces compounding gains in capability per training token.

Pre-training-scale extension. Apply the structure-rich-sparse regime (Group D analog) to *pre-training* rather than fine-tuning. Train two models from scratch on equal-FLOP budgets, one on standard pre-training data and one on Aeternum-scaffolded versions of the same corpora, and measure on standard reasoning benchmarks. The pre-training-scale extension is the most consequential prediction the present data motivates; it requires substantially more compute than fine-tuning and is positioned as a foundation-model-class experiment for follow-up work.

Larger seed replication. Run 5–10 seeds per arm on the new 2×2 to characterize the seed-fragility of CCR on Group D and confirm the C $\approx$ B $\approx$ D pattern at higher statistical power than the present 2-seed configuration provides.

Multi-base catastrophic-A replication. Reproduce Group A protocol on 3–4 base model families (Qwen, Mistral, Phi, larger Llama) to confirm that the catastrophic floor generalizes across architectures.

Output information-density transfer characterization. Test whether Group D's +81% information-density gain replicates on out-of-distribution prompts (math, code, legal text). If the effect is a robust style-transfer from training data, characterize the conditions under which it transfers and the conditions under which it does not.

Higher LoRA ranks. Test $r \in \{64, 128\}$ to assess whether increased adapter capacity reveals hidden multiplicative effects between density and structure that the rank-32 ablation cannot detect.

Beyond vision-language models. The provenance-binding framework that produces the NEST metadata is domain-agnostic. Preliminary anonymized work (Anonymous, 2026d) suggests that the same structure-rich-supervision principle may transform embodied AI training pipelines.

8 Conclusion

The original Density Imperative was correct in the direction of its empirical claims and underspecified in their mechanistic attribution. The two-cell design conflated density and structure; the new 2×2 ablation closes the design and reveals that either density or structure alone is sufficient to recover cognitive depth from the catastrophic Group A regime. The 24-percentage-point CogBench swing the original paper attributed to "density" is, in the new data, attributable to either dimension. Group A is uniquely catastrophic; the joint absence of density *and* structure is the only condition under which the base model is degraded below its un-fine-tuned baseline.

The strongest operational claim that survives the 2×2 is the catastrophic floor: **do not train on sparse, unstructured supervision**. Web-alt-text-grade fine-tuning data is not capability-neutral on this base size and corpus; it is actively destructive. The cleanest operational regime when sparse data is structurable is Group D: structure-rich, density-poor supervision that matches dense-structured supervision on cognitive depth, dominates it on capability retention and output information density, and costs an order of magnitude less per training record.

The mechanistic interpretation we identify but do not fully validate is the *elicitation hypothesis*: that sparse-structured supervision reorganizes the base model's existing pre-trained associations into reasoning-accessible form rather than teaching new associations. The hypothesis predicts that Group D should fail on a domain the base model has not seen, and we pre-register this as the experiment that would falsifiably bound it.

The complementary out-of-distribution finding from companion work (Anonymous, 2026e)—that the $11.1\times$ density advantage between Groups A and B does not detectably transfer to a 953-prompt OOD reasoning benchmark—refines the present paper's scope. The in-domain claim is not falsified; the implicit extrapolation to general reasoning capability is. Whether density or structure transfers OOD at larger base sizes or larger corpora remains open future work.

We release four LoRA adapters (Groups A, B, C $\times$ 2 seeds, D $\times$ 2 seeds), all evaluation outputs, raw aggregates, paired Wilcoxon contrast scripts, and an interactive comparison tool at an anonymized HuggingFace organization (URL withheld for double-blind review; restored at camera-ready) for community verification (Apache-2.0 license; DOI deferred until camera-ready). The Alexandria Aeternum dataset is available at an anonymized HuggingFace dataset repository (URL withheld for double-blind review; Zenodo DOI deferred until camera-ready).

Fine-tuning data quality is signed in only one direction: sparse, unstructured supervision destroys. Either density or structure rescues. Structure is the cheaper rescue path, and it preserves more of what the base model already knows.

Broader Impact Statement

This work refines a previously published claim that fine-tuning data quality is signed (the Density Imperative), demonstrating via a 2×2 ablation that the catastrophic floor (sparse, unstructured supervision) is the operationally critical finding, while density and structure are independently sufficient at the recovery axis. **Positive impacts.** The operational prescription "do not train on sparse, unstructured data" is robustly evidence-based and may redirect industry practice away from web-alt-text-grade fine-tuning regimes that the present data documents as actively destructive on this base size. The further finding that sparse-but-structured supervision is a viable cost-efficient regime opens a path for resource-constrained practitioners working with structurable sparse corpora (technical manuals, compliance documentation, scaffolded reasoning data). The full open release of all adapters, evaluation tooling, raw outputs, and the interactive comparison tool enables community verification and extension. **Potential negative impacts.** The cognitive-depth gain reported here is a 100-image in-domain measurement at a single base size; companion out-of-distribution work shows the density component does not detectably transfer to general reasoning. Practitioners should treat the present paper's claims as in-domain and seek their own out-of-distribution validation before deploying. The structural supervision regime presupposes the existence of an analytical schema for the domain; for domains without an established structural ontology, the cost of constructing one may exceed the cost of generating density. The cultural-heritage focus of the present experiments may not generalize to all domains without further validation; the preliminary multi-domain probe in the companion paper should be read alongside this work.

Acknowledgments

Acknowledgments, competing-interests disclosure, funding statement, and author contributions are deferred to camera-ready submission per double-blind policy. A companion submission to a peer-reviewed venue reports the out-of-distribution evaluation imported into this paper as Contribution 5 (§4.4); the anonymized citation (Anonymous, 2026e) is preserved for cross-reference.

References

A. Alemohammad, J. Casco-Rodriguez, et al. Self-consuming generative models go MAD. In *Proceedings of ICLR*, 2024.

Anonymous. Prior work on the provenance-binding theoretical framework underlying the present study. Technical report, Title, author, and venue redacted for double-blind review, 2026a. Anonymized self-citation; full bibliographic record deferred to camera-ready.

Anonymous. Original two-cell preprint on vision-language fine-tuning metadata quality, of which this paper is the 2×2 refinement. Technical report, Title, author, and venue redacted for double-blind review, 2026b. Anonymized self-citation; original preprint cited as prior work. Full record deferred to camera-ready.

Anonymous. Prior work on the provenance-embedded data-refinery architecture. Technical report, Title, author, and venue redacted for double-blind review, 2026c. Anonymized self-citation; full bibliographic record deferred to camera-ready.

Anonymous. Prior work extending the provenance-embedded framework to embodied ai training pipelines. Technical report, Title, author, and venue redacted for double-blind review, 2026d. Anonymized self-citation; full bibliographic record deferred to camera-ready.

Anonymous. Companion submission under double-blind review at the same venue. Technical report, Title, author, and venue redacted for double-blind review, 2026e. The cited companion work reports an out-of-distribution evaluation on a 953-prompt sequestered reasoning benchmark; bibliographic record deferred to camera-ready.

H. Chen et al. CounterVQA: Counterfactual visual question answering. *arXiv preprint arXiv:2511.19923*, 2025.

E. Dohmatob, Y. Feng, and J. Yang. A closer look at model collapse: From a generalization-to-memorization perspective. *arXiv preprint arXiv:2509.16499*, 2025.

S. Gadre, G. Ilharco, A. Fang, et al. DataComp: In search of the next generation of multimodal datasets. In *Proceedings of NeurIPS Datasets and Benchmarks*, 2023.

HuggingFace Team. FineWeb: A large-scale open web dataset. Technical report, HuggingFace, 2024.

Y. Li, Y. Du, K. Zhou, et al. Evaluating object hallucination in large vision-language models. In *Proceedings of EMNLP*, 2023.

G. Penedo, Q. Malartic, D. Hesslow, et al. The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data alone. In *Proceedings of NeurIPS*, 2023.

I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot, and R. Anderson. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759, 2024.

Z. Song et al. CogBench: A benchmark for measuring cognitive depth in vision-language models. In *Proceedings of NAACL*, 2025.

F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

S. Xie, H. Pham, X. Dong, N. Du, H. Liu, and Y. Lu. DoReMi: Optimizing data mixtures speeds up language model pretraining. In *Proceedings of NeurIPS*, 2023.

S. Yu et al. VULCA-Bench: A benchmark for visual understanding and language competency assessment. *arXiv preprint arXiv:2601.07986*, 2025.

C. Zhou, P. Liu, P. Xu, et al. LIMA: Less is more for alignment. In *Proceedings of NeurIPS*, 2023.