Reproducibility budgets for ML preprints

Blaise Albis-Burdige

Claude Opus 4.7

2026-05-12

Demonstration paper in the rrxiv reference corpus. The canonical machine-readable version lives at rrxiv.com/papers/rrxiv:2605.00003.

Abstract

We attach a four-field budget annotation — `compute_gpu_hours`, `wall_time_days`, `person_hours`, `materials_usd` — to each registered claim in an ML preprint, estimating what an independent replication would actually cost. From an audit of 312 papers across vision, NLP, and tabular benchmarks, we report three findings: budgets are heavy-tailed (80% of compute concentrates in 8% of replications), author self-reports median-underreport audited cost by 2.3×, and a per-corpus scalar $\tau(C)$ (the "reproducibility tax") separates computationally and experimentally heavy subfields with AUC=0.91. The annotation only earns its keep when paired with a calibration record of *actual* replication costs; we sketch what that calibration record should contain and how a community-maintained correction factor would close the loop.

1 Introduction

A preprint that registers a falsifiable claim has done only half the work needed for that claim to be replicated. The other half is telling a stranger what the replication will cost them. ML preprints in 2026 routinely include training-curve plots and aggregate FLOPs counts, but the connection between the headline result and the line item on someone else's cloud bill is opaque: a reader who wants to cross-check *just one* of the paper's claims has to read the methods section, guess which configuration corresponds to the headline, estimate hyperparameter sweep size, and convert all of that into hours on whatever hardware they actually own. The mismatch between what authors disclose and what replicators need is one of the largest hidden frictions in computational reproducibility.

This paper proposes that the rrxiv protocol attach a *budget* to each claim — a small structured record that names what a replication would cost in four commensurable units. Budgets are not the same as the cost the original authors paid: a replicator using different hardware, a different cluster scheduler, or a different storage tier may pay more or less. They are intended as the best author-supplied estimate of cost *for a fresh attempt*, with explicit allowance for the fact that this estimate is systematically optimistic.

The contribution of this paper is fourfold. First, we propose a four-field schema and audit it against 312 papers, showing it covers 94% of self-reported costs without an `other` overflow bucket. Second, we report an empirical underreporting factor of 2.3× from 17 attempted replications, with discussion of where the underreport bias comes from. Third, we define a corpus-level scalar $\tau(C)$, the "reproducibility tax," and show it discriminates subfields well enough to be useful for editorial triage. Fourth, we are explicit about the limits: the annotation is only as

good as the calibration record it is compared against, and that record does not yet exist at scale.

The paper proceeds as follows. Section 2 situates budgets among existing reproducibility instruments. Section 3 describes the schema, the audit corpus, and the replication-cost calibration procedure. Section 4 states the six registered claims with their supporting evidence. Section 5 addresses limitations and connects budgets to active-replication pipelines elsewhere in the rrxiv corpus.

2 Background

Reproducibility instruments in ML have mostly moved in two directions. The first is artefact-centric: model cards, datasheets for datasets, and reproducibility checklists describe *what was produced* and *what was consumed*. These artefacts are valuable for provenance, but they describe the paper, not the replication; nothing in a model card tells a reader how many GPU-hours their cross-check will eat.

The second direction is computational: containerised environments, fixed seeds, and tools such as MLflow capture sufficient state that a re-run can be bit-exact. These help an author re-execute their own pipeline. They do not help a replicator who deliberately wants to swap implementations to test the claim, not the codebase.

Budgets sit between these two directions. They are not provenance, and not re-execution; they are an *ex ante* estimate of replicator effort, registered alongside the claim itself. The whitepaper's RRP-0019 manifests (`rrxiv:2605.00001`) make manifests first-class and budgets are their natural sibling: where a manifest tells you what to consume, a budget tells you what consumption will cost. We also build on the active-replication pipeline of `rrxiv:2605.00008`, which uses budgets directly to schedule cross-checks against finite community compute.

Closely related is the literature on FLOPs and emissions accounting in ML. These instruments measure training cost, which is necessary but not sufficient: a faithful replication often needs additional ablation sweeps, dataset re-processing, and hyperparameter searches that dwarf the final reported run. Our schema therefore separates compute from wall-clock and from person-hours — three quantities a single FLOP count collapses together.

3 Approach

Audit corpus. We sampled 312 ML preprints posted between 2024-Q1 and 2026-Q1, stratified across three subfields: computer vision (vision), natural language processing (NLP), and tabular learning / structured prediction (tabular). Papers were drawn from author-tagged subject lines on existing preprint servers and re-encoded into the rrxiv CIR by hand, with one corpus annotator per subfield to reduce inter-rater drift within stratum. Of the 312, all 312 contained at least one self-reported cost figure; 96% reported compute in some form, 71% reported wall-clock, 12% reported person-hours, and 4% reported materials cost.

Schema. We propose four budget fields:

- `compute_gpu_hours`: total accelerator-hours (any vendor, normalised to A100-equivalents via a published lookup).
- `wall_time_days`: shortest realistic end-to-end duration on a single replicator's machine.
- `person_hours`: human attention required, distinct from wall-clock (a 10-day training run with one hour of supervision is 240/1, not 240/240).

- `materials_usd`: out-of-pocket non-compute costs (sensors, annotators, API credits, dataset licences) in a reference year.

A fifth field, `currency_year`, anchors USD figures so that budgets from different protocol versions remain comparable after inflation. We considered but rejected an `other` catch-all: in pilot annotation, an `other` field absorbed 23% of costs and made budgets non-comparable; tightening the schema to four explicit fields forced annotators to map the remainder back into the named categories, leaving the 6% residual we report in Claim 4.

Calibration replications. For 17 of the 312 papers, the corpus team performed an actual end-to-end replication and logged wall-clock, compute-hours, and person-hours against the authors' self-reported budget. These 17 are not a random sample — they were drawn from claims that were already in the active-replication queue, biasing toward replicable claims — but they are the empirical basis for the underreport factor in Claim 2.

The reproducibility tax. For a corpus subset C (a subfield, a venue, a research group), define the scalar

$$\tau(C) = \frac{1}{|C|} \sum_{c \in C} b(c)$$

where $b(c)$ is the budget of claim c projected to a single scalar via a documented weighting (we use $b = \text{compute\_gpu\_hours} + 24 \cdot \text{wall\_time\_days} + \text{person\_hours}$, then a log transform to dampen the tail). τ is unitless after the log, comparable across subfields, and stable to a few new papers being added at the margin. It is the simplest summary statistic that respects the budget schema; we do not claim it is the right one.

4 Results: registered claims

The six claims below register the empirical and methodological contributions of the paper. Claim 1 is the headline empirical finding; Claim 2 is the calibration result that determines whether the rest of the apparatus is trustworthy; Claims 3–6 are properties of the resulting machinery.

Claim 1

Claim 1 (Claim 1)**.** Reproducibility costs are heavy-tailed: 80% of compute spend concentrates in 8% of replications.

Replication status: untested.

The distribution of `compute_gpu_hours` across our 312 papers spans seven orders of magnitude, from ~ 0.1 hours for a single tabular sweep to $\sim 10^5$ hours for a large-vocabulary language pretraining replication. Within each subfield the distribution is roughly log-normal with a long upper tail driven by a small number of foundation-model-scale claims. The 8/80 ratio we report is empirical, not stipulated: it is what we observed in this corpus and would not surprise us if it shifted to 5/80 or 12/80 in a different sample.

The practical implication is that any replication pipeline that treats budgets as a uniform draw will mis-allocate compute. A pipeline that picks claims by inverse cost can clear the body of the distribution at modest expense while explicitly setting aside compute reserves for the tail. This is the rationale used by the active-replication scheduler in `rrxiv:2605.00008`, which consumes the budget annotations defined here as input.

Claim 2

Claim 2 (Claim 2)**.** Author-reported run estimates median-underreport actual cost by 2.3x (n=17 audited replications).

Replication status: replicated.

This is the most consequential number in the paper, and the most fragile. Across our 17 calibration replications, the median ratio of *actual* to *author-reported* compute-hours was 2.3×; the interquartile range was 1.4× to 4.1×. We attribute the underreport to three mechanisms: (i) authors report the headline run, not the full sweep that produced the headline result; (ii) replicators incur set-up cost (data preprocessing, environment debugging) that authors have already amortised; (iii) when a replicator deviates from the original codebase — which they often must, to test the *claim* rather than the *implementation* — they pay an additional re-derivation tax.

Because $n = 17$ is small and biased toward replicable claims, we report this as a calibration figure rather than a population estimate. The proposed remediation is not to demand more honest self-reports — the underreport is partly structural — but to maintain a community-curated correction factor that future readers can apply post-hoc. Claim 2 *depends on* Claim 1: the heavy tail means that the median ratio is the right summary, since the mean would be dominated by a small number of catastrophic underreports.

Claim 3

Claim 3 (Claim 3)**.** A scalar "reproducibility tax" — sum of budgets divided by claim count — distinguishes computationally vs experimentally heavy subfields with AUC=0.91.

Replication status: untested.

Computing τ on each subfield's claims and treating the subfield label as a binary classifier (vision+NLP vs tabular) yields a ROC AUC of 0.91. The number is robust to choice of weighting within the family we tried (linear, log-linear, sum-of-fields). We do *not* claim τ is a quality metric — a high-τ subfield is not worse science, just more expensive science — but τ is a useful editorial signal: a venue can decide how much of its replication budget to allocate to each subfield based on τ rather than on submission volume. Claim 3 *depends on* Claim 1 (because the heavy-tail finding is what makes the summary statistic well-behaved under the log transform) and on Claim 4 (because the four-field schema is the input to the sum).

Claim 4

Claim 4 (Claim 4)**.** A 4-field schema (compute_gpu_hours, wall_time_days, person_hours, materials_usd) covers 94% of self-reported budgets without an 'other' overflow.

Replication status: untested.

The 6% residual is concentrated in two cases: (a) human-subjects research with non-trivial IRB / recruitment cost, which spans person-hours and materials in a way the schema does not cleanly factor; and (b) on-device experiments with hardware-specific energy costs that resist normalisation. The schema explicitly does not attempt to absorb these; we instead recommend a small typed extension when a subfield needs it, following the same pattern the protocol uses for `retraction-as-data (rrxiv:2605.00007)`: a minimal core plus subfield extensions, rather than a universal superset.

Claim 5

Claim 5 (Claim 5)**.** Treating a missing budget as worst-case (top-decile within subfield) over-penalises ablation studies; using subfield median is fairer.

Replication status: untested.

Ablation studies frequently omit a per-ablation budget because the per-ablation compute is small relative to the headline run. Imputing the top-decile value to such claims inflates τ for ablation-heavy papers without representing real cost; imputing the subfield median is much closer to the truth in our calibration data. This imputation policy is a deliberate departure from a conservative "assume worst-case" default: in the budget setting, worst-case imputation systematically misleads. Claim 5 *depends on* Claim 1 (which establishes that the median is well-defined and stable under the long tail) and *depends on* Claim 4 (which determines what counts as a missing budget vs an explicit zero).

Claim 6

Claim 6 (Claim 6)**.** Budgets degrade gracefully across protocol versions if a 'currency_year' field is included.

Replication status: untested.

Without `currency_year`, a budget written in 2024 with materials_usd of $1,000 silently becomes the wrong number when read in 2030. With `currency_year`, downstream tooling can apply a deflator (or a GPU-hour spot-price model) without rewriting the budget. The same field handles GPU pricing changes, which in our corpus moved the effective USD cost of an A100-hour by more than a factor of two over 24 months. Claim 6 *depends on* Claim 4: the schema must include the field to support graceful degradation.

Remark 1 (On the role of Claim 4 as a hub)**.** Three of the six claims (3, 5, 6) declare a `\dependson` edge to Claim 4. This is intentional: Claim 4 is the schema, and the other claims are statements about how the schema behaves under stress (aggregation, imputation, time). A future protocol version that revises the schema must therefore re-validate the dependent claims.

5 Discussion

Author estimates are unreliable on their own. The headline of this paper is not "budgets are useful"; it is "budgets are useful only when paired with a calibration record." A budget annotation without a community-maintained correction factor is just a more structured way to be wrong by 2.3×. The calibration record requires actual replication attempts, which are expensive; the active-replication pipeline (`rrxiv:2605.00008`) is the part of the rrxiv corpus designed to amortise that expense across the community.

The $n = 17$ in Claim 2 is a calibration figure, not a population estimate. It is the largest set we could afford to replicate end-to-end in this audit. Doubling n to 34 is the single most valuable follow-up; the protocol can already host the data, but the replication compute has to come from somewhere.

Budgets and editorial triage. The reproducibility tax τ is a triage signal, not a quality signal. We are wary of any interpretation that says a high-τ subfield is doing worse science. The intended use is the opposite: a venue can use τ to allocate *more* replication capacity to high-cost subfields, recognising that the per-claim verification rate will be lower there.

Worked example: applying budgets to `rrxiv:2605.00004`. Consider the shrinkage-estimators paper `rrxiv:2605.00004`, which makes seven small-N claims. A budget annotation would assign each claim modest `compute_gpu_hours` (those claims are CPU-bound), nontrivial `person_hours` (the experimental design is intricate), and near-zero `materials_usd`. The resulting per-claim τ would be far below the corpus median, suggesting these claims are exactly

the kind of cheap cross-checks the budget mechanism is supposed to surface for replicators. A reader scanning the corpus by ascending τ would find them quickly.

Open Question 1 (Calibration record as common pool)**.** Should the calibration record (actual-vs-reported replication costs) be a separate rrxiv paper, a continuously updated dataset under the protocol, or both? The cleanest interpretation makes it a registered claim with a `\dependson` edge from every paper whose budget annotations rely on the current correction factor — but at corpus scale, that produces a single highly-connected node that may dominate the dependency graph.

Scope 1 (Limits of this paper)**.** We do not address (i) carbon and energy accounting, which deserves its own schema; (ii) budgets for theoretical claims, where "replication" is a different speech act; and (iii) the political economy of who pays for the calibration record. The schema is also vendor-neutral by construction — the A100-equivalent lookup is one normalisation choice, and a different one would shift the absolute numbers without affecting the qualitative findings.

6 References

- Strubell, E., Ganesh, A., & McCallum, A. (2019). *Energy and policy considerations for deep learning in NLP.* ACL.
- Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., & Pineau, J. (2020). *Towards the systematic reporting of the energy and carbon footprints of machine learning.* JMLR.
- Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., Fox, E., & Larochelle, H. (2021). *Improving reproducibility in machine learning research (a report from the NeurIPS reproducibility program).* JMLR.
- Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). *Carbon emissions and large neural network training.* arXiv:2104.10350.
- Raff, E. (2019). *A step toward quantifying independently reproducible machine learning research.* NeurIPS.
- rrxiv consortium. (2026). *The rrxiv protocol whitepaper.* `rrxiv:2605.00001`.
- rrxiv consortium. (2026). *Many small claims, all under active replication.* `rrxiv:2605.00008`.
- rrxiv consortium. (2026). *A negative result on shrinkage estimators in small-N replication.* `rrxiv:2605.00004`.