When Retrieved Context Looks Better Than It Is: Verbosity, Visual Substitution, and a Hidden Precision Tax in a Frozen Vision–Language Model

A measurement-and-mechanism study on one frozen VLM and a cultural-heritage corpus (the "weights-to-context" question)

Tad MacPherson
Metavolve Labs, Inc.

2026-06-09

Abstract

Prior work in this program established that the *quality of training data* governs the capability of a fine-tuned vision–language model (VLM): sparse, unstructured supervision catastrophically degrades cognitive depth, while density and structure are independently sufficient to rescue it. Those results placed the knowledge in the *weights*. We ask whether the same levers operate at *inference time*, supplied to a *frozen* model through retrieved context—reframing the question from *instillation* to *elicitation*. Holding Llama-3.2-11B-Vision-Instruct frozen and varying only the context injected before a generation-quality query, we evaluate eight context arms over n=100 held-out images from the Alexandria Aeternum corpus. Our central finding is a measurement-validity result: **a verbosity-confounded composite metric and a length-controlled content metric disagree on the sign of the apparent effect.** Scored with our inherited *richness composite* ($\mathrm{RC} = 0.6 \cdot \mathrm{coverage} + 0.4 \cdot \mathrm{depth}$, $\mathrm{depth} = \min(1, n_{\mathrm{words}}/50)$), *any* context—including *random* exemplars—raises the score, reversing the base/random ordering. Scored with concept *coverage* (response length regressed out), the ordering is monotone: exact curation (an oracle) dominates (+0.155 coverage, $p<0.001$); relevance-bearing retrieval beats random (+0.018 to +0.028, $p<0.001$); and generic "priming" context *raises* the composite *without* raising content (−0.010 coverage, marginal under a mixed model). A weak visual-over-text edge does not survive correction and is not claimed. Output *formatting* (structured vs. length-matched prose) does *not* transfer: a marginal single-seed trend (p=0.08) is non-significant and near-zero at two further seeds (p=0.88, 0.99). Two further controls deepen the picture: an *image-ablation* test (real image vs. blank placeholder) shows the model genuinely uses the image, but retrieved context significantly crowds out that visual reliance (negative interaction, every arm, both seeds, $p<0.001$); and a three-lineage *faithfulness* panel (a Claude judge swarm, GPT-5.4, Gemini-3.1-Pro) shows retrieval's recall comes at a steep, previously-invisible precision cost—groundedness collapses from 0.60 (base) to 0.19 (retrieval), unanimously across all three judges, with random context leaving ∼93% of claims hallucinated. At inference time the levers are *relevance* and, far above it, *curation*—not formatting—and a verbosity-confounded metric can manufacture an "elicitation" effect that content-grounded measurement reveals to be largely hallucinated. All results are from a *single* frozen model (Llama-3.2-11B-Vision) on *one* richly-annotated cultural-heritage corpus; we therefore frame the contribution as a measurement-and-mechanism study of this setting, not a universal law, and scope the claims accordingly.

Plain-language summary. We test whether an image-reading AI can do better *without retraining*, just by being handed relevant background text at answer time. On one art corpus: relevant text helps a little, an ideal hand-picked record helps a lot, and neat *formatting* of the text does not help. But a common scorecard was misleading—even *irrelevant* text raised the score by making the model write more. Two further tests sharpen the lesson: removing the image (replacing it with a blank) shows the model truly uses the picture, yet added text makes it lean on the picture *less*; and a three-judge faithfulness panel shows the added text makes the model *hallucinate*—confidently stating things that are not true of the image. The takeaway: better *curation* beats prettier formatting, and evaluation must measure *grounding*, not just wordy recall.

1 Introduction

The program preceding this paper produced two results about *training* VLMs. The **Density Imperative** [1] showed that the informational density of supervision governs capability; a 2×2 ablation [2] showed that organization (explicit analytical schemas) is an independently sufficient lever, alongside a provenance/ integrity framing [3]. Both shared a condition: the knowledge lived in the *weights*, instilled by fine-tuning—an expensive, static, per-model intervention.

This motivates the present question: *must dense, structured knowledge be compressed into the weights to confer its benefit, or can the same levers operate at inference time, supplied to a frozen model through its context window?* The distinction is mechanistic. In training, the model generalizes from compressed structure; at inference, it must read and use structure presented in-context with no parameter change. There is no guarantee a principle observed in the weights transfers to the context. We therefore reframe the question as one of **elicitation rather than instillation**: holding the model frozen, can retrieved dense, structured context recover the capability that dense, structured training conferred?

In carrying out this test we encountered—and report as a first-class result—a *measurement* problem that anyone studying retrieval-augmented generation with richness-based metrics should heed. The metric we inherited rewards output length by construction, and under it *every* intervention that lengthens output "works," including injecting random context. Only when we re-scored on a length-controlled content metric did the real structure of the effect appear.

Contributions.

1. **A measurement-validity result.** On a verbosity-confounded composite metric, retrieval-augmented context (even *random* context) appears to elicit capability—the metric reverses the base/random ordering; on a length-controlled content metric, that appearance is largely an artifact. We quantify the confound and report which conclusions survive it.

2. **Curation, not retrieval cleverness, is the dominant inference-time lever**—with an honest caveat. Exact curated context (an oracle) lifts content coverage by +0.155 ($p<0.001$), far above any retrieval arm. The oracle injects the very gold record that coverage is scored against, so this is best read as an upper bound on what *any* text context could supply (a "text-substitutes-vision bandwidth" ceiling), not a clean capability bound; that the model still attains only 0.226 coverage with the answer key in context indicates the bound is non-trivial.

3. **Relevance transfers; generic priming does not.** Relevance-bearing retrieval beats a random-context control on coverage ($p<0.001$), while random context *lowers* coverage relative to no context even as it *raises* the composite. A visual-over-text advantage is present but

weak (+0.010, image-level p=0.010, at the Bonferroni edge) and single-seed; we report it as suggestive only.

4. **Formatting does not transfer, replicated across seeds.** A structured-vs-prose contrast, token-matched to remove the length confound, is null; a marginal single-seed trend (p=0.08) does not replicate and is near-zero at two further seeds—reflexively confirming this program's own caution against single-seed evaluation.

5. **The image is used, but context substitutes for it.** An image-ablation control (real vs. blank image) shows the frozen model genuinely uses the image, yet retrieved context significantly reduces that visual reliance (negative interaction across every arm, two seeds, p<0.001)—"relevance transfer" is partly visual cognition and partly text substitution.

6. **Relevance transfer carries a hidden precision tax.** A three-lineage faithfulness panel (Claude judge swarm, GPT-5.4, Gemini-3.1-Pro) shows retrieved context *collapses* groundedness (base 0.60 → retrieval 0.19, unanimous; random leaves ∼93% of claims hallucinated)—the precision penalty that recall-only metrics structurally hide.

2 Related work

Length/verbosity bias in evaluation. That automatic and model-based metrics reward output length is well documented: verbosity bias in LLM-as-judge and arena-style evaluation [6, 7], in RLHF reward models [5], and in preference labeling [8]. Length-controlled AlpacaEval [4] debiases by regressing length out of the score—the same de-confounding move we apply to coverage. Our contribution is not the technique but its consequence in a *retrieval* setting: a length-confounded richness composite does not merely *rank* verbose answers higher, it *reverses the sign* of a retrieval comparison (random context appears to help while reducing content), and we quantify how much of an apparent "elicitation" effect is verbosity.

Knowledge in weights vs. context. The trade-off between instilling knowledge by fine-tuning and supplying it in-context has been studied directly [9] and underlies retrieval-augmented generation [10] and nearest-neighbor LMs [11]. We adopt this weights-vs-context lens for a frozen VLM and ask specifically whether the *density/structure* levers established in training transfer to retrieved context.

Demonstration retrieval and (visual) in-context learning. That relevant retrieved exemplars beat random ones is a founding result of demonstration-selection work [12, 13] and of visual in-context learning and exemplar retrieval for VLMs [14, 15, 16]. Our relevance-vs-random result replicates this in our setting; the novel angle is that the apparent effect is metric-dependent and, on a content metric, modest.

RAG evaluation and concept-recall metrics. Faithfulness- and relevance-grounded RAG evaluation [17, 18] operationalizes the content-grounded measurement we advocate; our coverage metric is a concept-recall measure in the lineage of SPICE [19], and shares its known precision blind spot, which we treat as a limitation.

3 Background: the instillation baseline

On the held-out generation-quality suite the unmodified base model scored 0.388 on the composite metric. Sparse, unstructured supervision (Group A) induced a catastrophic collapse to 0.176; dense, structured supervision (Group B) recovered to 0.422. A 2×2 ablation showed density and structure are *independently sufficient*: Group C (dense, unstructured) reached 0.429 and Group D (sparse, structured) 0.409, both statistically indistinguishable from B. These training-time figures are reproduced from the source papers (not re-derived here) and serve only to frame the inference-time question. (The instillation-suite base score 0.388 is from those papers' evaluation; it is not the same run as this paper's frozen-base composite of 0.378 on the 100-image elicitation set in Table 1.)

4 Method

Model and state. Llama-3.2-11B-Vision-Instruct, completely *frozen*—no parameter updates, no adapters. The sole manipulated variable is the context injected before each query. Decoding: `do_sample=True`, temperature 0.3, max 320 new tokens.

Data. The Alexandria Aeternum corpus (Met-derived; 111-field "NEST" deep-analysis metadata). Retrieval pool ≈10,000 records; n=100 held-out eval images, four question types per image (narrative, mental-state, focal-point, causal-depth).

Two metrics (the crux). The inherited composite—which we call the *richness composite* (RC)[1]—is $\mathsf{RC} = 0.6 \cdot \text{coverage} + 0.4 \cdot \text{depth}$, where $\text{depth} = \min(1, n_{\text{words}}/50)$. *Forty percent of this score is a monotone function of output length* (and saturates at 50 words), and coverage itself is recall without a precision penalty—so any intervention that lengthens output inflates the score regardless of relevance. We therefore report three quantities per arm: (i) the composite RC (for comparability with the inherited analysis); (ii) **concept coverage**, the content metric; and (iii) a **length-residual coverage**, in which coverage is regressed on response word-count (OLS, pooled) and the residual isolates content from verbosity.

Statistics. Per-arm values are means. Paired contrasts report the mean Δ with a 95% confidence interval and the Wilcoxon signed-rank p-value, computed both at the *item* level (image×question, up to 400 pairs) and at the *image* level (per-image mean, n=100 independent units; CIs are image-level). Because the four questions per image are not independent, the image-level test is the conservative one we rely on; with five primary contrasts a Bonferroni threshold is 0.010. Because the signed-rank test still assumes independent paired units, we additionally confirm every primary contrast with a **cluster bootstrap** (5,000 resamples of *images*, the nesting unit), which respects the repeated-measures structure: all primary effects retain 95% CIs that exclude zero (random−base $[-0.016, -0.005]$; clip−random $[+0.022, +0.034]$; bm25−random $[+0.012, +0.024]$; oracle−base $[+0.139, +0.170]$), matching the Wilcoxon conclusions. We additionally fit a **linear mixed-effects model** (coverage ∼ arm + question-type, random intercept per image; n=2000 items, 100 images). It confirms the transfer contrasts and the ceiling (bm25−random +0.018, p=0.002; clip−random +0.028, p<0.001; oracle−base +0.155, p<0.001) but finds the *small* priming contrast only **marginal** (random−base −0.010, p=0.08). We therefore down-weight the strong

[1] RC is our own internal metric, inherited from an earlier pipeline. A name previously used for it coincided with the unrelated published *CogBench* benchmark; that collision is unintentional, and we rename it here to avoid any confusion.

form of the priming claim: the metric-divergence result does not depend on it, since random context raises RC regardless of whether it *lowers* coverage or merely *fails to raise* it.

Arms (8). **base** (no context, floor); **oracle** (the eval record's own gold NEST, ceiling); **random** (relevance control: real exemplars, irrelevant); text-retrieval via BM25 in three formattings—**bm25-structured** (`lbrain_bm25_k3`), **bm25-prose** (`unstructured_bm25_k3`), **bm25-prose-matched** (`bm25_matched_k3`, truncated to the structured arm's exact length); and visual-retrieval via CLIP image-similarity in two formattings—**clip-structured** (`lbrain_clip_k3`), **clip-prose** (`unstructured_clip_k3`). Contrasts labelled "clip"/"bm25" below use the structured arm of each retriever.

Leakage controls. The retrieval query uses only shallow catalog fields (title/creator/date/department) or the eval *image* itself—never the gold NEST. Eval ids and perceptual (pHash) near-duplicates are excluded from the retrieval pool at run time (threshold 6), so a positive retrieval result cannot be a near-duplicate leak.

Seeds. The main eight-arm matrix is run at one sampling seed (n=100). The formatting contrast (the title's structure question) is additionally replicated at two further seeds to test single-seed fragility. The image-ablation matrix (below) is run at two seeds (42 and 7).

Image ablation and faithfulness (added controls). To separate visual cognition from text completion, each key arm (base, random, bm25, clip, oracle) is re-run with the real image replaced by a neutral gray *blank* placeholder, retrieved context held fixed; the 2×2 (image × context) design isolates the image's contribution via the interaction. To recover the precision that coverage (pure recall) omits, we score **faithfulness**—the fraction of a response's concrete claims grounded in the verified gold record—with a three-lineage panel: a swarm of Claude judges plus GPT-5.4 and Gemini-3.1-Pro, each blind to the arm, under an identical strict rubric (extract claims; mark each supported or unsupported against the record; faithfulness = supported/total), aggregated as the panel mean. The faithfulness proxy is text-grounded (against the record, not the pixels), so it is conservative about visually-true-but-unrecorded claims; the cross-lineage agreement and effect magnitude make the direction robust regardless.

5 Results

5.1 The metric divergence

Table 1 reports all three quantities per arm. Read down the RC column and the arms look nearly flat above a raised floor—random context alone lifts the composite by +0.028 (relative to base), and the retrieval arms cluster just above it. Read down the *coverage* column and a different, ordered story appears: **oracle** ≫ **visual** > **text** > **base** > **random**. The length-residual column confirms it is content, not verbosity: random is *worst* (pure length, no content), the retrieval arms are positive relative to random, and oracle is far above everything. Figure 1 makes the confound visible: per item, the composite rises with output length (R^2=0.36, slope +0.003/word), so RC compresses arms that the content metric separates.

5.2 Primary contrasts (coverage, length-robust)

Table 2 runs the comparisons that matter on coverage, as mean Δ with 95% CIs and both item- and image-level significance.

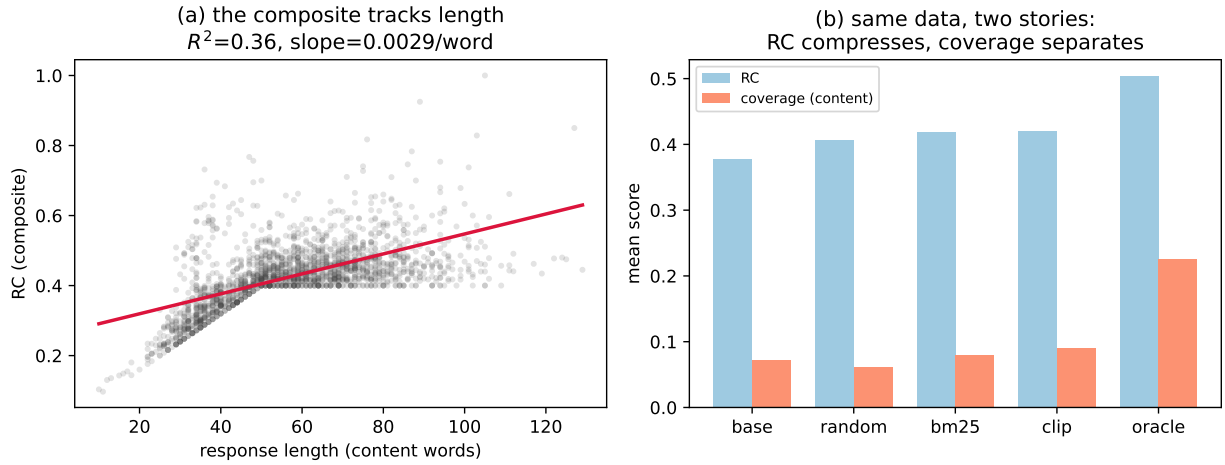

Figure 1: The metric divergence. **(a)** Per item (∼2000 points across arms), the inherited composite RC rises with response length. **(b)** Per arm, RC compresses the arms into a narrow band while content *coverage* separates them—and *inverts* the base/random ordering: random context *raises* RC but *lowers* coverage.

Two results deserve emphasis. First, the **priming confound**: random context *raises* the composite (+0.028) while *not* raising coverage—which, if anything, trends down (−0.010; significant in paired/bootstrap tests, marginal in the mixed model). Generic exemplars make the model write more and score higher on a length-sensitive metric *without* recovering more relevant content—the cleanest demonstration of the confound. The decisive test for "does retrieval elicit capability" is therefore retrieval *minus random*, not retrieval minus base—and on that test, relevance-bearing retrieval (both modalities) wins on coverage. Second, the **curated ceiling**: oracle context is in a different regime entirely (coverage 0.226 vs. $\leq$ 0.090 for any retrieval). Because the oracle injects the gold record that coverage scores against, we read this as the bandwidth ceiling of *text* context, not a capability claim; the dominant practical lever is nonetheless *curation quality*, not retrieval mechanism. We therefore treat oracle throughout as a *calibration/ceiling* condition—a sanity bound on what any text context could supply—*not* a substantive arm comparable, apples-to-apples, with BM25 or CLIP retrieval. We also flag here, early, a construct caveat that recurs below: because coverage rewards reproducing the record's controlled vocabulary, part of the measured "relevance transfer" may be *vocabulary propagation* rather than improved visual understanding; the faithfulness analysis (§5) bounds how much.

5.3 Formatting does not transfer (multi-seed)

The title's structure question—does *organizing* retrieved context (explicit schema vs. free prose) help, once length is matched?—is answered in the negative. Table 3 shows the token-matched structured-vs-prose contrast across three seeds. A marginal trend at the first seed (+0.005, p=0.08) does not replicate: the effect is near-zero and non-significant at both additional seeds. The formatting lever, which was decisive *in the weights*, does not survive to inference-time context—and the single-seed trend that might have been reported as positive is exactly the kind of artifact this program has elsewhere cautioned against.

Arm	RC	Coverage	Coverage (length-resid.)
base	0.378	0.072	−0.017
random	0.406	0.062	−0.035
bm25-prose-matched	0.422	0.075	−0.023
bm25-structured	0.419	0.080	−0.017
bm25-prose	0.421	0.076	−0.023
clip-structured	0.420	0.090	−0.006
clip-prose	0.432	0.088	−0.010
oracle	**0.504**	**0.226**	+0.130

Table 1: Per-arm means, frozen model, n=100 images (≈400 image-question items). RC compresses the arms; coverage and its length-residual separate them into a monotone relevance/curation gradient. A *positive* length-residual means content above what response length alone would predict; a negative one, content below it.

Comparison	Mean Δ cov.	95% CI	p (item)	p (image)	Verdict
oracle − base	+0.155	[+0.139, +0.171]	< 0.001	< 0.001	curation dominates
clip − random	+0.028	[+0.022, +0.034]	< 0.001	< 0.001	visual relevance transfers
bm25 − random	+0.018	[+0.012, +0.024]	< 0.001	< 0.001	text relevance transfers
clip − bm25	+0.010	[+0.003, +0.017]	0.004	0.010	visual > text (n.s. after corr.)
random − base	−0.010	[−0.016, −0.004]	0.002	0.004	**priming: no recall gain**[†]

Table 2: Key paired comparisons on coverage (means; image-level CIs). The image level treats each image as one unit (n=100). With five contrasts the Bonferroni threshold is 0.010: every effect clears it except *clip − bm25*, which sits exactly at the edge—hence we treat the visual-over-text ordering as suggestive only. The length-residual column of Table 1 preserves the same ordering, confirming the effects are content, not verbosity. [†]The small random−base contrast is significant in paired and bootstrap tests but *marginal* (p=0.08) under the mixed-effects model; the metric-divergence conclusion does not depend on it (see §Statistics).

5.4 Image ablation: the model uses the image, but context crowds out vision

To test whether the apparent "relevance transfer" is genuine visual cognition or text-only completion, we re-ran each arm in a **blank-image** condition: the real artwork is replaced by a neutral gray placeholder (identical input shape, zero visual information), holding the retrieved context fixed. The visual contribution of an arm is its coverage drop when the image is removed (img − blank); the key quantity is the *interaction* $(\mathrm{X}_{\mathrm{img}} - \mathrm{X}_{\mathrm{blank}}) - (\mathrm{base}_{\mathrm{img}} - \mathrm{base}_{\mathrm{blank}})$, which asks whether retrieved context changes how much the image matters. Table 4 reports seed 42 (full five-arm matrix); seed 7 replicates every sign and significance.

The reading is unambiguous: the model is *not* blind—vision contributes in every arm—but retrieved text and the image are **partial substitutes**. The more (text) context the frozen model is given, the less it leans on the image; at the oracle limit (the target's own record in context) the image is redundant. "Relevance transfer" is therefore neither pure visual cognition nor pure text-parroting, but a measured middle in which context induces a quantified, significant shift away from visual reliance.

Seed	structured − prose-matched (mean Δ coverage)	p (Wilcoxon)
seed 42	+0.005	0.08
seed 7	+0.000	0.88
seed 123	+0.000	0.99

Table 3: Token-matched formatting contrast, replicated across three sampling seeds (p is item-level; the seed-42 image-level p is also non-significant, 0.13). The first-seed trend vanishes; structure does not transfer at inference time.

Arm	cov (image)	cov (blank)	visual Δ (img−blank)	interaction
base	0.072	0.030	+0.041 (p<0.001)	— (ref.)
random	0.062	0.055	+0.007 (p=0.002)	−0.034 (p<0.001)
bm25	0.080	0.065	+0.015 (p<0.001)	−0.027 (p<0.001)
clip	0.090	0.080	+0.010 (p=0.002)	−0.032 (p<0.001)
oracle	0.226	0.218	+0.009 (n.s., p=0.55)	−0.033 (p<0.001)

Table 4: Image-ablation matrix, coverage, seed 42 (n=100, narrative+3 question types; paired Wilcoxon). Removing the image cuts base coverage by ∼58% (0.072 → 0.030): the frozen model genuinely uses the image. But the image's marginal contribution shrinks 3–6× once retrieved context is present, and the negative interaction is significant for *every* arm; with the oracle answer-key the image adds nothing significant. seed 7 replicates (base +0.042, clip +0.011, random +0.007, oracle n.s.; interactions −0.032 to −0.036, all p<0.001).

5.5 Faithfulness: relevance transfer carries a precision tax

Coverage is recall and has no precision penalty, so it cannot see hallucination. We therefore scored **faithfulness**—the fraction of a response's concrete claims that are *grounded* in the verified gold record—with a three-lineage judge panel (a swarm of Claude judges, GPT-5.4, and Gemini-3.1-Pro), n=50 images/arm on the narrative question, seed 42 image condition, each judge blind to the arm and held to an identical strict rubric. Table 5 reports the result.

Arm	Claude (swarm)	GPT-5.4	Gemini-3.1	**Panel**
base	0.672	0.504	0.617	**0.598**
random	0.078	0.052	0.085	**0.072**
bm25	0.261	0.193	0.271	**0.242**
clip	0.307	0.199	0.289	**0.265**
oracle	0.894	0.833	0.795	**0.841**

Table 5: Faithfulness (claims grounded in the gold record / total claims), three-lineage panel. **All three lineages independently rank base > every retrieval arm.** Panel: base 0.598 → retrieval-mean 0.193 (Δ = −0.405). Retrieved context *collapses* groundedness—random context drives ∼93% of claims ungrounded; even relevant retrieval (bm25/clip) stays below half of base—while only exact curation (oracle) remains faithful.

This is the precision penalty that recall-only coverage structurally hides, and the panel is *unanimous*: not because LLM judges became reliable, but because the effect is large enough to survive cross-lineage scrutiny (cf. [6]). Combined with the coverage results, the full picture is that retrieved

context yields longer, higher-*recall*-looking answers whose added content is largely *hallucinated*—imported from the retrieved exemplars rather than read from the image. Relevance transfer is real but expensive; only curation is both high-recall and faithful. Figure 2 plots this trade-off directly.

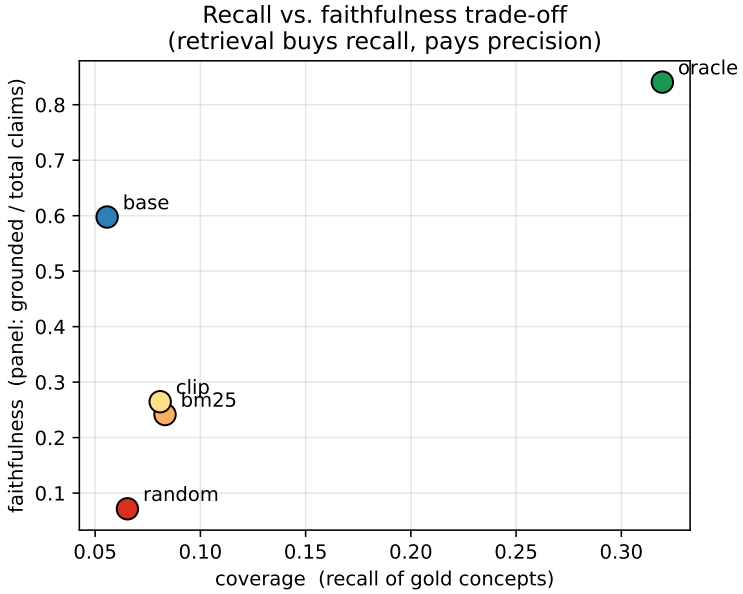

Figure 2: The recall–faithfulness trade-off (per arm; coverage on the narrative question vs. panel faithfulness). Retrieval (bm25/clip) buys recall over base but pays a steep precision cost; random context is low on both. The desirable **top-right region—high recall *and* high faithfulness—is reached only by *curation*** (the oracle, calibration); BM25/CLIP retrieval reach neither.

6 Discussion

On a frozen VLM, six things hold. (i) A large latent ceiling exists and is reachable only by *exact curation* (oracle +0.155 coverage); retrieval recovers a small fraction of it, and the ceiling is best understood as a text-bandwidth bound, not a capability bound. (ii) A meaningful slice of apparent "elicitation" is generic-context *priming* that a verbosity-sensitive metric rewards while content *declines*—a cautionary result for the common practice of scoring RAG with richness composites. (iii) *Relevance* does transfer at inference, on a content metric, for both text and visual retrieval; our data hint that visual similarity carries more relevance than keyword text for this visual task, but that cross-modality ordering is weak, does not survive multiple-comparison correction, and rests on a single seed, so we do not claim it. (iv) *Formatting* does *not* transfer: the structure lever that was decisive in training is null in context across three seeds. (v) The model *does* use the image, but retrieved text and vision are *partial substitutes*—added context significantly reduces visual reliance (the ablation interaction), so "relevance transfer" is a measured middle between seeing and parroting, not the former alone. (vi) That transfer carries a *precision tax* invisible to recall: faithfulness collapses under retrieval (base 0.60 → retrieval 0.19, unanimous across three judge lineages), because the model imports content from the exemplars; only exact curation is both high-recall and faithful.

The practical implication for memory systems is direct: invest in retrieval *relevance* (seman-

tic/visual) and especially in *curation* (the oracle-equivalent), do *not* expect gains from merely structuring retrieved context, and *measure groundedness, not just recall*—because a frozen VLM repays vague or irrelevant context with confident hallucination. Evaluate with a content-grounded, length-controlled, *precision-aware* metric—otherwise priming masquerades as capability and hallucination masquerades as recall.

Practical effect size. The retrieval gains are statistically clear but modest in absolute terms: relevance-bearing retrieval adds +0.018 to +0.028 coverage over random on a base of ∼0.07—a real relative lift, but small beside the curation jump (oracle +0.155) and far below the training-time density/structure gains in [1, 2]. Set against the precision tax on the other side of the ledger (faithfulness 0.60→0.19), a practitioner should ask whether a ∼0.02 recall gain justifies the added system complexity and the hallucination cost. Our reading for this setting: *curation* repays the investment; generic retrieval does not.

Out-of-distribution expectation. The Alexandria Aeternum corpus is unusually text-favorable: human curators have densely documented exactly these works (111-field NEST records), so text can substitute for vision more readily than it would elsewhere. In domains with sparser, noisier, or more informal retrieval pools—natural scenes, medical imaging, web documents—we hypothesize the two opposing effects move in *opposite* directions: the visual-substitution effect should *weaken* (there is less dense, on-target text to lean on), while the precision tax could *worsen* (noisier pools offer more off-target content to import and hallucinate). We therefore expect the qualitative pattern—verbosity confound, formatting null, and a precision cost to retrieval—to generalize, but the *magnitudes* to be domain-conditional; a second-domain replication is the natural next step.

7 Limitations

(1) Single model, and the main eight-arm matrix is single-seed at n=100; the curation, relevance, and priming effects survive image-level testing, but the *cross-modality* ordering (visual > text) is marginal (p=0.010) and single-seed—reported as suggestive only. The formatting contrast and the image ablation *are* multi-seed. (2) The four questions per image are non-independent; we address this with image-level tests, a cluster bootstrap, and a mixed-effects model (§Method), which agree except in down-weighting the small priming contrast. (3) The composite's 50-word depth cap is arbitrary, but the base/random reversal does not hinge on it: it is driven by a real word-count gap (base ≈ 48 vs. random ≈ 60) that would persist if the cap rose, and we also report the cap-independent length-residual. (4) Single, unusually text-favorable domain (cultural-heritage art); magnitudes—and possibly the direction—of the oracle/retrieval effects may be domain-conditional, and out-of-distribution transfer is not claimed (see the OOD hypothesis in §Discussion). (5) **Construct validity of coverage.** Coverage is *recall* against one controlled-vocabulary record with no precision penalty, so it rewards echoing the record's jargon, and part of "relevance transfer" may be vocabulary propagation rather than visual understanding. We address this directly—the image ablation (§4) shows genuine but context-diminished visual use, and the faithfulness panel (§5) supplies the precision axis—with two residual caveats: the ablation's blank-gray image is itself out-of-distribution (a no-image, noise, or wrong-image control would be cleaner), and faithfulness is text-grounded (n=50, one question, one seed; a pixel-grounded judge and a human-validated subsample would strengthen it). An embedding-normalized coverage and a second domain are the principal remaining future work. (6) The oracle injects the record coverage scores against, so its +0.155 is a metric upper bound (auto-recall under paraphrase), not a capability bound.

8 Conclusion

The density/structure imperative does not transfer cleanly to inference-time retrieval as a *formatting* effect. What transfers is *relevance*—atop a generic priming floor that a verbosity-confounded metric mistakes for capability, and beneath a high curated-context ceiling. The headline is as much methodological as empirical: on a frozen model, the metric you choose decides whether you conclude that retrieval elicits capability or that it mostly elicits words—and, under a precision-aware metric, words that are largely hallucinated. The image is used but substitutable; relevance transfers but at a precision cost; curation alone is both faithful and high-recall. For practitioners the takeaway is blunt: *invest in curation; generic retrieval does not repay its precision cost*—and measure grounding, not just recall, or the gain you see may not be there.

Reproducibility. Frozen-model inference config, the eight-arm harness, the run-time pHash exclusion, retrieval logs, the inherited composite scorer, the offline length-control re-scorer (coverage + length-residual + paired contrasts), the blank-image ablation harness and its interaction analysis, and the three-lineage faithfulness panel (rubric + per-judge scores) are released; corpus and held-out eval set under CC BY 4.0.

A Per-question-type coverage

Table 6 disaggregates coverage by question type (seed 42, image condition). The descriptive types (focal-point, narrative) carry most of the signal; *mental-state* inference is near the floor even for the oracle (0.101 with the answer key), because affective content is sparsely represented in the metadata record—a reminder that concept-recall grades the descriptive question types more validly than the interpretive ones.

Arm	narrative	mental-state	focal-point	causal-depth
base	0.056	0.043	0.132	0.056
random	0.065	0.029	0.109	0.043
bm25	0.083	0.046	0.126	0.062
clip	0.081	0.053	0.152	0.072
oracle	0.319	0.101	0.297	0.187

Table 6: Per-question-type mean coverage (seed 42, image condition).

B Methods detail

Retrieval. BM25 (k_1=1.5, b=0.75) over shallow catalog text (title/creator/date/department); CLIP ViT-B/32 image-similarity over the corpus; k=3 exemplars, position-shuffled. Context is the top-k exemplars' deep NEST fields, rendered structured or as length-matched prose. **Decoding.** `do_sample=True`, temperature 0.3, max 320 new tokens; the ablation adds two seeds (42, 7). **Coverage** is concept recall against the eval record's deep NEST fields; **length-residual** pools an OLS of coverage on word-count. **CIs** are image-level (normal-approx) and confirmed by a 5,000-resample image-level cluster bootstrap. **Faithfulness panel.** Each response is scored by a swarm of 10 Claude judges plus GPT-5.4 and Gemini-3.1-Pro, blind to arm, under an identical rubric:

extract the concrete claims, mark each supported/unsupported against the gold record, faithfulness = supported/total; the panel value is the mean of the three lineage means. Exact prompts, judge rubric, per-arm token distributions, retrieval logs, the ablation harness, and the re-scorers are released with the code.

References

[1] T. MacPherson. The Density Imperative: How Semantically Dense Metadata Reshapes Foundation Model Behavior. *Zenodo*, 2026. `10.5281/zenodo.18667735`.

[2] T. MacPherson. The Density Imperative, Refined: A 2×2 Ablation Reveals Independent Sufficiency of Density and Structure in Vision–Language Fine-Tuning. *Zenodo*, 2026. `10.5281/zenodo.20162589`.

[3] T. MacPherson. The Entropy of Recursion: Provenance-Verified Data and the Preservation of Cognitive Integrity. *Zenodo*, 2026. `10.5281/zenodo.18436975`.

[4] Y. Dubois, B. Galambosi, P. Liang, T. B. Hashimoto. Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators. *arXiv:2404.04475*, 2024.

[5] P. Singhal, T. Goyal, J. Xu, G. Durrett. A Long Way to Go: Investigating Length Correlations in RLHF. *arXiv:2310.03716*, 2023.

[6] L. Zheng et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *NeurIPS Datasets & Benchmarks*, 2023.

[7] P. Wang et al. Large Language Models are not Fair Evaluators. *arXiv:2305.17926*, 2023.

[8] K. Saito, A. Wachi, K. Wataoka, Y. Akimoto. Verbosity Bias in Preference Labeling by Large Language Models. *arXiv:2310.10076*, 2023.

[9] O. Ovadia, M. Brief, M. Mishaeli, O. Elisha. Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs. *EMNLP*, 2024.

[10] P. Lewis et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *NeurIPS*, 2020.

[11] U. Khandelwal, O. Levy, D. Jurafsky, L. Zettlemoyer, M. Lewis. Generalization through Memorization: Nearest Neighbor Language Models. *ICLR*, 2020.

[12] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, W. Chen. What Makes Good In-Context Examples for GPT-3? *arXiv:2101.06804*, 2021.

[13] O. Rubin, J. Herzig, J. Berant. Learning to Retrieve Prompts for In-Context Learning. *NAACL*, 2022.

[14] Z. Yang et al. An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA. *AAAI*, 2022.

[15] Y. Zhang, K. Zhou, Z. Liu. What Makes Good Examples for Visual In-Context Learning? *NeurIPS*, 2023.

[16] J.-B. Alayrac et al. Flamingo: a Visual Language Model for Few-Shot Learning. *NeurIPS*, 2022.

[17] S. Es, J. James, L. Espinosa-Anke, S. Schockaert. RAGAS: Automated Evaluation of Retrieval Augmented Generation. *arXiv:2309.15217*, 2023.

[18] J. Saad-Falcon, O. Khattab, C. Potts, M. Zaharia. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. *arXiv:2311.09476*, 2023.

[19] P. Anderson, B. Fernando, M. Johnson, S. Gould. SPICE: Semantic Propositional Image Caption Evaluation. *ECCV*, 2016.