Worked Examples Beat Task Descriptions

What distillation actually adds, and why the variance most people measure is the wrong one

Matt Liotta · August 2026

Abstract

Retrieving teacher-authored procedures at inference time is an established way to lift a small model: prior work builds a corpus of step-by-step guides from clustered training questions and retrieves them per query, reporting 5–9% gains across three domains (arXiv:2510.13935). We do not propose that architecture. We ablate it.

Those corpora are generated from *worked instances of the task*, and whether the instances are load-bearing has not been tested. Episodes are roughly 12× more expensive to distil from than a one-line task description, so if the description suffices, the examples are ceremony.

We find episodes are worth **+18.7pp** over the best description-only alternative (64.0% ± 3.3% vs 45.3% ± 7.8%, against a 38.7% ± 5.0% floor), a gap **3.1× the pooled generation noise**. Episodes also produce *more stable* artifacts — the description-only condition varies nearly twice as widely across generations, meaning a practitioner gets whichever procedure their single call happened to produce.

Reaching that result required correcting a methodological error we made twice, and which we believe is common. Our first two experiments generated **one** artifact and evaluated it three times, reporting `± 0.0%` error bars. Those bars described *evaluation* noise while the quantity that actually moved between experiments was *generation* noise. The two passes returned opposite answers (+0.0pp and +8.9pp for the same comparison). Measured properly — N independent generations, one evaluation each — artifact-generation variance (6.0pp) **exceeds** model-evaluation variance (5.0pp). Any evaluation of LLM-authored artifacts that repeats evaluation over a fixed artifact is measuring the smaller of two noise sources and reporting confident intervals about the wrong quantity.

Three secondary results, two of which qualify the prior work. A *generic* procedure actively harms (−8.9pp), so the benefit requires task-appropriateness rather than mere structure. Governance bypass rises from 0% to 13.3% for **both** artifact conditions equally — a real safety cost of the technique that attaches to having a procedure at all, and which the instruction-retrieval literature does not measure. And a capability-gradient hypothesis — that curation helps small models more, which the prior work's ">= 3B parameters" result invites — **failed**: of five models across three families, only one showed any premium, with no gradient.

Scope is narrow and load-bearing: one weak model, one task family, 15 evaluation timelines. We demonstrate that the episodes-versus-description gap **can** be large; we do not estimate how large it typically is, and §6.3 is our own evidence against extrapolating.

---

1. Introduction

Give a small model a procedure and it does better. That much is established (§2). The question with economic content is *where the procedure comes from*.

Two sources, differing by an order of magnitude in cost:

- **Distillation from worked episodes.** Show a frontier model N solved instances and ask for the transferable procedure. Cost: N episodes rendered into context, plus a generation call.
- **Description-only authoring.** Show the same frontier model a one-line description of the task and ask for the same thing. Cost: one short call.

If these produce equivalent procedures, the episode-gathering step in existing instruction-retrieval pipelines is theatre — nobody should pay 12× for it — and the framing of memory-as-transferable-experience loses its empirical support.

We found — twice, with confident-looking error bars — that episodes added nothing. Both results were artifacts of a sampling error described in §4, and both were wrong.

2. Related work

Instruction retrieval is not ours. Prior work already proposes the architecture this paper evaluates. *Big Reasoning with Small Models* (arXiv:2510.13935) constructs an "Instruction Corpus by clustering similar training questions and using a teacher model to generate generalizable guides that pair domain background with explicit step-by-step procedures", which the small model retrieves at inference "without any additional fine-tuning". It reports gains of 9.4%, 7.9% and 5.1% in medicine, law and mathematics for models of at least 3B parameters. That is the same mechanism we implement.

We are not proposing it; we are ablating it. Their corpus is built from clustered *training questions* — worked instances of the task. The question of whether the instances are load-bearing is not asked: there is no arm in which the teacher writes the same procedures from a task description alone. That ablation is this paper. Our answer (+18.7pp for episodes, §4) supports their design choice, and supplies the evidence for it that the original does not.

Two further gaps we address. Their gains hold "for models with at least 3B parameters", which invites a scale reading; we test five models across three families and find the benefit concentrated in a single 4B model with no gradient (§6.3). And neither that work nor the surrounding literature measures what the intervention costs in governance terms — we find a real one (§6.2).

Distillation. *Distilling Step-by-Step* (Hsieh et al., 2023) extracts natural-language rationales from a large model and trains a small one on them, and the broader rationale-distillation line follows that shape. The difference is where the transferred capability lives: in weights, versus in an inspectable artifact retrieved at inference. The artifact form is revocable, attributable, and scopeable — properties a weight update cannot offer — which is what makes the governance question in §6.2 askable at all.

Evaluation variance. Our §4 correction has a direct precedent. *ReliableEval* (arXiv:2505.22169) argues that "standard benchmarks typically report performance using a single prompt, raising concerns about the reliability of such evaluations", and proposes estimating how many prompt resamplings are needed for a meaningful result. Related work on reproducibility in reasoning evaluation makes the parallel point for random seeds, finding single-seed results on small datasets unstable.

Our case is a specific and sharper instance. ReliableEval resamples *meaning-preserving perturbations of a human-written prompt*: the perturbations are constructed, and the variation is introduced deliberately. When the artifact is **authored by a model**, the variation is not introduced — it is inherent, unavoidable, and easy to miss, because there is exactly one artifact and it looks like a fixed object. We show that treating it as fixed inverts a result (§3), and that its variance exceeds the model-sampling variance that evaluations do typically report (6.0pp vs 5.0pp).

Memory-evaluation confounds. *MemDelta* (arXiv:2606.29914) finds that reported gains in agent memory "often mix changes in the memory method with changes in the language model, embedding model, or retrieval pipeline", and that varying one component at a time can flip a conclusion. That is a confound in the *pipeline*; ours is a confound in the *artifact*. Both produce the same failure — a comparison whose sign depends on an uncontrolled variable — and a system can be exposed to both independently.

3. Setup

Task. `repair_propagation` from StateBench: a value is corrected, and conclusions derived from the old value must be recomputed rather than reused. Chosen because the weak model has genuine headroom there (≈39% unaided) rather than the ceiling it hits on other tracks, where no lift is measurable.

Models. Weak: `mlx/qwen3-4b:4bit`, local. Author/distiller: `gpt-5.6-sol`. Judge: `gpt-4o-mini`, pinned globally and never inherited from the system under test.

Contamination control. Distillation draws from the **train** split and evaluation from **dev**, so episode sets are disjoint by construction. A scrub pass then rejects any artifact containing an evaluation-set literal — entity, amount, or date — because templates reuse entity pools across splits, and disjoint episodes do not guarantee disjoint content. Without this, "transfer" could be answer leakage.

Conditions.

Condition	Artifact source
`weak_alone`	none
`distilled`	frontier model, 12 worked episodes
`description_only`	frontier model, one-line task description, no episodes
`generic` (§6.1)	hand-written, task-agnostic
`targeted` (§6.1)	hand-written, task-specific

4. The methodological correction

Our first two experiments each ran **three evaluation passes over one generated artifact** and reported the spread as an error bar. Results:

Pass	`distilled`	`description_only`	Conclusion drawn
A	53.3%	53.3% ± 0.0%	episodes add **+0.0pp**
B	55.6%	46.7% ± 0.0%	episodes add **+8.9pp**

Two passes, opposite conclusions, and both reported `± 0.0%` on the description-only arm.

That `± 0.0%` was the diagnostic and we initially read it as precision. Three evaluations of a *fixed* artifact measure only weak-model sampling. The artifact itself is drawn from a distribution, and it is the artifact that differed between passes: pass B's generation was simply worse.

The corrected design points the same compute at the right variance — **N independent generations, one evaluation each** — and compares distributions rather than point estimates. This is not a refinement. The naive design returned the wrong sign.

The general claim. For any evaluation of LLM-authored artifacts — prompts, procedures, plans, distilled skills, generated rubrics — the artifact is a random variable. Repeating *evaluation* while holding the artifact fixed produces error bars that describe the wrong quantity, and here that quantity was the *smaller* of the two (5.0pp evaluation vs 6.0pp generation). Confident intervals over the smaller source, applied to a comparison driven by the larger, is how both of our early passes reached the wrong answer with narrow bars.

5. Result

Five independent generations per condition, one evaluation each, 15 dev timelines.

Condition	Accuracy	Range	SFRR	GBR
`weak_alone`	38.7% ± 5.0%	33.3–46.7%	50.7%	**0.0%**
`distilled`	**64.0% ± 3.3%**	60.0–66.7%	69.3%	13.3%
`description_only`	45.3% ± 7.8%	40.0–60.0%	52.0%	13.3%

Episodes are worth +18.7pp, 3.1× the pooled generation standard deviation of 6.0pp. The distributions barely touch: distilled's worst generation (60.0%) equals description-only's best.

Episodes also stabilise. Description-only spreads ±7.8% against distilled's ±3.3% — roughly twice the variance. For a deployment this matters more than the mean: you get one generation, not a distribution, and the description-only condition's 20pp range means the procedure you happen to receive may be barely better than nothing.

That stability result is only visible under the §4 design. The naive design cannot see it at all, because it never generates a second artifact.

6. Secondary results

6.1 Structure is not enough; task-appropriateness is required

Two hand-written controls, three evaluation runs each:

Control	Accuracy	vs floor
generic ("answer from currently valid state…")	26.7% ± 5.4%	**−8.9pp**
targeted (repair-specific)	60.0% ± 0.0%	+24.4pp

A generic procedure is **worse than no procedure**. So the effect is not "text in the context helps" — a procedure must fit the task or it actively misleads. The generic one advises using the most recent value, which is close to wrong advice on a track that requires recomputation.

The targeted control scores well, but we authored it *after* seeing the distilled artifacts and treat it as contaminated in its own favour. It bounds what a domain expert might achieve; it does not establish it.

6.2 The governance cost is not distillation-specific

Governance Bypass Rate — asserting content the state layer excluded, scored against the engine's audit record with no judge — is **0.0% unaided and 13.3% for both artifact conditions, identically**.

So procedures cause a weak model to surface withheld state regardless of their provenance. A procedure instructing the model to trace corrections through their dependents evidently also induces

it to reach for state it should not have. This is a cost of the technique, not of distillation, and it is worth stating plainly: the accuracy gain comes with a real governance regression.

6.3 A failed hypothesis: no capability gradient

We predicted curation would help in inverse proportion to model capability, which would have given artifact transfer an economic motivation. Across five models and three families, matched subset, judge held constant:

Model	Family	Premium	runs
`qwen3-4b`	Qwen	**+16.3pp**	3
`qwen3-8b`	Qwen	−3.3pp	3
`gemma-4-12b`	Google	+3.3pp	3
`gpt-5.4-mini`	OpenAI	+0.7pp	3
`gpt-5.6-sol`	OpenAI	−0.0pp	3

Four of five sit within ±3.3pp of zero, and only `qwen3-4b` clears its own noise (5.6× pooled σ; the next largest is `gemma-4-12b` at 1.6×, which is not significant). **Only the 4B model shows a premium**, and a second small local model gets nothing. There is no gradient and no threshold — there is one outlier. Whether that reflects scale, that specific model, or an interaction with the context format is unresolved, and a single model cannot distinguish them.

The tiers also illustrate §3 from the other direction. Every one of them shrank toward zero as runs accumulated: `gemma-4-12b` read +5.9pp at n=1, +2.0pp at n=2 and +3.3pp at n=3; `qwen3-8b` read +3.9pp at n=1 and −3.3pp at n=3, reversing sign. Single-run tier results were uniformly more encouraging than replicated ones.

We report this because the hypothesis motivated the work, and because a single supporting datapoint would have been easy to present as a trend.

7. Discussion

7.1 What a practitioner should take from this

If you are building an instruction-retrieval or distilled-skill system, three things follow.

Gather the episodes. They are worth ~19 points over what the same teacher writes from a description, and the description-only shortcut is tempting precisely because it looks equivalent when you test it once. On this task it was not.

Generate the artifact more than once and keep the better one. Description-only spans 20 points across generations; distilled spans 7. Since generation is cheap relative to serving, sampling several

and selecting is close to free and recovers most of the downside. Nothing in the literature we surveyed does this, because nothing in it measures generation variance.

Do not assume a generic procedure is a safe default. A task-agnostic procedure scored *below* no procedure at all (§6.1). Shipping one "just in case" is a live way to make a system worse.

7.2 The safety cost is a property of the technique

Governance bypass rose identically for both artifact conditions (§6.2). This is not a distillation problem to be engineered away by better distillation — it attaches to putting a reasoning procedure in front of a weak model at all.

The mechanism is plausible on inspection: a procedure instructing a model to trace corrections through their dependents is also instructing it to go looking, and a model that goes looking finds state it was not meant to use. The intervention makes the model more capable *and* more acquisitive, and the accuracy headline reports only the first.

Anyone deploying this should measure the second with an instrument the procedure cannot influence. We used a metric scored against the state layer's own audit record, which requires no judge and is insensitive to how the answer is phrased. A phrase-list or judge-based safety metric would be scored on text the procedure helped write.

7.3 Why we report a failed hypothesis at length

§6.3 records a prediction that did not survive: that curation helps in inverse proportion to model capability. It motivated this work, and one supporting datapoint existed.

We report it because the failure mode is instructive rather than embarrassing. The hypothesis was plausible, the first datapoint confirmed it, and four subsequent models did not. Had we stopped at one model — as the compute budget invited — we would have published a clean gradient story built on a single observation. The same pattern recurred at every scale of this project: the description-only comparison inverted between passes, and every capability tier shrank toward zero as runs accumulated. Single-run results were uniformly more encouraging than replicated ones.

That regularity is worth more than the specific hypothesis. If it generalises — and §4 gives a mechanism for why it should, since the encouraging draw is the one that gets written up — then the appropriate prior on any single-run result in this literature is that it overstates.

8. Limitations

These are the paper's boundaries, not a research agenda. The two most consequential — a second task family and an uncontaminated expert control — were considered and deliberately not run. We state them as open rather than forthcoming.

A single task family. This is the binding limitation. Every headline number comes from `repair_propagation` on `qwen3-4b` over 15 timelines. We claim episodes beat descriptions *on this*

task, and nothing wider. §6.3 is our own evidence against extrapolating: a capability hypothesis that looked reasonable held for one model out of five. A reader should treat +18.7pp as a demonstration that the gap **can** be large, not an estimate of how large it typically is.

The expert-authored baseline is unresolved. §6.1's targeted control scores 60.0% — *above* distillation — but we wrote it after seeing the distilled artifacts, so it cannot be treated as independent. The question it was meant to answer, *could a domain expert match distillation without episodes?*, is therefore open. We consider this the strongest surviving deflationary account of our result: if the answer is yes, the contribution narrows from "episodes are necessary" to "episodes are a cheap substitute for expertise". Resolving it requires an author with no exposure to the distilled artifacts, and we did not run that.

Description quality is a confound we could not fully remove. `repair_propagation`'s description — "corrections cascade to derived conclusions" — nearly *is* the procedure. We attempted an impoverished-description arm, but the frontier model inferred the method anyway, producing guidance about "treating corrections as superseding earlier information" from a description that only said information changes over time. A task whose method cannot be inferred from any honest description would be a cleaner test; we did not find one.

Over-generalization is untested, and the attempt failed rather than being skipped. The intended control — applying skills to a family whose correct behavior conflicts — was degenerate: the weak model scores 0.0% on that family with or without skills, leaving no headroom for harm to appear. So we cannot say whether these artifacts misfire when applied where they should not. Given §6.1's finding that a *wrong* procedure is worse than none (−8.9pp), the potential for harm is real and unmeasured.

No cost measurement. The local inference path reports `usage: null`, so token accounting for the weak arm reads zero and the cost ratio is unavailable. The 12× generation-cost difference is by construction, not measured end-to-end.

Judge and distiller share a provider family, though not a model. A cross-family judge control was not run.

9. Conclusion

Worked examples are worth paying for. They produce procedures ~19 points better and about twice as stable as what the same model writes from a task description alone. Existing instruction-retrieval systems build their corpora from worked instances; this is the evidence that the choice matters, which those systems assert rather than demonstrate.

The methodological result may be the more portable one: when the thing under test is generated by a model, the generation is the experiment. We measured the wrong variance twice, with narrow error bars both times, and got opposite answers. The fix costs nothing — the same compute, redistributed — and it is the difference between a result and its negation.

Finally, the technique carries a governance cost that its accuracy numbers conceal. A procedure that helps a weak model reason also helps it reach for state it was not given. That cost is identical whether the procedure came from episodes or a description, and it should be measured with an instrument the procedure cannot talk its way past.

References

Hsieh, C.-Y., et al. (2023). *Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes.* arXiv:2305.02301.

Big Reasoning with Small Models: Instruction Retrieval at Inference Time. (2025). arXiv:2510.13935.

ReliableEval: A Recipe for Stochastic LLM Evaluation. (2025). arXiv:2505.22169.

A Sober Look at Progress in Language Model Reasoning: Pitfalls and Paths to Reproducibility. (2025). arXiv:2504.07086.

Wang, K. (2026). *MemDelta: Controlled Baselines and Hidden Confounds in Agent Memory Evaluation.* arXiv:2606.29914.

Wu, R., et al. (2026). *MemHarness: Memory Is Reconstructed, Not Replayed.* arXiv:2607.28272.

Liotta, M. (2026). *The Correct Answer Violates: Measurement Validity in Agent-Memory Evaluation.*

Appendix A: Reproduction

```
# The corrected design: N generations, one evaluation each
uv run python experiments/with_car_secret.py -- \
    python experiments/artifact_variance.py --samples 5 --limit 15

# Controls, three evaluation runs each
uv run python experiments/with_car_secret.py -- \
    python experiments/skill_transfer.py --track repair_propagation --runs 3
```

Results: `experiments/results/artifact_variance.json`, `skill_transfer_v2.json`, `capability_curve.json`.