AN APPROACH TO QUANTIFYING THE IMPACT OF ARTIFICIAL INTELLIGENCE ON THE SCIENCES: PERSPECTIVES FROM BIOMEDICINE

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine learning (ML) is now ubiquitous in the biomedical sciences, but prior evidence on how well it is implemented comes almost entirely from manual systematic reviews that do not scale beyond a few hundred papers. We present a literature-scale audit: an LLM-based pipeline that screened 27,200 open-access biomedical papers from OpenAlex — 11,736 classified as using ML — against failure modes drawn from the leakage taxonomy of Kapoor & Narayanan (2023), together with a frequency-matched non-ML comparison arm, a cross-model-family validation of both classification labels, a blind human review of a stratified validation set, and a sandboxed execution audit that followed every code-availability claim in the cohort to its source. Against the human standard the classifier was 79.4% accurate (95% CI 67.7–89.2), or 85.4% (74.8–94.3) once review articles are recoded under the pre-specified rule. Among assessable ML papers, 40.3% report performance metrics inadequate to the claim they support and 13.4% show evidence of illegitimate features. Of 2,056 papers claiming available code, 640 (31%) still resolve, 18 release code matching the paper's own methods, and exactly one builds and executes in an isolated sandbox — and even that one leaves its headline result unverifiable, because the released code omits the analysis that produced it. Yet ML papers accrue 78% more citations than matched non-ML papers after adjustment (IRR 1.78, 95% CI 1.64–1.94), appear in venues with 3.5× the median impact factor, and suffer no citation penalty when their availability claims fail audit. A forensic deep read of 200 papers shows the automated flags mark inflated claims (median reported improvement 16.5% versus 4.3%, $p = .0003$) rather than textual evidence of leakage itself. We report that as a failure analysis of the instrument as much as a measurement of the literature: a screen built to detect one epistemic property reliably detects a different one, and only the execution-grounded tier separates them. We release the pipeline, prompts, and per-paper audit records.

1 INTRODUCTION

ML is now an instrument of everyday biomedical research, and the volume of literature employing it has grown sharply over the past decade (Priem et al., 2022). The enthusiasm has a well-documented dark side: ML is easy to implement incorrectly in ways that lend weight to spurious claims — and traditional methods often match or outperform it (Vabalas et al., 2019) — through data leakage, inappropriate metrics, and a family of pitfalls specific to biomedical data (Kapoor & Narayanan, 2023; Lones, 2024; Whalen et al., 2022). These are not theoretical concerns: nearly all published COVID-19 prediction models were found at high risk of bias (Wynants et al., 2020), and none reviewed for COVID-19 imaging were judged fit for deployment (Roberts et al., 2021). Reporting frameworks such as MI-CLAIM (Norgeot et al., 2020) and TRIPOD-AI/PROBAST-AI (Collins et al., 2021) standardize such appraisal, but applying them requires an expert to read each paper — which has confined every prior assessment to a few hundred hand-curated papers. The largest prior effort of this kind, by Kapoor & Narayanan (2023), aggregated 17 field-specific reviews to identify 294 affected papers, then hand-analysed the 12 papers in one field that shared complete code and data, finding errors in 4.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Large language models change the economics: they approach human-level accuracy on literature screening and extraction (Khraisha et al., 2024) at a per-paper cost measured in cents. That makes metascience a natural application domain for AI — if the resulting measurements can be validated and their constructs stated honestly. This paper attempts both.

Contributions. (1) A literature-scale, fully auditable measurement of ML implementation quality in biomedicine: 27,200 open-access papers screened directly against five failure modes from an established leakage taxonomy — two orders of magnitude beyond the largest prior leakage audit, which aggregated others' reviews rather than screening papers itself — with a frequency-matched non-ML comparison arm that prior audits lacked. (2) A release audit that follows every code-availability claim in the cohort through link resolution, code–paper correspondence, environment build, and sandboxed execution — measuring, for the first time at this scale, what survives of a reproducibility claim with no human rescue effort. (3) A failure analysis of the pipeline as a scientific-AI system: both classification labels re-derived by two independent model families on a 3,000-paper subsample, validated against a blind human review of a stratified 60-paper sample, and a 200-paper forensic deep read separating what the cheap automated flags detect (inflated claims) from what they are assumed to detect (leakage mechanisms) — a negative result about the instrument, established from its own released records. (4) The pipeline, versioned prompts, per-paper audit sidecars, and analysis of record, all public (see Data and code availability).

2 METHODS

Figure 6 (Appendix A) gives an overview. Every work drawn into the study carries a manifest entry; every exclusion carries a machine-readable reason code from a closed vocabulary; and every processing stage writes a per-work JSON sidecar to a public bucket, so each reported number is traceable to the records that produced it. Appendices A–E give full detail for each stage; this section states the design.

2.1 CORPUS FRAME AND ENRICHMENT SAMPLING

The frame is the subfield layer of the OpenAlex topic taxonomy (Priem et al., 2022): all Life Sciences and Health Sciences subfields plus Biomedical Engineering, minus 25 subfields out of scope for a biomedical audit, leaving 81. Within that boundary we retained articles published 2010 or later that expose an open-access PDF — approximately 8.7 million works, and the principal generalization limit of the study. The ML arm is drawn under an enrichment proxy (secondary AI/CV topic assignment, or ML vocabulary in title or abstract); the non-ML arm is drawn frequency-matched to the ML arm on subfield $\times$ year. The proxy is a sampling device only: every reported analysis uses the classifier label, never the arm.

2.2 CLASSIFICATION AND ITS VALIDATION

A confidence-gated cascade labels each paper ML / non-ML / unclear, and each ML paper ML-primary / ML-supported, from OCR full text. The cascade buys the necessary throughput but contains no internal independent rater, so its validity rests on two external controls that answer different questions. A *stability* control re-classifies a random 3,000-paper subsample end to end with two model families independent of the cascade's (Section 3.6). An *accuracy* control — a blind human review of 60 papers, labelled before seeing model output and stratified across agreement and disagreement cells — estimates accuracy against a human standard, which is the quantity the cascade's validity ultimately rests on; agreement between models bounds stability, not correctness. Both are reported in Section 3.6, with full detail in Appendix B.

2.3 FAILURE-MODE SCREENING

Five checks are operationalized from the leakage taxonomy of Kapoor & Narayanan (2023). Three are judgment checks, where a model returns the verdict directly with a confidence rating, a verbatim supporting quotation, and a rationale; two are extraction-plus-rule checks, where the model extracts values only and the verdict is computed deterministically in code. Every prompt states that absence of evidence is `UNCLEAR`, never `FAIL`. Appendix C gives each check and its adequacy rule.

2.4 RELEASE AUDIT: FROM CLAIM TO SANDBOXED EXECUTION

Every code-availability claim in the cohort is followed through four tiers: L0, the claim itself; L1, the link resolves with artefacts present; L2, the released code matches the paper's own description of its methods; L2.5, a buildable environment specification; L3, execution in an isolated sandbox. The standard is a no-rescue standard: reproducibility as published, with no human repair. The L2+ readers deep-audit GitHub-hosted claims in canonical `owner/repo` form; other hosts pass L1 unaudited, so the funnel below L1 is a floor, reported as such.

2.5 FORENSIC DEEP READ

A subsample of $n = 200$ ML papers — 50 per cell of a 2×2 design over any-flag $\times$ era (2010–2020 vs. 2021+) — received a deep structured read recording affirmative textual evidence of a specific leakage mechanism (full-dataset preprocessing, feature selection before the split, temporal or group leakage), whether seeds were reported, and a headline performance value paired with its baseline. The deep read recovered a value-and-baseline pair for 121 of 200 papers (61%), against 22% for the shallow cohort-wide extractor — a $2.8\times$ per-paper yield, so extraction depth rather than the literature limits how many papers a claim-inflation estimate can draw on. The forensic read and the cheap checks measure different constructs, and Section 3.5 turns on that difference.

2.6 STATISTICAL ANALYSIS

The primary comparison contrasts ML with non-ML papers; the sub-comparison contrasts ML-primary with ML-supported. Both use classifier labels, not sampling arm; the post-classification groups are not individually matched, and comparability rests on the shared subfield $\times$ year frame, the year cutoff, and regression adjustment for year and subfield, with adjusted and unadjusted estimates reported side by side. Citations are modeled by negative binomial regression (reported as incidence rate ratios), venue impact by Mann–Whitney U on OpenAlex's two-year mean citedness, and categorical outcomes by χ^2 or Fisher's exact test; failure-mode families are corrected by Benjamini–Hochberg FDR at $q = .05$. The furthest audit tier reached enters the adjusted citation model as an ordinal predictor. Author count and ML-specialist co-author were fitted as moderators. Every test emits a structured record, and all reported numbers derive from a single frozen analysis of record (Appendix E).

3 RESULTS

3.1 COHORT

The classifier assigned $11{,}736$ distinct papers to the ML group and $15{,}423$ to non-ML, leaving 41 (0.2%) `UNCLEAR`; within ML, $9{,}557$ rows are ML-primary and $2{,}479$ ML-supported. (The analysis of record computes over $27{,}679$ classified rows, $\sim 1.7\%$ of them duplicated manifest rows — since repaired; duplicated rows carry identical verdicts, so within-group rates are unaffected; Appendix G.) Published years span 2010–2026.

3.2 FAILURE-MODE PREVALENCE WITHIN THE ML CORPUS

Figure 1 reports each check as a prevalence among ML papers with a definite verdict (Table 2, Appendix C, gives exact counts). 40.3% of assessable ML papers report performance metrics inadequate to the claim they are offered in support of; 13.4% show evidence of features unavailable at genuine prediction time. The test-set-size rate is extreme (92.8%) but also the most selected — a paper enters that denominator only by reporting enough statistics to reconstruct an interval — so we report it without extrapolating. Failures are not concentrated in a subset of bad papers: 42% of ML papers with at least one definite verdict fail no check and 41% fail exactly one, and pairwise co-occurrence runs 0.9–$1.3\times$ independence, so the checks behave as near-independent detectors of distinct problems. No composite score is constructed (Appendix E).

Three of five checks show no significant subtype difference after correction, and the fourth, inappropriate metrics, is significant but negligible in size (40.9% vs. 37.9%, Cramér's $V = 0.03$). The

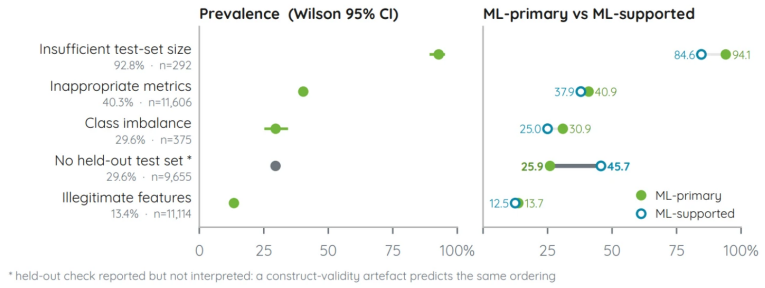

Figure 1: Failure-mode prevalence within the ML corpus. Left: per-check prevalence with Wilson 95% intervals; denominators are check-specific and differ by a factor of forty. Right: the same checks by ML subtype. Only the held-out-test-set check separates the subtypes materially, and it is the one check whose direction we do not interpret (see text).

centrality of the ML contribution therefore does not meaningfully predict these failure modes — an interpretable null. The exception, the held-out-test-set check (ML-primary 25.9% vs. ML-supported 45.7%), is the one check on which ML-primary papers appear better on any measure in this study, and its direction follows the same ordering a construct-validity artefact would produce; a rigor gradient and an applicability artefact predict the same result here, and this analysis cannot separate them, so we report it uninterpreted (Table 3, Appendix C).

3.3 Release audit: one claim in two thousand survives

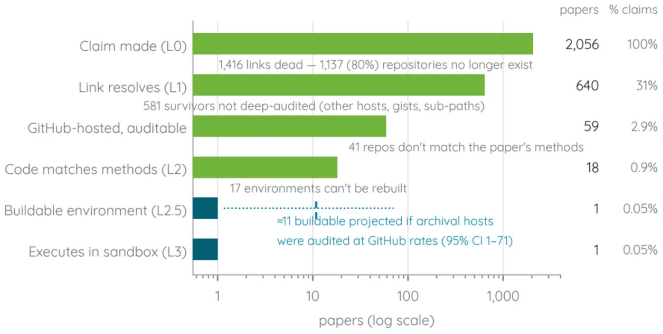

Figure 2: Survival of code-availability claims across audit tiers (distinct papers, tallied from the released audit sidecars). The projected value is an extrapolation of GitHub-observed rates over all surviving links, not a measurement; its interval is wide because the build rate rests on one success in eighteen.

2,056 papers made a followable code-availability claim (Figure 2; counts in Table 5). 640 (31.1%) still resolve with artefacts present. Of the 1,416 dead claims, 1,137 — 80.3% of dead claims, and 55.3% of every claim audited — point at a GitHub repository that no longer exists (or was never public). Of the 94 surviving GitHub-hosted links, the 59 in canonical `owner/repo` form were deep-audited; 18 of those released code matching the paper's own description of its methods; one produced a buildable environment specification; and that one executed successfully in the sandbox (the `additivityTests` R package, OpenAlex W2013748161).

Two readings of that figure are available and only one is supported. The measured quantity is survival of an availability claim to unassisted execution under a no-rescue standard — not how

often biomedical ML results replicate. On that stricter question the count is zero, not one: the paper that executed ran the released package's own documented example on synthetic data, and its correspondence record states that the repository implements the paper's statistical tests but not the simulation study that produced its headline metric, logged `UNVERIFIABLE` rather than replicated. Applying GitHub-observed rates to all 640 surviving links projects on the order of 11 buildable papers cohort-wide ($\approx 0.5\%$ of claims) — an order of magnitude above the measured floor, still vanishingly small.

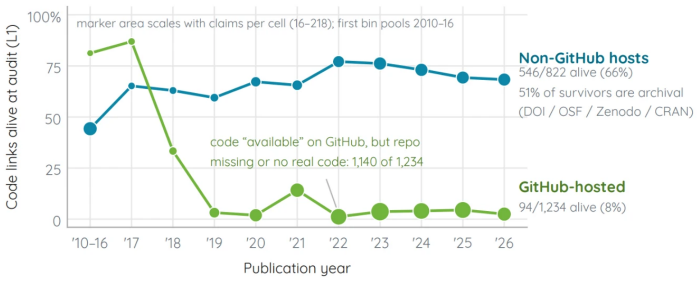

Figure 3: Share of claimed-code links alive at audit, by publication year: GitHub-hosted versus all other hosts. Marker area scales with claims per cell; sparse early years are pooled. GitHub claims — 60% of all claims, and concentrated after 2018 — fail near-totally from 2019 on; half of the surviving non-GitHub links are archival deposits.

Dead availability claims are a hosting problem. Splitting claims by host class localizes the attrition (Figure 3). GitHub-hosted claims are 60% of all claims audited (1,234 of 2,056) and fail at 92% (94 alive); claims on all other hosts survive at 66% (546 of 822). The failure mode is not gradual link rot: GitHub claims concentrate in 2019–2026 (median year 2023) and fail near-totally within a few years of publication — consistent with repositories that were deleted quickly or never made public — while the small population of pre-2018 GitHub claims mostly survives. Non-GitHub dead links, by contrast, do skew older than surviving ones ($p = 4 \times 10^{-5}$), the signature of genuinely aging lab and project pages. Archival infrastructure (DOI-resolved Zenodo, OSF, CRAN and kin) accounts for half of all surviving non-GitHub links, and no archival link was found dead — though dead non-GitHub links carry no recorded host, so archival mortality is bounded, not measured (Appendix D). Mandated archival deposits appear to outlive author-hosted repositories — an immediately actionable finding for journal availability policy (Section 4).

3.4 The attention economy does not price verifiability

ML papers are cited 78% more than non-ML papers after adjustment for year and subfield (IRR 1.78 $[1.64, 1.94]$, $p \approx 4 \times 10^{-41}$); the adjusted estimate exceeds the unadjusted 1.43, so the premium is not an artefact of ML work being more recent or better-placed (Figure 4). The venue gap is larger still: median journal impact factor 3.13 against 0.89 (rank-biserial $r = .35$). Retraction is rare in both groups but roughly twice as likely among ML papers (0.23% vs. 0.12%, OR 2.02, $p = .018$), and within ML is concentrated among ML-primary work (OR 7.02, $p = .026$ — 27 events against 1, a direction rather than a magnitude). The premium persists under both fittable moderators (author-count interaction IRR 2.10; ML-specialist co-author IRR 1.60, fitted on $n = 26{,}849$), so neither explains it away.

Among audited papers, the furthest audit tier reached does not predict fewer citations: 1.36 per tier step ($[0.995, 1.855]$, $p = .054$), a trend that if anything runs positive. There is no evidence in this cohort that failing an availability audit carries a citation cost — an absence of evidence, stated as such. One methodological result rides alongside: an interim estimate on a partial audit snapshot ($n = 679$) was significantly inverse (0.67, $p = .023$) and reversed on audit completion; mid-pipeline estimates from pipelines of this kind should be treated as provisional by default.

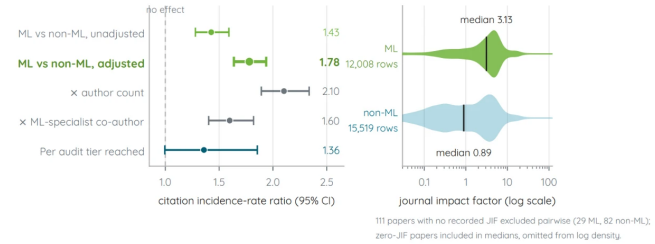

Figure 4: Left: citation incidence rate ratios with 95% CIs. The adjusted ML premium exceeds the unadjusted one, so the premium is not an artefact of ML work being more recent; the audit-tier coefficient is the only interval crossing unity. Right: journal impact factor by group (log scale), medians marked; that panel's ns are analysis-of-record rows with a recorded impact factor, so they exceed the distinct-paper counts in Section 3.

3.5 Flags mark inflated claims, not leakage evidence

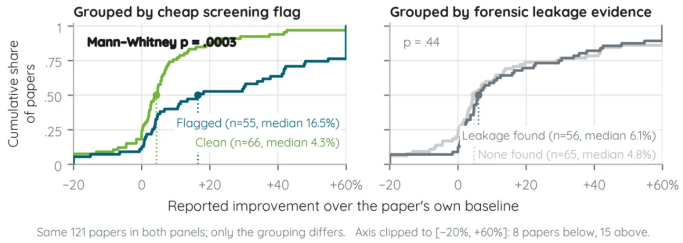

Figure 5: The same 121 forensic-subsample papers in both panels; only the grouping differs. Left: reported relative improvement grouped by cheap screening flag. Right: the same quantity grouped by affirmative forensic leakage evidence. The screen separates large claims from small ones; the leakage construct does not.

In the forensic subsample ($n = 200$, zero read errors), 98 papers (49%) showed affirmative textual evidence of a leakage mechanism, and 6 (3%) reported a random seed. The central result is a dissociation (Figure 5): papers the cheap checks flagged report a median improvement over their own baseline of 16.5%, against 4.3% for unflagged papers ($p = .0003$) — yet those same flagged papers are no more likely to show affirmative leakage evidence than unflagged ones (52% vs. 46%), and papers with leakage evidence do not report larger improvements (6.1% vs. 4.8%, $p = .44$). The corpus-wide meta-analytic estimate through the shallow extractor is uninformative ($n = 86$ papers with value and baseline; all $p \geq .20$).

The cheap checks are therefore a detector of inflated claims rather than of leakage, notwithstanding the taxonomy they are drawn from. An elevated flag rate licenses a claim about how large reported effects are — not about the mechanism that produced them. Establishing the latter requires comparing reported values against faithful reruns, which this study reaches for exactly one paper.

3.6 The measurement instrument is accurate and reproducible where it matters

On the 3,000-paper subsample, the cascade's ML/non-ML labels reproduce across independent model families at $\kappa = 0.81$–0.87 (agreement 93.5% with `llama-3.3-70b`, 90.8% with `gpt-4o-mini`, against 94.0% between the two independents); where the independents agree

with each other ($n = 2{,}794$), the cascade matches them 94.8% of the time, a 5.2% ceiling on cascade-specific stage-1 disagreement. The ML-primary/ML-supported split reproduces far less well ($\kappa \approx 0.5$), with one independent model placing the boundary substantially differently, and the sub-comparison should be read with that in mind (Table 1, Appendix B).

Stability is not accuracy, so the cascade was also scored against a human standard. A single rater labelled a blind sample of 60 papers drawn across three strata (all-models-agree, cascade-dissents, models-disagree), reweighted to the composition of the 2,973-paper agreement subsample. Under the rater's applied interpretation — classifying by subject matter — stratified cascade accuracy was 79.4% (95% CI 67.7–89.2). Recoding the eleven ML-topic review articles as `non_ml`, per the pre-specified rule that discussing machine learning does not constitute machine-learning use, gives 85.4% (95% CI 74.8–94.3); the difference between the two figures arises solely from those eleven papers. Per-stratum accuracy was 86.7% (26/30) where all models agreed, and 20% in each of the two oversampled disagreement strata (4/20 and 2/10). Cohen's κ on this sample is 0.114 and is not interpretable as an agreement statistic, since the disagreement strata are deliberately oversampled. The sample is small and single-rater; both readings carry wide, substantially overlapping intervals and should not be treated as distinguishable (Appendix G).

Tool-concentration results — declining Herfindahl indices as subfields mature, no support for a tool-monoculture story — are in Appendix F.

4 Discussion

Verifiability is not priced. Every incentive-bearing signal in this cohort runs the same direction: ML papers are cited more, placed in higher-impact venues, and retracted more; papers whose availability claims fail audit lose nothing. The reward system prices the claim of ML and, apparently, nothing about whether it can be checked — the quantitative shape of a market failure manual audits have long suggested but could not measure at scale. Interventions aimed at readers (checklists, badges) address the wrong side of that market unless they also touch citation and venue incentives.

Verification infrastructure, not author virtue. More than half of every code-availability claim we audited points at a GitHub repository that no longer exists — and the failure pattern indicates repositories deleted quickly or never made public, not links that aged out. Meanwhile archival deposits persist, and the one claim that executed pointed at a GitHub repository for a package also published on CRAN — infrastructure that enforces buildability as a condition of existence, which is plausibly why it built at all. The actionable version of this finding is boring and cheap: journals requiring a DOI-addressed archival deposit (Zenodo/OSF-style) at acceptance, rather than a URL, would have preserved most of what this cohort lost — and, because the loss occurs at or near publication time, a check at acceptance is exactly when it would bite.

What an automated rigor screen measures. Our forensic result cuts both ways for AI-for-metascience. The flags are real signal — they identify papers whose claimed improvements are four times larger — but they are not evidence of the mechanism they are named for. Any deployment of LLM screening as a rigor instrument should state its construct this plainly: cheap screens rank papers by how far their claims outrun their reporting; only execution-grounded audits convert that ranking into a measured inflation tax. Sandboxed, no-rescue execution keyed to each paper's own headline metric — what the L3 tier prototypes — is the infrastructure that would.

Relation to prior work. Kapoor & Narayanan's twelve hand-analysed papers (Kapoor & Narayanan, 2023) and our execution count should not be compared directly: their twelve were selected for having complete code and data, whereas ours measures survival of availability claims as published, with no rescue — the complementary lower bound.

5 Limitations

The cohort is the retrievable open-access subset (27,200 of 49,179 sampled), truncated by host rate-limiting rather than design; round-robin ordering preserves stratum balance, but results generalize only to fetchable OA biomedical literature. All labels come from proprietary LLMs; classification

is a single-family cascade whose validity rests on the cross-family stability study and a single-rater blind spot-check of 60 papers, and ambiguity routes to UNCLEAR, never FAIL, so a hedged or unconventionally described failure is more likely lost from the denominator than miscounted as a pass. The spot-check is small enough that its two readings (79.4% and 85.4%) have substantially overlapping intervals, and no human–human agreement figure is available, so the ceiling on achievable classifier accuracy is unknown. The five checks are one slice of the documented pitfall space, two ran on a $1{,}700$-paper subset, and all five presuppose a predictive design with no NOT_APPLICABLE verdict available — which is why no cross-arm failure contrast is reported and why the held-out check's subtype difference is left uninterpreted. The release audit reads GitHub-hosted claims only past resolution and applies a no-rescue standard, so it is a floor; its claim-inflation contrasts are claims-based, the true tax awaiting execution at scale. Venue impact uses OpenAlex two-year mean citedness, not field-normalized. Both comparisons are observational: frequency-matched and regression-adjusted, not individually matched or randomized. Full statements in Appendix G.

6 Conclusion

ML implementation quality in biomedicine has a measurable floor, and the reward system pays the ML premium without inspecting any of it: two in five assessable ML papers support their claims with inadequate metrics, more than two-thirds of code-availability claims lead nowhere, and not one paper in the cohort could be carried to a verified replication of its own headline number. The pipeline that measured this is cheap, auditable, released, and explicit about what its flags do and do not mean — the standard, we argue, for any AI-for-metascience instrument.

Data and code availability

The processed corpus — per-work OCR text, classification, failure-mode, audit and forensic sidecars, and every execution artefact — is already hosted in public, credential-free object storage, and the pipeline library ships with versioned prompts, the frozen analysis of record, and 128 unit tests. URLs are withheld here as author-identifying; they appear in the camera-ready and can be supplied to reviewers via the area chairs.

Use of large language models

LLMs are this study's measurement instrument, not an incidental aid: Gemini 2.5 Flash and Pro perform classification, the judgment checks, both extraction passes, and the forensic read; GPT-4o-mini and Llama-3.3-70B are the independent raters in the agreement study. Every prompt is versioned and released. LLM assistance was also used in drafting this manuscript; all reported numbers were recomputed from the released artefacts rather than taken from generated text, and the authors are responsible for the content.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300.

Bland, J. M. and Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476):307–310.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Collins, G. S., Dhiman, P., Andaur Navarro, C. L., Ma, J., Hooft, L., Reitsma, J. B., Logullo, P., Beam, A. L., Peng, L., Van Calster, B., van Smeden, M., Riley, R. D., and Moons, K. G. M. (2021). Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*, 11(7):e048008.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36.

Harris, C. R. et al. (2020). Array programming with NumPy. *Nature*, 585:357–362.

Kapoor, S. and Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804.

Kerby, D. S. (2014). The simple difference formula: an approach to teaching nonparametric correlation. *Comprehensive Psychology*, 3:11.IT.3.1.

Khraisha, Q., Put, S., Kappenberg, J., Warraitch, A., and Hadfield, K. (2024). Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Research Synthesis Methods*, 15(4). https://doi.org/10.1002/jrsm.1715

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Lones, M. A. (2024). How to avoid machine learning pitfalls: a guide for academic researchers. *arXiv preprint arXiv:2108.02497v5* (originally posted 2021; revised August 2024).

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60.

McKinney, W. (2010). Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 56–61.

Norgeot, B., Quer, G., Beaulieu-Jones, B. K., Torkamani, A., Dias, R., Gianfrancesco, M., Arnaout, R., Kohane, I. S., Saria, S., Topol, E., Obermeyer, Z., Yu, B., and Butte, A. J. (2020). Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nature Medicine*, 26(9):1320–1324.

Priem, J., Piwowar, H., and Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv:2205.01833*.

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., Weir-McCall, J. R., Teng, Z., Gkrania-Klotsas, E., AIX-COVNET, Rudd, J. H. F., Sala, E., and Schönlieb, C.-B. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, 3:199–217.

Seabold, S. and Perktold, J. (2010). statsmodels: Econometric and statistical modeling with Python. In *Proceedings of the 9th Python in Science Conference*, pages 92–96.

Vabalas, A., Gowen, E., Poliakoff, E., and Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLOS ONE*, 14(11):e0224365.

Virtanen, P. et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272.

Whalen, S., Schreiber, J., Noble, W. S., and Pollard, K. S. (2022). Navigating the pitfalls of applying machine learning in genomics. *Nature Reviews Genetics*, 23(3):169–181.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.

Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., Bonten, M. M. J., Dahly, D. L., Damen, J. A., Debray, T. P. A., et al. (2020). Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ*, 369:m1328.

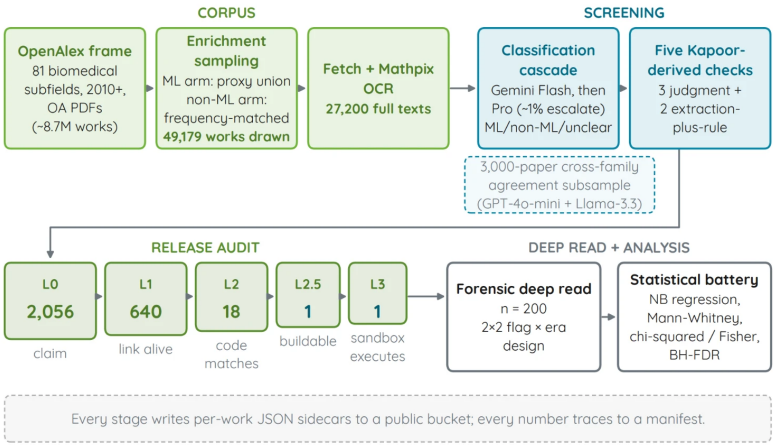

Figure 6: The qml_v1 pipeline. A stratified, ML-enriched draw from an 81-subfield OpenAlex frame is fetched, OCR'd, and classified by a confidence-gated cascade; classified papers receive five failure-mode checks; every code-availability claim is followed from the claim itself through four audit tiers ending in sandboxed execution; and a 200-paper forensic deep read provides the construct check on the automated flags.

A Corpus construction and provenance

A.1 Frame boundary

Selection keys on the subfield layer of the OpenAlex topic hierarchy rather than the topic layer: topics are machine-generated clusters and unstable across OpenAlex releases, whereas subfields are a curated taxonomy that can be stated and defended. The frame comprises all Life Sciences and Health Sciences subfields plus subfield 2204 (Biomedical Engineering), minus 25 subfields dropped as out of scope for a biomedical audit — chiefly agriculture, veterinary science, allied health, nursing administration, dentistry, and complementary medicine — leaving 81. The full drop list, with a recorded reason per subfield, is released with the corpus design tables; the cuts removed approximately 15% of the frame's mass but only approximately 10% of its ML population. Within the boundary, works of `type:article` published 2010 or later with an open-access PDF location (`best_oa_location.pdf_url`) are retained — approximately 29% of the boundary, on the order of 8.7 million works. The frame snapshot (query, retrieval date, cursor order) is recorded with the draw; retrieval and processing ran to 2026-08-28.

A.2 Enrichment signals

The ML-arm proxy has two signals, used for sampling only. Signal A (citation cluster): a secondary topic in subfield 1702 (Artificial Intelligence) or 1707 (Computer Vision), catching papers embedded in the ML literature that never use the phrase "machine learning". Signal B (vocabulary): ML terminology in the title or abstract, catching ML-supported papers living entirely inside biomedical citation neighbourhoods; OpenAlex caps a search at five boolean operators, so longer term lists were issued as multiple queries and de-duplicated. The signals overlap by 18–25% (of signal B and signal A respectively; 12% by Jaccard); their union is approximately 582,000 works ($\approx$ 2% of the full subfield boundary), of which on the order of 170–260,000 also expose a fetchable PDF. Subfield allocation is proportional to the subfield's ML-union count, capped at 2,500 works per subfield; the specified floor of 50 is not binding in the realised allocation, where 27 small subfields fall below it. At least 30% of the ML draw was required to fall in 2010–2020, so that analyses requiring a complete five-year forward citation window are not starved by post-2019 ML volume; the realised share was 30.1% in aggregate, though 41 of 81 subfields fall below 30% individually. A fixed seed plus the recorded frame snapshot make the draw exactly reproducible and allow incremental top-up without re-drawing.

A.3 INCLUSION CRITERIA AND MANIFEST

Records were admitted subject to three ordered criteria, the first failure determining a record's fate: (1) resolvable identifier (exclusions `metadata_incomplete`, `duplicate_paper_id`); (2) publication year present and $\geq$ 2010 (`missing_publication_year`, `publication_year_before_cutoff`); (3) full-text availability (`missing_pdf_url`, `pdf_fetch_failed`, `invalid_file`, `extraction_failed`). A fourth criterion used in our pilot — requiring a reconstructable abstract — was deliberately dropped: it served a title-and-abstract rater that v1 does not use (all classification runs on OCR full text) and cost approximately 42% of retrieved records for no analytic benefit. Every record receives a manifest entry recording provenance and fate; the first exclusion recorded is final; the manifest is reconciled after every stage by verifying pulled = included + excluded; and all reported counts derive from this manifest. Figure 7 shows retention.

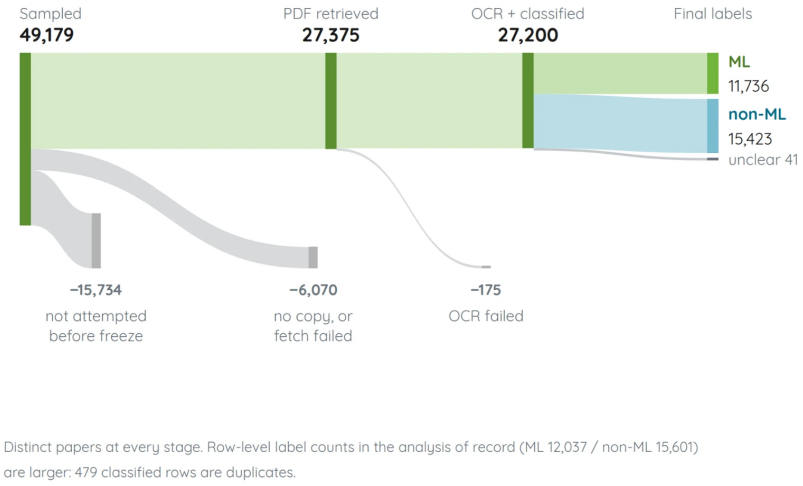

Figure 7: Retention from sampling frame to classified cohort, distinct papers at every stage.

A.4 FULL-TEXT ACQUISITION

Full text was recovered through a two-vendor router rather than a single engine. The Mathpix Files API is preferred — it accepts whole PDFs in batch and returns Markdown with LaTeX mathematics preserved, which matters for the methods sections the failure-mode checks read — but it is reserved against a monthly page cap; once that cap is consumed the router falls through to Datalab, which converts per file. Over the $27{,}200$ documents in the analysed cohort the realised split, read back from the released `.ocr.json` sidecars, is $25{,}707$ documents (94.5%; $325{,}632$ pages) through Datalab and $1{,}493$ (5.5%; $18{,}547$ pages) through Mathpix — so Datalab, the fallback by design, did the large majority of the work in practice. Transport and conversion failures are recorded per work with their engine and reason in `data/ocr_failures.csv` (511 records over 379 works, most of which succeeded on retry).

A.5 ACHIEVED COHORT AND TRUNCATION

Tier-1 ingestion retrieved $27{,}375$ PDFs; $27{,}200$ completed OCR and classification; 175 works lacked a usable OCR result. Ingestion was frozen when the OpenAlex content host began hard-throttling hosted-PDF downloads; the works remaining in the draw are predominantly hosted-only. By sampling arm the analysed rows are $16{,}574$ ML-arm and $11{,}105$ non-ML-arm (60/40 rather than 50/50, because ML-arm works had richer hosted-PDF coverage); by classifier label — what every analysis uses — the distinct-paper cohort is $11{,}736$ ML, $15{,}423$ non-ML, and 41 unclear. The analysis of record carries $27{,}679$ classified rows including 479 duplicated rows ($\sim 1.7\%$, subsequently repaired; the full tier-1 manifest carries 595 duplicates across all $49{,}774$ rows), duplication being roughly

twice as common among ML papers; duplicated rows carry identical verdicts, so within-group rates are unaffected while row-level analyses slightly overstate precision. Tier 2 (extraction-only breadth pass over the same 81-subfield frame with no ML enrichment) reached 32,087 papers at freeze; Tier-1 works are a subset of Tier-2 processing, so no work is extracted twice.

A.6 Venue and author enrichment

OpenAlex work payloads embed dehydrated source objects without `summary_stats`, so journal impact factor is unavailable as retrieved. Distinct source identifiers were collected across retained papers and the full Source entities fetched once per venue; the two-year mean citedness (OpenAlex's open JIF analogue) and h-index were written back onto each work's recorded source. Venues with no reported summary statistics remain unenriched and are excluded pairwise from JIF comparisons only. An author-entity pass supplies the ML-specialist co-author moderator.

B Classification and validation detail

B.1 Label definitions

A paper counts as ML if it trains, fine-tunes, or evaluates a model that learns from data; applies a pretrained model as part of its method; or proposes a new ML algorithm or architecture. Classical statistics, rule-based systems, deterministic simulation, and purely theoretical analysis do not count, nor does a passing mention in related or future work, nor a survey about ML in which the authors run no model. Stage 2 assigns `ML_PRIMARY` (ML is the main contribution or object of study) versus `ML_SUPPORTED` (ML is an instrument serving a non-ML contribution). Papers typed `NON_ML` receive their subtype deterministically and never reach stage 2; a subgroup audit reports any violation. Papers `UNCLEAR` at stage 1 are excluded from all group comparisons; papers `UNCLEAR` at stage 2 remain in the primary comparison but are excluded from the subtype sub-comparison. Escalation Flash→Pro ran at $\sim$1.8% of papers at stage 1 and $\sim$1.2% at stage 2 (2.3% at either stage; 4% in pilot); `UNCLEAR` finals were 41 papers (0.2%). Every prompt is a versioned YAML configuration released with the library.

B.2 Cross-family agreement study

A random subsample of 3,000 papers was re-classified end to end — both stages, same versioned prompts, same 40,000-character text prefix — by `gpt-4o-mini` (OpenAI) and `llama-3.3-70b-instruct` (Meta), neither sharing a model family with the cascade. Unparseable responses were rare and carried through as such: 21 papers (0.7%) for `llama-3.3-70b`, 7 (0.2%) for `gpt-4o-mini`.

At stage 1 the cascade is not an outlier among the three raters (its agreement with each independent family matches their 94.0% agreement with each other) and is not biased toward either label (ML share 41.9%, between `llama`'s 39.5% and `gpt-4o-mini`'s 44.5%). On the 2,794 papers where both independents agree, the cascade matches them 94.8% of the time (95% CI 94.0–95.6), dissenting 144 times; that 5.2% is the tightest bound this design places on cascade-specific stage-1 error. Stage 2 is markedly less reproducible: part is κ's conservatism when a category (ML-supported) is rare, but part is genuine boundary disagreement — `llama-3.3-70b` assigns ML-supported to 29.5% of papers against the cascade's 13.6% and `gpt-4o-mini`'s 12.8%, almost entirely one-directional (183 moved primary→supported, 9 the reverse). Where the two independents agree ($n = 892$), the cascade matches them 94.3% of the time. The practical consequence: the ML/non-ML split every result rests on reproduces at $\kappa \approx 0.87$; the primary/supported split only the sub-comparison rests on reproduces at $\kappa \approx 0.5$.

B.3 Human blind spot-check

A single rater labelled a blind sample of 60 papers for type and subtype — label first, compare second — drawn across three strata (all-models-agree, cascade-dissents, models-disagree) and reweighted to the composition of the 2,973-paper agreement subsample. This estimates accuracy against a

Table 1: Pairwise agreement on the 3,000-paper subsample. Stage 1 is ML/non-ML/unclear (n = 2,973 after unparseables); stage 2 is ML-primary vs. ML-supported among papers the cascade assigned a definite subtype (n = 1,099). κ interpretation per Landis & Koch (1977).

Stage	Rater pair	Agreement	Cohen's κ (Cohen, 1960)
1	cascade vs. `llama-3.3-70b`	93.5%	0.865 (almost perfect)
1	cascade vs. `gpt-4o-mini`	90.8%	0.813 (almost perfect)
1	`llama` vs. `gpt-4o-mini`	94.0%	0.877 (almost perfect)
2	cascade vs. `llama-3.3-70b`	82.5%	0.502 (moderate)
2	cascade vs. `gpt-4o-mini`	89.4%	0.537 (moderate)
2	`llama` vs. `gpt-4o-mini`	81.2%	0.458 (moderate)

human standard, the quantity the cascade's validity ultimately rests on: agreement between models bounds stability, not correctness, since models can agree and all be wrong.

Under the rater's applied interpretation — classifying by subject matter — stratified cascade accuracy was 79.4% (95% CI 67.7–89.2). Recoding the eleven ML-topic review articles as `non_ml`, per the pre-specified rule that discussing machine learning does not constitute machine-learning use, gives 85.4% (95% CI 74.8–94.3); the difference between the two figures arises solely from those eleven papers. Per-stratum accuracy was 86.7% (26/30) where all models agreed, and 20% in each of the two oversampled disagreement strata (4/20 and 2/10). Cohen's κ on this sample is 0.114 and is not interpretable as an agreement statistic, since the disagreement strata are deliberately oversampled. The sheet, the key, and the scoring script are released with the corpus.

Informal review of model outputs with evidence visible was conducted during development as a sanity pass and is reported as such; a rater who can see the model's answer is not blind to it.

C Failure-mode checks and adequacy rules

C.1 Judgment checks

For three failure modes a language model returns the verdict directly with a confidence rating, a verbatim supporting quotation, and a brief rationale; each prompt states that absence of evidence is `UNCLEAR`, not `FAIL`, and that verdicts must not be inferred from venue reputation. *No held-out test set* (`kapoor_L1.1`): whether reported performance is measured on data excluded from both training and tuning; k-fold cross-validation counts only if no tuning or feature selection preceded it on the full dataset, else `UNCLEAR`. *Illegitimate features* (`kapoor_L2`): whether any input feature proxies for or derives from the outcome label — information unavailable at genuine prediction time. *Inappropriate performance metrics* (`kapoor_bad_metrics`): whether reported metrics can conceal weak performance or fail to support the actual claim, with task-family scaffolding (accuracy alone under imbalance; discrimination alone for a probability claim with no calibration measure (Brier, 1950); correlation for an agreement claim where Bland–Altman or a concordance coefficient is indicated (Bland & Altman, 1986); single-operating-point metrics for a general discrimination claim; unnormalised regression error; overlap-only segmentation metrics where the claim depends on boundaries; classification metrics on censored time-to-event data; detection mAP without its IoU threshold).

C.2 Extraction-plus-rule checks

For the remaining two failure modes the model extracts only and never judges adequacy; the verdict is computed deterministically in code. *Class imbalance*: the model reports the target's class distribution exactly as stated (empty for continuous targets); a paper is flagged when the smallest class is under 10%. *Insufficient test-set size*: the model extracts metrics, test-set sizes, per-class breakdowns, comparators, dispersion statistics, and claim type, each with a verbatim snippet; values absent from the text return `NOT_REPORTED`, kept distinct from a missing answer, and estimation is forbidden. Both ran over a 1,700-paper subset; applicability gating leaves analysable denominators of 436 (test-set size) and 819 (class imbalance) across both arms, and 292 and 375 respectively within the ML

corpus, which is what Table 2 reports. The deterministic logic is covered by 66 of the library's 128 unit tests.

C.3 Dual extraction and conservative reconciliation

Every extraction ran twice (`gemini-2.5-flash` and `gemini-2.5-pro`) and was reconciled before any rule was applied — a stability check, not independent replication, and stated as such. Numeric values within relative tolerance 0.02 are averaged; values outside tolerance take the more conservative reading (smaller sample sizes; larger dispersions); categorical disagreements route to `UNCLEAR` with a flag; a confidence interval is credited only if both passes report one; repeated metric names retain the entry hardest to beat. Transport failure after five retries is recorded as `API_FAIL`, distinct from extraction failure, so infrastructure problems are never degraded into substantive `UNCLEAR` verdicts.

C.4 Test-set adequacy

Adequacy is assessed against the paper's own claim using intervals reconstructed from reported statistics ($z = 1.96$); reported intervals are used when present, and a paper whose reported quantities cannot bind an interval is recorded not-assessable rather than assigned a verdict. Proportion metrics use the Wilson score interval (Wilson, 1927); AUROC uses the Hanley–McNeil variance (Hanley & McNeil, 1982), assigning the smaller class count to the positive class where unidentified (slightly conservative); multiclass problems bind the binomial on the smallest class; otherwise a normal interval from the standard error of the estimate (never from the outcome standard deviation, which would yield an authoritative-looking but incorrect interval). The rule then branches on claim type: standalone precision (interval width ≤ 0.10 at stringency 1.0), superiority (interval separation in the claim's direction, inverted for lower-is-better), non-inferiority/equivalence (interval within the paper's own stated margin; a margin-less claim is unassessable). The paper-level verdict is the worst case across reported metrics.

C.5 Prevalence and subtype tables

Table 2: Failure-mode prevalence within the ML corpus. n = ML papers receiving a definite verdict on that check (analysis-of-record rows).

Check	n	**Failures**	**Rate**
Inappropriate performance metrics	11,606	4,678	40.3%
No held-out test set	9,655	2,854	29.6%
Illegitimate features	11,114	1,493	13.4%
Class imbalance	375	111	29.6%
Insufficient test-set size	292	271	92.8%

Table 3: Failure-mode rates by ML subtype. p_{adj} is Benjamini–Hochberg corrected across the five checks.

Check	**ML-primary**	**ML-supported**	**OR**	p_{adj}
Inappropriate performance metrics	40.9%	37.9%	1.14	0.018
No held-out test set	25.9%	45.7%	0.42	4.0×10^{-60}
Illegitimate features	13.7%	12.5%	1.11	0.16
Class imbalance	30.9%	25.0%	1.34	0.29
Insufficient test-set size	94.1%	84.6%	2.88	0.075

D RELEASE AUDIT DETAIL

D.1 SANDBOXED EXECUTION

L3 executes each buildable specification as an isolated batch job, one container per repository, in two phases: phase one has network access to clone the repository, install dependencies, and fetch data onto a shared volume; phase two runs the entry command with networking disabled, so untrusted third-party code never has egress. Jobs carry hard timeouts ($\sim$4h CPU, $\sim$8h GPU); stdout, stderr, and artefacts are archived keyed by paper identifier. Where runtime permits, three seeds are run and the median taken; single-seed runs are noted per record.

D.2 REPLICATION TOLERANCE AND OUTCOME TAXONOMY

The tolerance band and outcome taxonomy were frozen before the first execution batch. One run reproduces the paper's headline metric — the number its abstract leads with, as extracted at Tier 2 — and receives exactly one ordinal code (Table 4). Tolerances are metric-family specific: ± 0.02 absolute for bounded $[0, 1]$ discrimination and agreement metrics; ± 0.05 absolute for correlation-type metrics; $\pm 5\%$ relative for scale-dependent error metrics and anything else (flagged for review). Where a paper reports an interval or seed variance wider than our band, the paper's band governs: we do not hold a paper to a stricter standard than it holds itself. The ± 0.02 choice approximates combined seed and hardware nondeterminism for typical biomedical ML models, deliberately generous so that estimated claim inflation measures inflated claims rather than nondeterminism. Every execution record stores reported and reproduced values, seeds, metric family, tolerance, outcome code, and an environment-specification hash.

Table 4: Execution outcome taxonomy. "Replicated" means R0 or R1; R2 is reported separately; RX is excluded from claim-inflation estimates but retained in the funnel, never silently folded into either side.

Code	Name	Criterion
R0	Exact	Within ± 0.005 of reported, or bit-exact if deterministic
R1	Replicated	Within metric-specific tolerance, same direction
R2	Qualitatively replicated	Outside tolerance, central comparative claim holds
R3	Not replicated	Central claim fails, or degradation $> 2\times$ tolerance
RX	Indeterminate	Metric not comparable (data drift, different split, ambiguity)

D.3 HOST ATTRIBUTION

Audit sidecars record the resolved URL only for links that resolved (643: 640 passing L1 plus 3 resolving to no real code), so host class for dead links is recovered from the reason code: `REPO_NOT_FOUND` (1,137 claims) and `NO_REAL_CODE` (3) are returned by the GitHub API and therefore attribute the claim to GitHub, while `LINK_DEAD` marks a non-GitHub URL whose host was not recorded. This yields 1,234 GitHub-hosted claims (94 alive, 7.6%) and 822 non-GitHub claims (546 alive, 66.4%). Among surviving non-GitHub links, archival hosts (`doi.org`, Zenodo, OSF, CRAN, Bioconductor, Figshare, Dryad) account for 281 (51.5%); no archival link was found dead, but because dead non-GitHub links carry no host, archival mortality is bounded above only weakly and is not reported as zero. Of the 94 surviving GitHub links, 59 are in the canonical `owner/repo` form the L2 reader parses; the remaining 35 (gists, sub-paths, and other URL shapes) pass L1 unaudited along with all non-GitHub survivors. The L3 execution outcome is recorded in the execution artefacts archived by the batch runner, keyed by paper identifier, rather than in the audit sidecar itself.

Table 5: Availability-claim survival, distinct papers, tallied from the released audit sidecars. The analysis-of-record row table carries 2,093 rows, 37 of which are duplicates; ordinal-tier regression estimates were computed on those rows.

Tier	Papers	Share of claims
Claim made (L0)	2,056	100%
Link resolves, artefacts present (L1)	640	31.1%
GitHub-hosted, auditable (L1+)	59	2.9%
Code matches the paper's claims (L2)	18	0.9%
Buildable environment spec (L2.5)	1	0.05%
Executes in sandbox (L3)	1	0.05%

D.4 Funnel table

E Statistical analysis detail

E.1 Tests and reporting

Citation counts: negative binomial regression, group of interest plus publication year and subfield covariates, unadjusted and adjusted IRRs with 95% CIs and the fitted dispersion. JIF: Mann–Whitney U (Mann & Whitney, 1947) with rank-biserial correlation (Kerby, 2014), medians and per-group n. Categorical outcomes: χ^2 or Fisher's exact (any expected cell < 5), odds ratios with Cramér's V under one consistent table orientation. Failure-mode families: Benjamini–Hochberg (Benjamini & Hochberg, 1995) at $q = .05$ within comparison, across the five checks, primary and sub-comparison corrected separately; unrunnable tests are excluded from the correction family. Statistical computing: SciPy (Virtanen et al., 2020), statsmodels (Seabold & Perktold, 2010), pandas (McKinney, 2010), NumPy (Harris et al., 2020).

E.2 Why no composite score

Rates are reported per failure type with exact binomial intervals and check-specific denominators. The decision was made in advance on interpretive grounds and the data support it empirically: pairwise co-occurrence between checks runs 0.9–1.3× independence, so a composite would average five distinct constructs rather than aggregate five views of one. A composite is also incomparable across groups in principle because exposure is unequal: papers with more applicable checks have more opportunities to fail.

E.3 Moderators

Author count and ML-specialist co-author were fitted as interaction terms within the adjusted models (the latter fitted on $n = 26{,}849$, the rows carrying a resolved specialist flag); the premium persists in both (IRR 2.10 $[1.89, 2.34]$ and 1.60 $[1.40, 1.82]$); note the interaction models are fitted at their own reference levels and on different n, so neither is a like-for-like strengthening of the 1.78 main estimate. The third pre-specified moderator — citation of key methodological references — requires a reference-list enrichment pass that was not performed and is reported as unevaluated rather than as a null result (Appendix G).

E.4 Audit-tier estimate and the snapshot reversal

The furthest audit tier reached enters the adjusted negative binomial model as an ordinal predictor; the executed stratum is far too small for a binary replicated-versus-not contrast. The complete-audit estimate is IRR 1.36 per tier step ($[0.995, 1.855]$, $p = .054$). An interim estimate on a partial snapshot ($n = 679$) was significantly inverse (0.67, $p = .023$) and reversed on completion; all audit-linked estimates are therefore reported only from the complete set.

F TOOL CONCENTRATION

Per-subfield methodological concentration was computed from the Tier-2 tool extractions as the Herfindahl–Hirschman index and top-five share over subfield × year cells. Concentration falls as subfields mature: the mention-weighted index across 846 cells averages 0.050 under the design-reweighted estimator (0.025 unweighted; mean top-five share 0.29 and 0.23), and in Biomedical Engineering falls from 0.163 in 2010 to 0.011 by 2026 in the analysis of record, while distinct tool counts grow steadily (Figure 8; the figure's independent canonicalisation puts absolute levels lower and suppresses sparse early cells, so those two values are not readable off it). The hypothesis that methodological tooling converges and propagates shared mistakes is not supported. The associated interaction estimate (any-flag on concentration, OR 1.38 $[1.15, 1.66]$) is released for completeness but is exploratory only: its outcome is a composite over unequal exposure, computed across the full cohort rather than within ML, and expressed per unit of an index whose observed mean is 0.067 — far below the range the estimate extrapolates over.

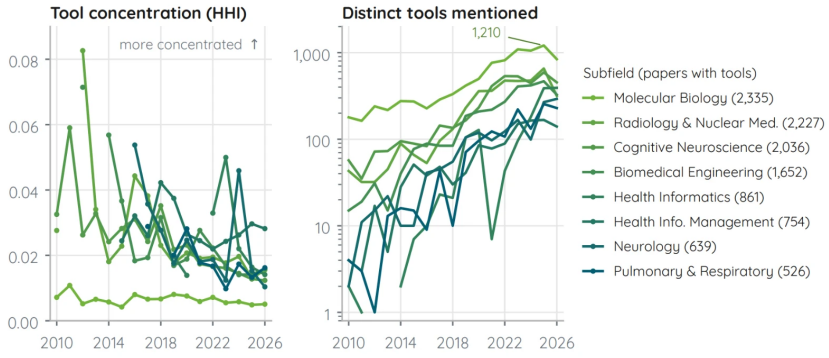

Figure 8: Tool concentration by subfield and year for the eight largest subfields (cells under 40 tool mentions suppressed). Left: Herfindahl–Hirschman index. Right: distinct tools mentioned (log scale). Tool names are matched case-insensitively here rather than through the pipeline's canonicaliser; absolute levels differ from the analysis of record, the trend does not.

G FULL LIMITATIONS

Taxonomy coverage. The scan operationalizes five checks from a single taxonomy (Kapoor & Narayanan, 2023) and reports each as a separate rate. Failure modes catalogued elsewhere — full-dataset preprocessing, hyperparameter leakage across folds, temporal and distributional shift, batch and site confounding (Lones, 2024; Whalen et al., 2022) — are not covered at cohort scale, though the first and third are recorded by the forensic read at $n = 200$. The scan does not speak to questionable research practices in the broader sense (selective reporting, undisclosed multiplicity); that component of the original analysis plan was not built, and is reported as unbuilt rather than as null. A paper passing all five checks has avoided five specific failure modes, not been certified sound. Extending the taxonomy and adding a QRP pass is the most direct future work this study motivates — and the availability funnel now quantifies, rather than assumes, the central obstacle to doing so.

Subset coverage. The extraction-plus-rule checks ran over a $1{,}700$-paper subset, with applicability gating reducing analysable denominators to a few hundred; their null subtype results should be read as underpowered.

Construct validity of the checks. All five checks presuppose a predictive design and offer no `NOT_APPLICABLE` verdict. The expectation that non-predictive papers fall to `UNCLEAR` holds only partially; a substantial share of non-ML papers received definite verdicts, most visibly on the held-out-test-set check, which fires more often on non-ML than ML papers. We therefore report no cross-

arm failure contrasts (those in the released results table are flagged non-interpretable), restrict failure rates to the ML corpus, and leave the held-out check's subtype difference uninterpreted, since ML-supported papers sit between ML-primary and non-ML papers in how central their predictive claims are. Adding an applicability verdict and re-running the affected checks is the cheapest correction available to a future version.

Classifier validation is based on a single rater. The 60-paper blind check was completed by one rater, so no human–human agreement figure is available and the ceiling on achievable classifier accuracy is unknown. The sample is small ($n = 60$) and the resulting confidence intervals are correspondingly wide, spanning roughly 20 percentage points; the two interpretations reported in Section 3.6 have substantially overlapping intervals and should not be read as distinguishable. Accuracy in the two oversampled disagreement strata was 20%, and because those strata carry low weight, the headline figure is driven mainly by the 30 all-agree papers.

The ML/non-ML definition was underspecified for review articles. The classification rule stated that merely citing or discussing machine-learning work does not constitute machine-learning use, but did not address review and survey articles explicitly. In the blind check the rater classified reviews by subject matter — treating reviews *about* machine learning as `ml` and reviews on other topics as `non_ml` — while the cascade applied the method-use rule. Review, essay, and non-article items account for 15 of the 28 rater–classifier disagreements. Both figures in Section 3.6 are reported for this reason. The corpus itself was not filtered to exclude reviews: the sampling frame required `type:article`, which in OpenAlex does not separate review articles from primary research, and no document-language filter was applied, so a small number of non-English items are present. Prevalence estimates should be read as applying to the corpus as sampled, which includes review articles.

The classifier. Every label is produced by proprietary hosted models. Classification is a single-family confidence-gated cascade — a weaker independence guarantee than a two-family agreement gate, since correlated failure within a family is more plausible; the extraction checks' two passes are likewise tiers of one family and are a stability check, not replication. Validity rests on the blind human spot-check reported in Section 3.6, with cross-family stability at $\kappa \approx 0.87$ (stage 1) and ≈ 0.5 (stage 2). Ambiguity is routed conservatively (`UNCLEAR`, not `FAIL`; not-run recorded as not-applicable). Because an `UNCLEAR` verdict removes a paper from the denominator as well as the numerator, this makes the count of detected failures a floor rather than the rate itself: a failure described in hedged or unconventional language is more likely to be dropped as untested than miscounted as a pass. Model behaviour is a moving target: results reflect the specific snapshots enumerated in Appendix B and carry no multi-seed stability estimate for single-pass stages.

A citation-based moderator was pre-specified but not evaluated. The analysis plan included `cites_key_methodological_refs` — whether a paper cites key methodological or reporting-standards references — as a candidate moderator. It was not evaluated. The reference set was never operationalised: which works should count as "key methodological references" across many biomedical subfields was not resolved, and the configuration was left empty rather than populated with an arbitrary list, with the field held at `None` rather than defaulting to `False`, so no reported result treats "unconfigured" as "does not cite". A post-hoc check indicates the moderator would also have been underpowered. Reference lists are available via OpenAlex for approximately 82% of cohort papers (median 39 references per paper), but citation to any individual candidate reference is rare within this corpus: of $49{,}179$ papers, 11 cite Kapoor & Narayanan (2023) and 21 cite the CONSORT-AI reporting guidelines, with pooling across plausible candidates yielding on the order of 40 exposed papers. Any future attempt would additionally face non-random missingness in reference availability, which tracks publisher deposit practice and conflates "did not cite" with "no reference data", and the fact that citing a methodological reference is not evidence of adherence to it.

Cohort attrition. Only papers with a retrievable OA PDF and successful extraction enter analysis ($27{,}200$ of $\sim 49{,}000$ sampled); exclusions are recorded, not imputed. Truncation by host rate-limiting is non-random, but round-robin ordering preserves stratum balance, so the effect is wider intervals rather than differential composition. OA availability is not independent of venue or era —

outcomes this study analyses — so results generalize to fetchable OA biomedical literature. Cohort membership inherits OpenAlex's subfield assignments and their errors. The analysis of record contains $\sim$1.7% duplicated rows (2.6% of ML vs. 1.2% of non-ML papers, since repaired); identical verdicts leave within-group rates unaffected while inflating the ML arm's row count, so row-level analyses slightly overstate precision.

The audit. The no-rescue standard makes the funnel a lower bound on what released artefacts could yield under effort (an explicit extrapolation is reported); conversely, the L2+ readers deep-audit only GitHub claims in canonical `owner/repo` form (59 of 640 surviving links), so measured survival below L1 is specific to that subset. Host class for dead links is recovered from reason codes rather than recorded URLs, and archival mortality is consequently bounded, not measured (Appendix D). Extending the readers to archival hosts is the highest-value extension.

Claim inflation is claims-based. The forensic read records what papers report; the true tax — reported minus faithful rerun — requires re-execution at a scale this study does not reach. Because the screening flags and the forensic leakage construct do not coincide, contrasts against screening flags are rigor-flag contrasts, not measured leakage taxes. The corpus-wide meta-analytic version is further limited by shallow-extractor yield ($n = 86$).

Measures and design. Venue quality uses OpenAlex two-year mean citedness — an open JIF analogue, not field-normalized; venues without summary statistics are excluded pairwise. The comparisons are observational: frequency-matched at design, regression-adjusted at analysis, neither individually matched nor randomized; residual confounding cannot be ruled out. Test-set adequacy marks under-reporting as not-assessable rather than negative; to the extent under-reporting correlates with rigor, prevalence is understated.