RT-2026-04 · ACTIVE

Density over Volume: Validation Density as the Binding Constraint on Specialist Post-Training

DRAFT An open research track in progress. It sets out a question we are working on inside a live evidentiary environment, reports what has been measured so far, and names what remains untested. Figures will move as experiments run. We are sharing it to find collaborators, not to present a final conclusion.

Evidence: Measured Partner: Training compute · experiment ready

Specialist post-training usually scales corpus volume. We argue the binding constraint is validation density — expert-adjudicated claims, per-entity analyst flags, cross-model consensus. Ablation-guided layer pruning healed on a dense snapshot gained precision in every fold at a third fewer parameters, measured against distant-supervision labels rather than sealed expert gold. We also seal the benchmark generator rather than a fixed list.

Updated August 12, 2026

> **An open research track in progress.** It sets out a question we are working on inside a live evidentiary environment, reports what has been measured so far, and names what remains untested. Figures will move as experiments run. We are sharing it to find collaborators, not to present a final conclusion.

A production investigative platform emits extraction exhaust continuously. Every model call over every document produces spans, entities, and proposed relations, and almost none of it is ever looked at by a person. The standard move when a specialist model underperforms is to collect more of that exhaust. In a domain where the text is privileged, the analysts are few, and the failure mode is a confident wrong claim rather than a missed one, that move does not work. You cannot buy more corpus. What you can buy — expensively, one decision at a time — is validation.

The argument

We claim the rate limiter on specialist post-training is validation density: the fraction of training examples carrying a judgement from someone qualified to make it. Density is not annotation volume. It is the ratio of

adjudicated claims to emitted claims, and it is bounded by terminal analyst decisions, not by instrumentation. Our platform is heavily instrumented; the observation spine records far more than has ever been adjudicated. The gap between those two figures is the thesis.

Three things follow if that is right. Per-example value should be higher for validated claims than for volume-matched exhaust at equal token cost. The effect should concentrate on adversarial-prior cases — relations a model is primed to expect but the text does not support — because that is where unvalidated exhaust encodes the prior rather than the evidence. And dense-trained models should churn less between runs.

The second contribution is methodological and may outlast the first. A held-out benchmark that is a fixed list of documents can always be accused of selection, so we seal the *generator* instead: the eligibility snapshot is content-hashed, draws are seeded and stratified, and the sampler is deterministic — same disk state plus same seed yields the same draw hash. The invariance claim is cross-draw F1 stability within 0.05. No cherry-picked benchmark can make that claim, because it has only one draw.

The sampler is deliberately database-offline, reading per-version surfaces from disk, so a reviewer can reproduce a draw without reproducing our infrastructure.

What we measured

The first loop of the flywheel closed on a date-extraction slice. Ablation-guided layer pruning took an open-source span-extraction encoder from twenty-four layers to twelve, and a full fine-tune healed the result on a snapshot whose positives were cross-model consensus plus analyst high-impact flags. We are deliberate about the terminology: there is no teacher, no soft targets, no logit or hidden-state matching. This is structured pruning with recovery training, not knowledge distillation.

Layers were kept by single-layer knockout: each encoder layer was ablated in isolation on a small probe set and the twelve whose removal cost at least 0.02 date recall were retained. The tail layer alone accounts for −0.69 recall when dropped. Single-layer effects are not additive, which is why the pruned stack is healed rather than used directly — and the criterion is a selection heuristic applied faithfully, not a demonstration that the kept layers are causally necessary.

One design detail a reader should have before the numbers. Selection ran once, outside the fold loop, on a probe drawn from the same corpus later used for cross-validation. Selection was therefore not nested inside each fold, and probe documents could appear in an evaluation fold. We name that as selection leakage rather than argue it is negligible: it makes the compression figures exploratory on that axis, and the correct design — running layer selection inside the training side of each fold — is an outstanding item rather than a completed one.

Under five-fold cross-validation, **precision is the result and it is robust**: +0.131 ± 0.041 across folds, improving in five of five, pooled +0.128, 95% CI [+0.080, +0.183]. Garbage predictions fell by the same margin. The model is **−31% parameters**, **−35% encoder** and **−38% end-to-end latency**, and it emits **35.8% fewer predictions** (10,378 → 6,663).

Recall we cannot resolve. It moved −0.046 ± 0.100 across folds and −0.075 pooled, improving in one fold and falling in four; the 95% CI at five folds is [−0.170, +0.078]. That interval contains zero, contains the pooled estimate, and contains a loss twice as large — so this design cannot distinguish a substantial recall cost from none. F1 is in the same position (+0.078 ± 0.078 across folds, +0.056 pooled, CI [−0.019, +0.174]). We state this rather than pick the reading we prefer: **the study is underpowered on recall, which is not the same as recall being unaffected.** Precision, garbage-rate and F1 deltas are absolute; over-generation and latency are relative.

Intervals throughout are Student-t intervals over the five per-fold deltas (df = 4). We report them because they are the conventional summary and we would rather show the spread than a bare mean, with two caveats worth stating: five folds are not five independent experiments, because k-fold training sets overlap by construction, and our folds are deliberately unequal in size — so a t-interval over fold means understates that dependence and is sensitive to the largest fold. A document-set bootstrap would be the better instrument here and is not yet run.

A follow-up settles what the underpowered estimate could not, and the evidence points away from the pruned architecture as the primary cause of the recall cost.

Three architectural interventions move it nowhere. Re-selecting which twelve layers to keep — a middle-weighted list rather than the ablation-derived set — gives −0.044 against the original −0.046. Dropping one contiguous block instead of a scattered set gives −0.048, which is its own small finding: the layer-knockout selection procedure bought essentially nothing over a naive cut. Training twice as long gives −0.038, so this is not an under-training artifact. Across three seeds the pruning baseline lands at −0.046, −0.057 and −0.010 — direction stable, magnitude seed-noisy, no seed recovering.

One change to the loss moves it a great deal. Holding the prune and the folds fixed and down-weighting the implicit negatives — the spans nobody proposed, until now trained as hard negatives — gives **Δrecall +0.079, improving in five of five folds**. Compared fold-by-fold against the pruning arm on the same splits, that is **+0.125, 95% CI [+0.046, +0.205]**, and unlike every figure above it is resolved at this fold count. Pooled recall reaches 0.679, above the full-depth baseline's 0.596.

It is not a free lunch and we will not present it as one. The recovery costs roughly half the precision gain (+0.069 against +0.131), so this is a **different operating point, not a better model**, and the weight was one untuned value rather than a swept frontier.

The uncomfortable implication is the one worth stating plainly. If the negative weight is a dial, the precision gain this track opened with was also a reading of that dial. We set out to measure a property of compression and measured, in part, a property of our own loss function. The distinction that turns out to matter — between a span an expert rejected and a span nobody proposed — is the subject of our track on negative-space supervision, and this is the first measured evidence for it.

Two facts about the evaluation matter more than the point estimates. First, a single held-out split had shown recall down only 0.020; the full sweep put the pooled figure at 0.075 — **single-split evaluation understated the recall degradation by 3.75×**, and the "recall held within noise" conclusion we drew from that split was an artifact of which documents landed in it. Second, the folds are unbalanced by construction, because

splitting is by document set rather than by document: test-set sizes run from 110 to 590, the largest fold carries 43% of the evaluation and also the worst recall delta, and that is precisely why the pooled and per-fold recall figures diverge. Neither weighting is wrong; choosing between them after seeing both would be.

Resolving recall needs repeated cross-validation, not a larger claim. On our own variance, roughly three repeats of the five-fold sweep would give 80% power against an effect the size of the pooled estimate — a few hours of compute, and the experiment we would run before asserting a trade in either direction.

Two measurement-artifact controls belong in the same table as the results, and both are reported as past corrections rather than standing metrics. An output-token cap silently truncated roughly **29%** of one extractor's valid output, so any recall number taken before that cap was raised was deflated by a serving parameter rather than by the model. And a fleet-wide confabulation rate first reported at **5–13%** resolved to **0–1%** once measured at entity level against the model's own prompt substrate — a before-and-after of our measurement method, not a model improvement, and not a number we would quote as current without re-running it. Relatedly, **94%** of flagged date confabulations proved to be normalization artifacts: the same normalization that manufactures false consensus also manufactures false error.

The size of the corpus these results were drawn from is shared under NDA.

What this does not show

Every recorded metric above is measured against **distant-supervision consensus labels**, not a sealed hand-labelled gold set. For the date slice specifically, exact cross-model consensus is **zero, reproduced corpus-wide** — no two distinct models ever emit the identical date string — so the "consensus" behind our positives exists only after normalization. That is known-flawed for exactly the slice the flagship result sits on, and we would rather publish the mechanism than the comfortable version of it. The precision gain is unanimous across folds and we believe it; the recall cost is not adjudicable until a sealed expert set exists, and the two corpora must never share source material.

No sealed draw has been scored. Contamination control belongs downstream of the sampler and is specified there — but the replacement is not yet implemented. We originally placed a blanket exclusion at the benchmark layer, barring every unit that had appeared in any training fold. We retired it: a single shared pool serves several training and evaluation pipelines, one blanket exclusion cannot serve all of them, and it stripped the richest units out of the pool. The intended replacement is per-consumer exclusion — each training recipe or scorer excluding against its own manifest, at the pipeline stage rather than the benchmark stage.

That replacement is specified and not yet built, and we removed the old gate before building it. So the position today is worse than the architecture describes: there is no benchmark-level contamination guarantee, the per-pipeline guarantee does not exist yet, and a pool reporting zero contaminated units is reporting that trivially, because the exclusion set is empty. Nothing here may be described as contamination-controlled — by us, or by anyone citing us — until a draw is sealed and scored against a real manifest. We would rather publish that sequencing mistake than a diagram of the intended design. To keep the mechanisms distinct, because three exist and only one ever ran as code: the *legacy document-set seal gate*

— refuse-to-seal while the exclusion list was empty — is retired outright, not merely disarmed; the *per-consumer manifest exclusion* replacing it is specified and unimplemented; and *trace-level split control* is specification-only.

One of the three stratification dimensions was, until recently, near-degenerate: it keyed on a single dominant doctrine category per unit, and almost nothing resolved to one. It now takes the full set of tagged disciplines rather than a single winner, which should populate it for most units — units carrying few relations remain a known residual case, and the effect on effective stratum coverage has not been re-measured. We would not yet describe a draw as genuinely three-dimensional.

Pool formation turns out to be gated by instrumentation coverage rather than by labeling effort: the eligible pool can grow substantially without a single new label being applied, purely as upstream capture improves. A density argument therefore has to cost pool formation separately from annotation, or it will attribute an infrastructure win to human expertise.

Finally, inter-annotator agreement is not yet computable — and the reason is instructive. Historically, every human action collapsed to one generic identity across both the relational store and the graph. Doctrine authoring now captures analyst identity, and attribution is being threaded through flags, edge adjudications, integrity findings and the remaining mutation paths; historical records cannot be reconstructed retroactively. The limitation is missing attribution at write time, not missing annotators.

Relationship to prior work

That curation and quality can beat raw scale in instruction tuning is by now well established — self-guided selection, difficulty- and diversity-aware selection, and quality-focused reasoning data all report it, and a 2025 ACL short paper argues specifically that the field under-reports how data quality was established in the first place. We are not claiming quality-over-quantity as a novelty.

Our narrower claim is about *who* does the validating and *what fraction* of the corpus carries it. The selection literature largely scores examples with a model, a heuristic, or a reward proxy. Validation density asks what happens when the signal is a terminal decision by a domain expert whose judgement the model is trying to approximate, on privileged text that cannot be crowdsourced — and treats the adjudicated-to-emitted ratio, not the corpus size, as the quantity to optimize. The volume-matched control below is the experiment that would separate the two.

- Moon et al., *Call for Rigor in Reporting Quality of Instruction Tuning Data* — ACL 2025 (arXiv:2503.04807)
- Li et al., *From Quantity to Quality: Boosting LLM Performance with Self-Guided Data Selection for Instruction Tuning* — NAACL 2024 (arXiv:2308.12032)
- Chen et al., *MIG: Automatic Data Selection for Instruction Tuning by Maximizing Information Gain in Semantic Space* — Findings of ACL 2025 (arXiv:2504.13835)
- *Structure Trumps Size: Rethinking Data Quality for LLM Reasoning* — Findings of EMNLP 2025

Where this is going

Seal two independent draws and publish the cross-draw stability figure; that one number is the methodological credential this argument needs. Sweep the negative weight to map the precision–recall frontier properly, rather than reporting two points on it. Nest layer selection inside each fold so the compression figures stop being exploratory on that axis. Add a second extraction task, because one task, one architecture and one label scheme is not a general claim about validation density. Then run the comparison nobody has run: the volume-matched control — the same architecture and folds, trained once on expert-adjudicated claims and once on unvalidated exhaust.

About RedTorch

RedTorch, Inc.® is an investigative intelligence and applied AI firm, operating continuously since 2016 from Inglewood, California under California Bureau of Security and Investigative Services private investigator license 189236. Two practices run in parallel — investigative services (due diligence, litigation support, digital forensics, fraud investigation, asset recovery) and AI and strategy consulting — across more than twenty-five service lines. Casework runs on a platform we built, on GPU infrastructure we own. Privileged material never leaves our custody, and provenance, sanitisation and regulatory mapping are built into the platform's data model and operating workflows rather than added as a reporting layer.

RedTorch's platform is a closed-loop environment for evaluating, improving, and governing models on real evidentiary work. Evidence is captured with its lineage intact; machine claims are checked against their sources; analyst acceptances *and* rejections are preserved as supervision; the case graph rather than embedding similarity decides what a model reads; competence is measured from production exhaust; and the autonomy a model is granted is gated on that measurement. Each track in this collection documents one stage of that loop. We are building for systems that reason today and act tomorrow — a harder standard than benchmark performance, and a more useful one.

The corpus is what makes the research possible. It combines real privileged investigative text, an industry taxonomy, model-generated claims at scale, and a growing adjudication layer — analyst acceptances, dismissals, ruled-out hypotheses, doctrine assignments and integrity findings. Coverage and attribution vary by signal, and several of those gaps are themselves subjects of these tracks. What cannot be assembled after the fact is the pairing: a decade of privileged casework alongside the machine output produced against it and the expert judgement passed on that output. Corpus metrics are shared under NDA. This track works on the adjudication layer of that corpus — the part that cannot be grown by collecting more documents.

Collaborate

The unrun experiment is a token- and example-matched head-to-head — the same architecture and the same folds, trained once on expert-adjudicated claims and once on unvalidated extraction exhaust, scored on grounded precision over adversarial priors. RedTorch, Inc.® supplies the density: adjudicated claims, mechanism-typed flags, a deterministic sampler, and an existing cross-validation harness that has already produced a repeatable baseline. What we lack is training compute and a trainer service. Researchers and labs interested in this work can get in touch.