RT-2026-05 · ACTIVE

Negative-Space Supervision: Hard Negatives from Ruled-Out Investigative Hypotheses

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

Investigative work produces a supervision signal most pipelines throw away: hypotheses experts actively ruled out. We argue a negative's training value is its plausibility at the moment of rejection rather than its wrongness, and that mechanism-typed negatives outperform untyped ones at equal count. Confabulation becomes self-adversarial data. This is a position and experiment-design track — no training run has consumed a negative yet.

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.

An analyst spends most of a case ruling things out. A proposed link between two entities is examined and dismissed. An extracted date is flagged as unsupported by the page it supposedly came from. A line of inquiry is closed because the premise was wrong. Almost every extraction pipeline in production keeps the accepted links, writes the rejections to an audit log, and trains on the accepted set. The rejections are the more expensive signal: they cost the same expert attention and carry additional information, because a rejection is a judgement about something the system already found plausible enough to propose.

The argument

We treat the rejection stream as a first-class corpus with its own schema. Every negative normalizes to one record shape regardless of source: a granularity, a kind, a mechanism, a weight, a field naming which

objective the example is negative *for*, and a provenance origin that keeps derived and synthetic negatives from passing as primary ones. Eight kinds across five granularities. The trainable unit is not the bad content — it is context, rejected output, and reason for rejection.

The first claim is about weighting. **A negative's training value is its plausibility at the moment of rejection, not its wrongness.** A fabricated relation that survived cross-model consensus and looked properly grounded is a strong negative. An obviously wrong one teaches almost nothing. This inverts the usual instinct to weight negatives by confidence in the rejection.

We are not first to this principle, and the convergence is worth stating rather than eliding. Di et al. (2026) reach it independently for mathematical reasoning, arguing that "different errors carry fundamentally distinct learning values" and that samples lying close to correct solutions expose subtle weaknesses most effectively. Their route is synthesis: plausible negatives generated by reverse reinforcement learning against a composite reward. Ours is harvest — plausible negatives that a production system actually emitted and a domain expert actually rejected, with the failure mechanism attached at rejection time. Whether *adjudicated* plausibility carries the same training value as *synthesized* plausibility is an open and, we think, interesting question; that two independent lines arrived at the same weighting principle from opposite directions is the main reason we believe it.

The second follows from the first. If plausibility is what matters, then the *mechanism* of failure has to travel with the label, because "wrong" is uninformative while "compositional confabulation — two ungrounded spans fused into one claim" names a failure mode a model can learn to avoid. We predict mechanism-typed negatives beat count-matched untyped negatives.

Together these make confabulation into **self-adversarial training data**: the system produces its own hardest candidate negatives as an operational byproduct, a claim-integrity engine types them, and expert adjudication supplies the part that is never free — the rejection itself. A disjoint-root provenance guard stops a confabulation-derived negative from leaking back in as positive gold.

There is also a granularity above the edge that almost no one captures: the case. A case terminated because the client misrepresented the objective is a labeled adversarial premise — arguably the strongest negative an investigative firm holds, and one that trains something edge-level negatives cannot: premise detection.

What we measured

The state of play, stated up front: **no training run has yet consumed an adjudicated negative.** The two theses above are arguments with a corpus behind them, not results. What has been measured is the prerequisite — and it came out well enough to be worth the rest.

The signal is now at least countable. In the adjudication spine, 133 distinct relations carry an adverse analyst judgement — 125 dismissed, 8 ruled out — a small adverse minority against a much larger accepted pool. That is a real signal from real casework and it is small; the point is that it moved off zero and can be measured, not that it constitutes a training set. The corpus has now been materialized — entity-level negatives, confabulation findings, doctrine-failure tags and analyst notes — with two of the eight declared

kinds still marked pending by the materializer itself, which is the honest-partial-artifact discipline this track argues for. It has still not been *consumed* by a preference trainer, so the two theses above remain untested.

What exists is the baseline it must beat. A positives-only heal on the same architecture and folds produced a robust precision gain — +0.131 ± 0.041, five of five folds — while emitting 35.8% fewer predictions. Recall moved −0.046 ± 0.100 across folds and −0.075 pooled, which at five folds is not resolvable in either direction. Details, including why we decline to call that a trade, are in our track on validation density.

The distinction this track is built on has now been measured, on our own training data.

The compression run that produced that baseline lost recall, and the obvious suspects were all architectural. They were all wrong. Re-selecting which layers to keep changes nothing (−0.044). A contiguous cut instead of a scattered one changes nothing (−0.048). Training twice as long changes nothing (−0.038). A hidden-state distillation arm against the full-depth model as teacher changes nothing — paired per fold, +0.0002 ± 0.009 — because matching representations leaves the training signal that suppresses recall completely intact.

What moved it was the label scheme. On the same prune and the same folds, **down-weighting the implicit negatives** — the spans nobody proposed, until then trained as hard negatives — turned the recall delta from −0.046 to **+0.079, improving in five of five folds**. Paired against the pruning arm on identical splits that is **+0.125, 95% CI [+0.046, +0.205]**; pooled recall reaches 0.679, above the full-depth baseline. It costs about half the precision gain, so it is a different operating point rather than a strictly better model.

We want to be exact about what that shows and what it does not. **Treating "nobody proposed it" as equivalent to an explicit hard negative costs measurable recall.** That is the demonstrated result. Whether *expert-rejected* negatives carry still more useful supervision than merely down-weighting the unproposed is the next experiment — no training run has yet touched the adjudicated rejection set. Nor does this speak to *plausibility* or *mechanism*, the two theses this track actually advances. What it establishes is that the axis is real, that the distinction has a measurable price, and that the instrumentation to exploit it is worth building.

The mechanism behind the asymmetry is what makes the negatives arm worth running. The positives-only heal bought precision largely by *emitting less* — over-generation fell by more than a third — which is suppression, not discrimination. A model taught only what is correct becomes conservative near the decision boundary. Explicit negatives should move the boundary instead of retreating from it, which is why we predict recall recovers rather than falls further. That prediction is pre-specified with a numeric bar: **recover at least half of the 0.075 recall loss while retaining at least 80% of the 0.128 precision gain**, per fold across five folds. We will report the precision–recall frontier — PR-AUC, precision at fixed recall, recall at fixed precision — rather than the two endpoints alone, because a frontier shift is the claim and the endpoints are only where it happens to be sampled.

On magnitude we are deliberately silent. The literature supports the direction — Hamdan and Yuret (2025) report a substantially larger improvement per training example from negatives than from supervised fine-tuning, with near-misses exerting the greater influence — but we know of no defensible general multiplier and will not invent one.

Per-stack ratio bounds are pre-specified rather than tuned: **0.50** negatives for the span-extraction encoder, **0.15 / 0.20 / 0.25** swept for the supervised fine-tuning stack, and a preference-optimization weight invariant held in **[1.0, 1.33]**.

Two mechanism findings support the plausibility thesis. In one fully traced case, **one of eight models** fabricated a relation and cross-model dissent flagged it — dissent is a detector, not only a filter. And the pipeline amplified its own error: a fabrication first emitted at **0.60** confidence re-entered the next run's prompt and re-emerged at **0.85**, so cross-run reproducibility was partly manufactured by the loop. Separately, **94%** of flagged date confabulations were normalization artifacts, which is why provenance routing has to precede any training on this pool.

Pool size is shared under NDA.

What this does not show

No training run has consumed an adjudicated negative. The corpus has been materialized — and the materializer declares its own incomplete kinds on every output, which is the honest-partial-artifact discipline this track argues for — but nothing downstream reads it yet. Both theses above remain prospective.

The weighting thesis is **uncalibrated**. We do not know whether confidence at rejection predicts training value; that is the experiment, not the finding. The weight must also be computed only from signals recorded *before* the analyst rules on the item — model confidence, cross-model agreement, grounding score at emission. Any component derived from the adjudication itself would leak the target into the label, and we would rather define the weight narrowly than defend it later.

Negative selection is known to be delicate. False negatives and negatives that are hard for the wrong reason degrade training rather than improve it, which is the strongest argument for carrying the mechanism as a label and routing on it: an operator-defect negative and a genuine near-miss should not enter the same objective.

Rejection rationale is empirically absent, and the anatomy of the absence is worth stating precisely. The rejection write handler accepts an optional free-text note and would persist it into the event stream — but no interface sends one, so no persisted feedback event carries one, and no structured, queryable column exists for it anywhere. Reason-aware supervision is unavailable until that path exists end to end, so the hypothesis that phrasing-with-reason beats a naked reject label is untestable on our data as it stands.

The persisted dissent signal is lower fidelity than the interactive one: it records the confidence-floor rule but not the stance or relative-outlier rules the interface applies, so any statistic drawn from it undercounts reasoning-based dissent.

Mechanism classes are detector-dependent, and we found out the hard way. A corpus-wide re-scan under a new detector version reclassified the typed pool and eliminated two of the four classes outright; what survives is pure-confabulation and extracontextual-assertion findings spread across dozens of source sets, so the earlier single-source concentration is gone. This matters more than a stale caveat: the thesis that

typed negatives beat untyped ones presumes the type is a property of the failure, and a detector bump that zeroes two classes is evidence it is partly a property of the detector. Any test of that thesis has to pin the detector version, and cross-version comparability is an open problem we do not have a method for.

Inter-annotator agreement is not yet computable, and the reason is a systems fact rather than a staffing one. Historically, every human action collapsed to one generic identity in 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. Until that threading completes, most rejections cannot be checked against a second expert, because they are not attributable to a first.

Case-level negatives have no structured home: archived cases carry no machine-readable termination reason.

Relationship to prior work

Learning from rejected outputs is established, and the field has moved past treating all negatives as interchangeable. Hamdan and Yuret quantify how much a model gains per negative example relative to supervised fine-tuning; Di et al. show that plausible negatives outperform naively sampled ones, and that careless negative sampling can underperform the baseline it was meant to improve; a parallel retrieval literature documents the same hazard from the other side, where synthetic hard negatives cross a threshold and begin to hurt.

What we add is provenance rather than method. Almost all of that work constructs its negatives — sampled, synthesized, or mined. Ours are a byproduct of expert labour already being paid for: a domain specialist looking at real model output on privileged material and deciding it does not stand. That gives three things a constructed corpus cannot easily supply — the rejection is by someone qualified in the domain, the rejected item is one the deployed system genuinely produced, and the mechanism of failure is typed at the moment of rejection rather than inferred afterwards. It also brings a liability a constructed corpus does not have, which we state in full below.

- Hamdan & Yuret, *How much do LLMs learn from negative examples?* — arXiv:2503.14391
- Di et al., *Not All Negative Samples Are Equal: LLMs Learn Better from Plausible Reasoning* — arXiv:2602.03516
- *When Hard Negatives Hurt: Bridging the Generative–Discriminative Gap in Hard Negative Synthesis for Retrieval* — arXiv:2606.01304

Where this is going

Run the heal on the existing folds — one command against a harness that already produced the baseline. Then the two ablations that would settle it: mechanism-typed versus untyped at equal count, and high-plausibility versus low-plausibility at equal count. In parallel, wire click-time rejection capture, and give adverse case outcomes a structured vocabulary.

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 part of that judgement almost every pipeline discards: the rejections.

Collaborate

The specific unrun experiment is the negatives-aware heal against our positives-only baseline on identical folds, followed by the typed-versus-untyped ablation at matched negative count. RedTorch, Inc.® supplies adjudicated rejections with mechanism labels, cross-model dissent structure, provenance routing that excludes pipeline artifacts, and the cross-validation harness that produced the baseline. We lack training compute and a preference-optimization trainer. Researchers and labs interested in this work can get in touch.