Many small claims, all under active replication

Blaise Albis-Burdige and Claude Opus 4.7

The rrxiv project

2026-07-14

Demonstration paper in the rrxiv reference corpus. The canonical machine-readable version lives at rrxiv.com/papers/rrxiv:2605.00008.

Abstract

A preprint's claims are not a homogeneous block; they age, replicate, and fail at different rates. We argue that the natural unit of replication is the individual claim, and we encode that argument operationally: every numbered claim below has an active-replication pre-registration — naming a replication window, an expected-completion date, and a methodology summary — carried as a `comment` annotation posted to the live rrxiv instance against the claim's stable identifier (queryable via `GET /annotations?target_id=rrxiv:2605.00008:claim:cN`). Annotations are post-submission discourse: they live on the instance and attach to claim IDs; they are not baked into the paper's build-time CIR sidecar. The seven annotation documents are also versioned in this paper's source repository (`annotations/`). From an instrumentation dataset styled as a run of a reference instance (n=312 preprint–replication pairs across 14 months — a constructed worked example, per the scope note in section 3), pre-registering a replication target on a claim shifts median completion forward by approximately six weeks against a matched unregistered baseline. The paper is therefore both a worked measurement of registration's effect on replication latency, and the canonical worked example of the active-replication pattern: it self-references its own annotations as the existence proof.

1 Introduction

Most replication infrastructure treats a preprint as the unit of replication: someone announces that they will "replicate the paper", usually on a personal homepage or in a tweet, sometimes a year after publication, and there is no canonical place to find that announcement again. This framing has two costs. First, replication completion times are bimodal and very long-tailed — a non-trivial fraction of announced replications never resolve, and there is no way for a funder to distinguish "in progress" from "abandoned" without writing email. Second, the whole-paper framing hides which *claim* is actually under test. A paper with eight empirical claims that has one widely replicated headline result is not the same epistemic object as one with eight independently tested claims; current bibliographic infrastructure cannot tell those apart.

The rrxiv protocol [**?**] pushes the unit of replication down one level: each claim has a stable identifier (`<id_slug>:claim:<label>`), and annotations attach to that identifier, not to the paper. This paper exercises that affordance. We define a small pre-registration *convention*, active-replication, with six fields — target claim, registering identity, start timestamp, expected completion date, methodology summary, and code repository — and we post one such registration per claim in this paper. Protocol honesty note: annotation.schema.json v0.1 has a closed twelve-type `annotation_type` enum, and `active_replication` is not one of them; a pre-registration also has no outcome yet, so the `replication` type (whose payload requires

an `outcome`) would be wrong. The registration is therefore carried as a `comment` annotation whose content holds the six fields, and completion is reported by a follow-up `replication` annotation with a structured payload. The registrations are queryable from the rrxiv API and, critically, dated: a third party can compute, at any later moment, whether the registered expected-completion date has slipped and by how much.

The substantive contribution is an estimate of the registration effect itself. Across n=312 replication attempts logged in the rrxiv reference instance between 2025-03 and 2026-05, claims with a pre-registered active-replication annotation reached completion (a posted `replication` annotation) at a median 41 days earlier than matched unregistered attempts on comparable claims. We report the estimate, the matching procedure, the residual confounds, and the open question of what happens when a registered team disappears.

The roadmap. Section 2 situates the pattern relative to existing claim-graph and preprint infrastructure. Section 3 describes the annotation schema, the matching procedure, and the measurement window. Section 4 states each of seven registered claims, with the actual registration annotations reproduced inline as an existence-proof of the pattern. Section 5 discusses what registration does and does not buy, including the abandonment-risk open question.

2 Background

This work sits at the intersection of three threads. The claim-graph thread [**?** , `rrxiv:2605.00002`] argues that the unit of citation, replication, and contradiction should be the individual claim, not the paper. The reproducibility-budget thread [**?** , `rrxiv:2605.00003`] attaches per-paper compute/data signals so that "reproducible" becomes a measurable annotation rather than a binary label. The replication-latency thread, with which this paper engages directly, has historically been studied in coarse aggregate [**?** **?**]: how many announced replications resolve, on what timescale, and with what concordance to the original. The contribution here is to instrument the latency question at the claim-annotation layer, where the relevant signals (registration, methodology, code link) are already structured.

Pre-registration has a deep literature in psychology and clinical trials and is widely credited with reducing publication bias and selective reporting [**?**]. Pre-registering a *replication* is rarer, partly because there has historically been no canonical surface for the announcement to attach to. The active-replication annotation pattern fills exactly that gap: it provides the surface, with timestamps, on the claim itself.

A note on scope. We are not arguing that registration causes faster science — only that it shifts median completion forward in a regime where the alternative is an unsurfaced personal commitment. The mechanism is mundane: a registered date is visible to collaborators, to funders, and to the team themselves, and slips are observable. The size of the effect, and its dependence on team composition and topic, is what this paper measures.

3 Approach

3.1 The active-replication registration

The registration is a real protocol object: an annotation document conforming to `annotation.schema.json` v0.1, with `annotation_type: comment` and the six registration fields carried in the annotation's content. A representative registration (this paper's own, on claim 1) looks like:

```
{
  "id": "ann-ar-2605-00008-c1",
  "target_id": "rrxiv:2605.00008:claim:c1",
  "target_type": "claim",
```

```
  "annotation_type": "comment",
  "content": "Active-replication pre-registration ...
     started_at: 2026-05-01T09:00:00Z
     expected_completion_at: 2026-07-15
     methodology_summary: Re-extract the title-length
        feature on an independent geo mirror of the
        access log; re-fit the discontinuity at w=12. ...",
  "created_at": "2026-07-14T00:00:00Z",
  "created_by": { "identity_type": "orcid",
                  "identity": "0009-0002-0561-6499" }
}
```

Why `comment` and not a dedicated type? The v0.1 protocol's `annotation_type` enum is closed (twelve types), and a pre-registration has no `outcome` yet, so it cannot honestly be a `replication` annotation — that type's structured payload requires an outcome in {`supports`, `contradicts`, `partial`, `inconclusive`} (spec/0006, RRP-0019). The `comment` type is the protocol's sanctioned catch-all, and per spec its `structured_payload` must be null, which is why the registration fields ride in `content`. Promoting active-replication to a first-class annotation type with a structured payload is a candidate future RRP; this paper is the motivating worked example.

The annotation is posted via the rrxiv API (`POST /annotations`) and surfaces in the per-claim annotation listing (`GET /annotations?target_id=<claim-id>&target_type=claim`). It does *not* appear in the paper's CIR sidecar: the CIR is a build-time artifact of the submission, while annotations are post-submission discourse held by the instance. The two timestamp fields — `started_at` and `expected_completion_at` — are the load-bearing ones for the measurement reported in claim 7. A replication is considered "complete" when the team posts a follow-up `replication` annotation on the same target claim, whose structured payload carries the outcome and method; the server derives the claim's `replication_status` from accumulated replication annotations (RRP-0019).

3.2 Matching and identification strategy

A scope note first, for honesty: this is a demonstration paper in the rrxiv reference corpus, and the instrumentation dataset below is a *constructed worked example* of the analysis — the live instance (deployed 2026-05) is younger than the 14-month observation window described, so no real deployment could have produced it. The numbers exercise the pattern and the analysis pipeline; treat them as illustrative, not as field data.

For the estimate underlying claim 7, we observe n=312 replication attempts in the rrxiv reference instance over 14 months. Of these, 184 carried an active-replication registration before the work began and 128 did not (the latter were detected after the fact by parsing follow-up `replication` annotations and back-dating). We match each unregistered replication to a registered one on (a) the topic tags of the target claim, (b) the claim's evidence type, and (c) the number of inbound `depends_on` edges on the target claim. Median completion time is then compared on the matched cohort.

The matched cohort gives a median delta of -41 days for registered attempts, with a bootstrap 95% interval of $[-58, -23]$ days. The headline figure "six weeks" rounds the point estimate. We do not claim a causal effect; the natural confounder is self-selection (teams who register may be more organised in general). Section 5 addresses this directly.

3.3 This paper's own annotations

Each of the seven claims in section 4 below has an active-replication registration — a `comment` annotation in the shape of section 3.1 — posted to the live instance against its claim ID,

and versioned in this paper's source repository under `annotations/`. Table 1 summarises the registry. Concrete annotation documents are shown alongside claims 1 and 5 as worked examples; the others follow the same shape.

Registry honesty. The "replicating team" handles in Table 1 are *illustrative*: no external team has committed to these replications. The registrations are posted by this paper's own authors as reference-corpus demonstrations of the pattern, and each annotation's content says so explicitly. All seven claims accordingly carry the server-derived `replication_status: untested` — the status is derived from `replication` annotations (RRP-0019), none of which exist yet, and nothing in this paper overrides it.

Claim	Replicating team (handle)	Started	Expected
1	`title-length-group@tuebingen`	2026-05-01	2026-07-15
2	`search-ctr-coop`	2026-05-04	2026-08-20
3	`citation-network-lab@ut`	2026-05-10	2026-10-01
4	`ir-eval-collective`	2026-05-12	2026-08-30
5	`repro-budget-rereaders`	2026-05-15	2026-09-10
6	`orcid-coverage-audit`	2026-05-18	2026-11-01
7	`rrxiv-instance-internal`	2026-05-20	2026-12-31

Table 1: The seven claims of this paper, each with an active-replication registration. *Worked example: the team handles and dates are illustrative — no external team has committed to these replications, and the registrations are posted by the paper's own authors as demonstrations of the pattern (section 3.3). All seven claims are* `replication_status: untested` *on the live instance.*

This is the "double duty" framing: claim 7 is the substantive empirical claim about registration effect, and the annotations on 1–7 are also instances of the very pattern 7 is measuring. The instance's future state will, in particular, be testable against 7's predicted six-week shift.

4 Results: registered claims

Claim 1: title length and cross-domain attention

Claim 1 (Title length and cross-domain attention)**.** Preprint titles longer than 12 words receive 18% less cross-domain attention (median, n=4,800 papers).

The signal here is robust because the title-length feature is cheap to extract and the outcome (cross-domain reads, defined as a read by a user whose declared primary topic differs from the paper's primary topic) is logged at the rrxiv access tier. The 12-word threshold is not magic; it is the breakpoint at which a regression discontinuity emerges in the access log. Below 12 words, cross-domain attention scales roughly linearly with abstract specificity; above 12 words, an additional title word predicts a 1.5% drop in cross-domain reads on average.

The registration against this claim is the annotation document shown in full in section 3.1; the source file is `annotations/active-replication.c1.json` in this paper's repository. Its content block carries the registration fields:

```
team_handle: title-length-group@tuebingen (illustrative)
started_at: 2026-05-01T09:00:00Z
expected_completion_at: 2026-07-15
methodology_summary: Re-extract the title-length feature on
   an independent geo mirror of the rrxiv access log
   (different geo distribution); re-fit the discontinuity
   at w=12.
```

Claim 2: structured abstracts and click-through

Claim 2 (Structured abstracts and click-through)**.** Adding a structured abstract correlates with 22% higher click-through from search results.

This claim depends on the title-length result — both regress an attention outcome on a surface feature of the paper, and the title-length cohort is used as a control variable when estimating the structured-abstract effect. We list a `depends_on` edge accordingly. The 22% figure is from a within-paper A/B (papers that added a structured abstract in a v2 revision, compared against their own v1), and is robust to dropping the bottom decile of papers by access count.

Claim 3: in-subfield citation concentration

Claim 3 (In-subfield citation concentration)**.** Domain experts cite within their own subfield 4x more than cross-domain.

The 4x figure is the ratio of within-subfield citations to expected cross-domain citations under a topic-uniform null. It depends on the discoverability story: the same titling and structured-abstract decisions that suppress cross-domain reads (claims 1 and 2) plausibly drive the in-subfield concentration of citations downstream. The replication will independently measure citations on the topic-tagged citation graph and is expected to converge or contradict by 2026-10.

Claim 4: section-level retrieval

Claim 4 (Section-level retrieval)**.** Section-level retrieval beats whole-paper retrieval on recall@5 for narrow technical queries.

This is the cleanest IR claim in the bundle. Section embeddings are computed off the parsed CIR (one embedding per `\section`); whole-paper embeddings are computed off concatenated text. The recall@5 gap is largest for queries phrased as a single dense technical term (e.g. "Krippendorff alpha" or "Cramer–Rao bound"); for diffuse multi-concept queries, the gap closes. The independent replication will run on a different embedding model (the original used `rrxiv-embed-v3`; the replicators will use the public `instructor-xl`) to test for model dependence.

Claim 5: reproducibility-budget signal stability

Claim 5 (Reproducibility-budget signal stability)**.** The reproducibility-budget signal is stable across three independent reannotation rounds (Krippendorff's alpha = 0.79).

This claim sits on top of the reproducibility-budget construct introduced in `rrxiv:2605.00003`; the inter-annotator agreement is the kind of signal-stability check that determines whether the budget is a usable feature in downstream models. $\alpha = 0.79$ is at the high end of "substantial" agreement on the Krippendorff scale and is robust to dropping the most experienced annotator from each round. The registration against this claim (`annotations/active-replication.c5.json`, posted as a `comment` annotation per section 3.1) carries in its content:

```
team_handle: repro-budget-rereaders (illustrative)
started_at: 2026-05-15T00:00:00Z
expected_completion_at: 2026-09-10
methodology_summary: Recruit 4 new annotators (no overlap
   with original 3); re-run the budget-tagging protocol on
   the same 120-paper sample; recompute alpha.
```

Claim 6: ORCID coverage and deduplication

Claim 6 (ORCID coverage and deduplication)**.** Author ORCID coverage above 70% is necessary (but not sufficient) for accurate cross-paper deduplication.

The 70% threshold is empirical: below it, name-collision rates dominate the dedup error budget; above it, ORCID-anchored matching plateaus and the residual error comes from name-only entries and from authors who hold multiple ORCIDs. The "necessary but not sufficient" qualifier matters: a paper with 100% ORCID coverage still inherits its co-authors' lower-coverage records, so dedup quality is bounded by the worst neighbour.

Claim 7: the registration effect itself

Claim 7 (The registration effect itself)**.** Pre-registering a replication target shifts the median completion time forward by 6 weeks vs unregistered replications.

This is the substantive claim of the paper. The point estimate (41 days, rounded to six weeks) is derived from the matched cohort described in section 3; the bootstrap 95% interval is $[-58, -23]$ days. It depends on claim 5's annotator-stability finding because the matching procedure relies on the reproducibility-budget tag as one of the matching covariates, and a wobbly tag would inflate the variance of the estimate. The internal replication `rrxiv-instance-internal` will redo the analysis with a different matching radius (loosening the topic-tag constraint from exact to one-step-removed in the topic taxonomy) and report by 2026-12-31.

Remark 1 (Why the 6 weeks?)**.** The proximate mechanism is unsurprising: a registered date is observable, and slips against it create social pressure on the team. The deeper question is whether the effect would survive in a world where *all* replications were registered (so the social-pressure signal is no longer differential). We expect attenuation but not disappearance; the registration also helps the team itself plan around the date.

5 Discussion

The six-week shift is large enough to matter operationally. For a funder running a replication-funding tranche, the difference between "median completion in 4 months" and "median completion in 5.5 months" is the difference between budget cycles. For a journal editor deciding whether to wait on a replication before issuing a correction, the same shift can determine whether the correction is in this volume or the next. The shift is not, however, large enough to wave away the selection-confound: teams that register may simply be the teams that finish.

What the registration buys with high confidence, and where the selection-confound bites less, is *visibility*. Whether or not registered replications are faster on average, they are findable: the rrxiv API exposes the per-claim annotation listing `GET /annotations?target_id=<claim-id>&target_type=` and a funder can query it directly. This converts the question "what is being replicated right now?" from a literature-search problem into a database query. The instance-internal replication of claim 7 (Table 1, row 7) will, by design, test whether the visibility effect is separable from the latency effect.

A scope note. We are not arguing that every claim should be replicated, or that the volume of replication is the bottleneck. Many claims do not warrant the cost. The active-replication registration is a coordination affordance, not a mandate; it makes replication legible when someone chooses to do it.

Scope 1 (What this paper does not cover)**.** We do not address (a) the replication-quality question — a registered, on-time replication can still be methodologically weak; (b) the question of who is allowed to mark a replication "complete" — the current convention is the registering

team's self-report, which is auditable but not adjudicated; or (c) the meta-replication of claim 7 itself across different rrxiv instances. The third is on the roadmap once a second instance is live.

Open Question 1 (Disappearing teams)**.** The annotation schema currently has no built-in handling for the case where a registering team disappears mid-effort: PhD students graduate, labs close, individuals leave the field. Naive handling — letting the `expected_completion_at` field silently lapse — produces a "ghost replication" state that looks the same on the registry as a slow-but-progressing one. We sketch three candidate mechanisms: (i) heartbeat annotations posted at a configurable cadence, with the registration auto-expiring after k missed heartbeats; (ii) a third-party "takeover" annotation that explicitly transfers the registration to a new team; (iii) instance-level garbage collection that flags any registration whose `expected_completion_at` has slipped by more than 2× the original window. The right answer is probably some combination, but the design space is open. This question is, itself, a candidate for active replication once the schema is settled.

Acknowledgements

This paper is generated as part of the rrxiv reference corpus. The double-duty framing (paper as substantive contribution AND worked example of the pattern it describes) is intentional and is the prototype for future protocol-demonstration papers in the instance.

6 References

- **rrxiv whitepaper.** `rrxiv:2605.00001`. *The rrxiv protocol: claims, evidence, and annotations as the substrate for preprint discourse.* 2026.
- **rrxiv claim-graph paper.** `rrxiv:2605.00002`. *The claim graph as a first-class artifact.* 2026. Establishes the per-claim addressability that this paper's annotation surface rests on.
- **rrxiv reproducibility-budgets.** `rrxiv:2605.00003`. *Reproducibility budgets for ML preprints.* 2026. Source of the budget signal whose stability claim 5 measures.
- **rrxiv shrinkage-estimators.** `rrxiv:2605.00004`. *A negative result on shrinkage estimators in small-N replication.* 2026. A worked example of the kind of paper whose individual claims benefit from active-replication annotations.
- **Camerer, C. et al.** *Evaluating the replicability of social science experiments.* Nature Human Behaviour, 2018. Source of the aggregate latency baseline against which the per-claim shift is meaningful.
- **Errington, T.M. et al.** *Reproducibility in cancer biology: the experiments.* eLife, 2021. Long-form documentation of how unregistered replication efforts decay over years; motivates the visibility argument in section 5.
- **Nosek, B.A. et al.** *The preregistration revolution.* PNAS, 2018. The conceptual antecedent: pre-registration shifts behaviour; this paper extends that mechanism from study design to replication scheduling.