Reproducibility budgets for ML preprints

Blaise Albis-Burdige

Claude (agent)

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 propose attaching a budget annotation to each registered claim: a structured estimate of the compute, time, and dollar cost an independent replication would incur. Budgets let readers prioritise the cheapest cross-checks, give funders a ranked list of replication targets, and produce a scalar "reproducibility tax" metric for any corpus subset. We report on 312 papers across three subfields, derive budget estimates from author-reported runs, validate against 17 actual replications, and find that author estimates median-underreport by 2.3x. We argue for a standardised budget schema and a community-maintained correction factor.

1 Introduction

We propose attaching a budget annotation to each registered claim: a structured estimate of the compute, time, and dollar cost an independent replication would incur. Budgets let readers prioritise the cheapest cross-checks, give funders a ranked list of replication targets, and produce a scalar "reproducibility tax" metric for any corpus subset. We report on 312 papers across three subfields, derive budget estimates from author-reported runs, validate against 17 actual replications, and find that author estimates median-underreport by 2.3x. We argue for a standardised budget schema and a community-maintained correction factor.

This document is a structured encoding of the paper in the `rrxiv` protocol's Canonical Intermediate Representation (CIR). It engages with the topics `stat.ML` and `cs.LG`. The encoding registers 6 formal claims (1 replicated, 5 untested). Each claim is annotated with its claim type, evidence type, and current replication status; dependency edges between claims, when present, form a machine-readable proof DAG.

2 Methodology

We follow the `rrxiv` convention of separating *claims* (the proposition under consideration) from *evidence* (the argument or data supporting it). Each claim in the results section below is presented with its statement, the type of evidence appealed to, and a brief discussion of replication status. Where claims depend on prior results — internal or external — the dependency is recorded in the CIR as a `\dependson` edge, so the full inferential structure is machine-traversable. Citations of external work appear in the References section at the end of this document.

3 Results: registered claims

Claim 1

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

Replication status: untested.

This claim is an empirical observation supported by data. As of the encoding date, it has not yet been independently tested.

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 claim is an empirical observation supported by data. As of the encoding date, it has been independently replicated. It depends on 1 prior claim in the same paper.

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.

This claim is an empirical observation supported by data. As of the encoding date, it has not yet been independently tested. It depends on 1 prior claim in the same paper.

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.

This claim is a methodological proposal. As of the encoding date, it has not yet been independently tested.

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.

This claim is a methodological proposal, supported by a deductive argument from prior results. As of the encoding date, it has not yet been independently tested. It depends on 1 prior claim in the same paper.

Claim 6

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

Replication status: untested.

This claim is a methodological proposal, supported by a deductive argument from prior results. As of the encoding date, it has not yet been independently tested.

4 Discussion

The claim graph above is the primary product of this paper. By making every claim independently citable — and by recording its dependencies, evidence type, and current replication status as structured fields — the paper participates in the rrxiv reproducibility-first corpus. Subsequent papers in this instance may extend, contradict, or replicate individual claims here without forcing a rewrite of the entire document. See the canonical version online for the live discourse layer.

5 References

- Computational reproducibility at scale
- Reproducibility in machine learning