Reproducibility in Computer Science: An Empirical Multi-Year Replication Study of Neuro-Symbolic AI

Brandon Colelough*, Vladimir Martirosyan, Ishan Tamrakar, William Regli, Aditya Kumar,
Anh N. Nhu, Dhruv Dubey, Raj Ambavane, and Haowei Deng
Department of Computer Science, University of Maryland
College Park, Maryland, USA
*brandcol@umd.edu

***Abstract*—The current empirical Neuro-Symbolic AI (NSAI) literature is not reproducible enough to support cumulative science and empirical NSAI papers should be required at submission time to provide complete, versioned, and permanently archived artifact bundles. Reproducibility is the minimum standard for good science, yet the same-artifact rerun reproducibility of Neuro-Symbolic AI studies is almost nonexistent. We executed a multi-year audit that began with 5497 search hits, removed 3018 duplicates, and screened 2479 unique records. Title-and-abstract screening identified 1365 self-identified NSAI records. Full-text eligibility screening removed 61 records for off-topic, non-research, no-quantitative-evaluation, or inaccessible-full-text reasons, leaving 1304 eligible records. Of these, 849 lacked a verifiable public code artifact and 455 entered the same-artifact rerun audit. We fully or partially reproduced 85 studies (18.68 % of attempted reruns; 6.52 % of the eligible accessible-full-text audit corpus). 321 attempted reruns were blocked by missing non-code artifacts and 42 by missing or unusable code repositories. These figures quantify a persistent reproducibility deficit that survives even nominal "code available" declarations, and signal the need for enforced, versioned, and permanently archived artifact bundles in future NSAI publications.**

I. Introduction

Empirical Neuro-Symbolic AI papers should be required at submission time to provide complete, versioned, and permanently archived artifact bundles—including source code, data, model checkpoints or weights, environment specifications, evaluation scripts, and documentation, or a documented legal or ethical exception—because nominal code availability is insufficient for cumulative science. In this paper, we undertake a study of the same-artifact rerun of published results in the emerging field of Neuro-Symbolic Artificial Intelligence (NSAI). Specifically, we ask whether an independent team can rerun the computational pipeline released with a paper and recover the main reported result within a prespecified tolerance. NSAI is viewed by many as the emerging frontier that merges concepts from connectionist methods with those referred to informally as "Good Old Fashioned AI", consisting of logic and formal techniques for reasoning. The frontier science that aims to integrate these two techniques is highly dynamic, and while the number of papers that claim to be creating contributions in this area is exploding in number, this paper reports that the results in the vast majority of the papers are difficult or impossible to replicate. This indicates that the research area comprising "Neuro-Symbolic AI" is experiencing a crisis of replication and lacks the accepted community norms to enforce replicability as part of the publication of work. For the field to advance into the position of critical importance that many feel is inevitable, we should raise the bar on our scientific standards and require greater emphasis on how we can document the reproducibility so that others may verify our claims. We conclude the paper with some general recommendations based on our findings during this survey.

II. Background

As has been documented often in the recent literature, many areas of science are currently experiencing a "reproducibility crisis" [1] as studies have proven difficult or impossible to replicate [2], data sources are not available [3], and experimental assumptions are not made explicit [4]. Social sciences, in which these phenomena have been documented most extensively, have the additional issues posed by post hoc redesign of scientific hypotheses. Known as "Hypothesizing After the Results are Known", or "p-hacking", this occurs when a researcher forms or rewrites a hypothesis after seeing the data, and then presents that hypothesis as if it were specified before the data were collected [5]. The field of computing, in theory, should be highly reproducible, as algorithms, code, data, and other artifacts can be easily shared and adopted. Conferences and journals have begun to require data sharing and other best practices to improve reproducibility [6]. Without reproducible artifacts, computational research risks becoming unfalsifiable claims that cannot be independently verified and thus fall outside the bounds of science. We need to be able to rigorously and independently test hypotheses laid out in papers in the computer science community in the same way we do in other domains, so that the work can have more credibility.

A. Research Objective and Main Research Questions

Throughout this paper, reproducibility refers to obtaining substantially similar results using the original authors' released artifacts. Our Research Objective is ***To provide the first domain-wide code-level assessment of reproducibility in the Neuro-Symbolic AI literature, quantifying how often published systems can be (re)executed successfully and pinpointing the main drivers and barriers of reproducibility across time and venues***. To support this main research objective, we aim to fulfill our main research questions, including:

RQ1 What proportion of NSAI papers with publicly-available code can be fully or partially reproduced? ***RQ2*** How does artifact completeness (code / data / model weights) affect the probability of successful reproduction? ***RQ3*** Do reproduction outcomes vary systematically by publication year or venue family (conference, journal, preprint)?

B. Reproducibility Research in Computer Science

Reproducibility failures are well-documented across empirical disciplines. Landmark studies in psychology and medicine have shown that a substantial fraction of published findings cannot be independently replicated [7, 8], and large-scale surveys across the social and biomedical sciences trace these failures to non-disclosed analytical choices, unavailable data, and selective reporting [1, 2, 4]. The interested reader is referred to that broader literature for a full treatment. Computer science presents a structurally different case, as algorithms, code, and data can in principle be shared exactly, meaning empirical claims ought to be among the most verifiable in science. In practice, however, the same failure modes recur, as the studies below illustrate. Vanderdonckt and Vatavu introduced *Amplitum*, a contextual framework that augments generic replication taxonomies with explicit descriptors of participants, devices, and physical settings [9]. Their gesture-elicitation case study reproduced prior findings only after replicators matched the original laboratory environment and user cohort, underscoring that code and stimuli alone are insufficient when human behavior is the main dependent variable. Ferrari Dacrema *et al.* inspected 26 "state-of-the-art" (for the time) neural recommender papers (2015–2018) and were able to fully reproduce only 12 of the original 26, and in doing so discovered that *eleven of those twelve* fell behind well-tuned neighborhood, matrix-factorization, or sparse-linear baselines once a common experimental protocol was enforced [10]. Ferrari's analysis attributes the apparent performance gains to methodological weaknesses, including baselines left at default settings, inadvertent test-set leakage during epoch selection, and undocumented data-split choices. The 2023 ReproNLP shared task enlisted independent teams to reproduce evaluation metrics for forty-eight recent NLP systems using the artifacts released by the original authors. Fewer than half of the replication attempts matched the figures reported in the papers [11]. Divergences were traced to undocumented tokenisation, stochastic initialization, and missing pre-processing scripts. Henderson *et al.* evaluated several widely cited DRL algorithms and found that benchmark rankings could reverse when random seeds, hardware platforms, or training horizons were varied [12]. Their study therefore recommends reporting results over many seeds, applying formal significance tests, and disclosing every experimental detail. Pawlik *et al.* examined the longevity of public datasets and found that many links degrade, move, or silently change versions, rendering later experiments irreproducible despite nominal accessibility [13]. They argue for immutable storage, rigorous version identifiers, and provenance metadata. Replication breaks down whenever any piece of the experimental context is missing, and Neuro-Symbolic AI is

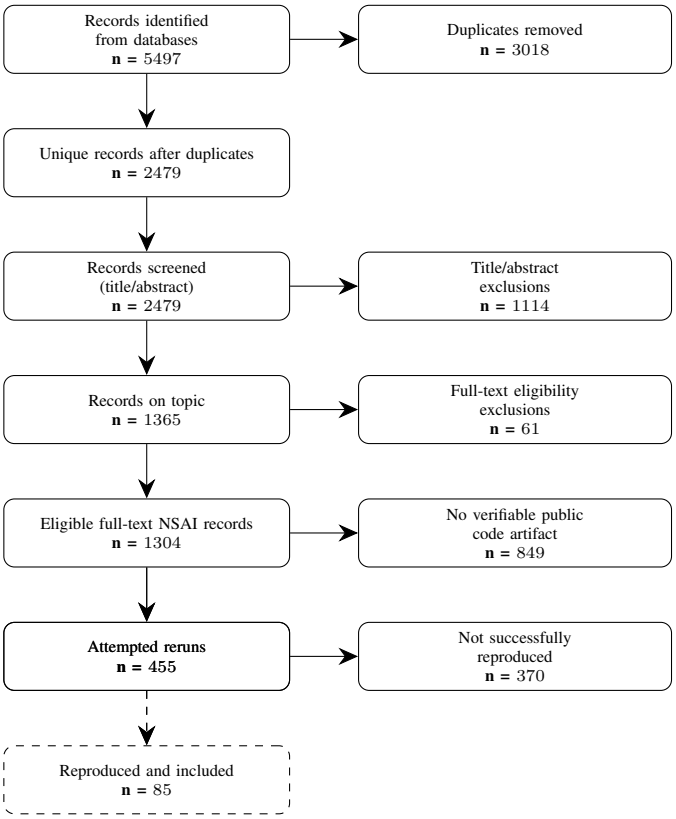

Fig. 1: Selection and eligibility flow for the NSAI audit. Late full-text eligibility exclusions are shown separately and are not counted as rerun outcomes. The dashed bottom box reports the number of successful reruns within the attempted set and is shown as an audit outcome.

no exception. An open-code link without hardware details, solver commits, pre-processing scripts, dataset splits, and hyper-parameter schedules offers little more than performative compliance. Gains attributed to the symbolic–neural fusion may vanish once baselines receive equal tuning or a knowledge base is revised; stochastic variation in the neural component could dominate the symbolic layer as well. Reliable reproduction, therefore, requires immutable, versioned artifact bundles, multi-seed evaluation, and full provenance for pretrained weights, logic programs, and curated knowledge graphs. Within the domain of computer science, empirical claims acquire the status of *science* only when independent researchers can interrogate and rerun the full computational pipeline. A lone GitHub URL is therefore insufficient, and authors must accompany source code with the complete suite of artifacts, **datasets, trained model checkpoints, environment manifests, and code documentation including hyper-parameter information**. Absent this level of disclosure, reproducibility deteriorates, methodological scrutiny becomes impossible, and the cumulative progress of the field stalls.

III. METHOD

A. Corpus Construction & Screening

This reproduction study targets same-artifact rerun reproducibility. This reproduction study began with a deliberately broad bibliographic sweep designed to capture the full breadth of the Neuro-Symbolic AI domain. Guided by a PRESS-validated query [14] centered on "neuro-symbolic OR NeSy OR NSAI", we queried nine major digital libraries—WEB OF SCIENCE, SCOPUS, PUBMED, EI COMPENDEX, IEEE XPLORE, ACM DL, SPRINGERLINK, GOOGLE SCHOLAR, and ARXIV on 23 May 2025, retrieving 5497 records. This broad sweep increased coverage of emerging or lexically idiosyncratic work that narrower queries often overlook. A de-duplication pipeline removed 3018 duplicates (55% of raw hits), leaving 2479 unique records. The substantial overlap across sources illustrates the cross-posting norm in NSAI research and underscores the need for multi-pass de-duplication.[1] Relevance screening was performed in a single-blind title-and-abstract pass. After a 20-record calibration, reviewers screened titles and abstracts and retained any paper whose authors explicitly described the work as "neuro-symbolic," regardless of application domain. 1365 papers satisfied this broad criterion. Code availability was not inferred from the abstract alone. Instead, repository verification was conducted during full-text eligibility assessment wherein annotators first inspected the full paper for repository or artifact links, and when none were present, they performed a structured external search using the paper title, author names, and method or domain keywords. This process eliminated 849 records for which no verifiable public code artifact could be identified and left 455 code-bearing studies for replication, an attrition of $70.55\,\%$ that exposes the gap between open-science claims and delivered artifacts. The stages reduced the literature from 2479 unique records to 1365 NSAI-relevant papers and further to 455 replication candidates that at minimum linked or could be matched to a code artifact. Figure 1 visualizes each reduction step, linking corpus construction directly to our objective of quantifying verifiable progress in Neuro-Symbolic AI.

B. Replication Protocol

1) Inclusion criteria: A paper entered the reproduction pipeline only when it satisfied *all* of the following; IC_1 **- Neuro-symbolic integration** – the study is self-described as Neuro-Symbolic. IC_2 **- Empirical evaluation** – the paper reports quantitative results on benchmarks, real-world datasets, or synthetic tasks and compares against baselines or ablations.IC_3 **- CS relevance** – the work contributes technical insight into the CS domain IC_4 **- Auditable code claim** – the paper provides a direct repository link or a uniquely identifiable public code artifact for the reported system, sufficient to permit artifact audit. pipeline. IC_5 **- Full text Available** – the paper has an accessible full text for audit.

[1]Pipeline tools: EndNote, Covidence, Zotero, SR-Accelerator, and Rayyan.

2) Exclusion criteria: Papers were removed if they violated *any* of the following; E_1 Not written in English. E_2 Literature review, review, survey, editorial, or otherwise not original empirical research. E_3 No verifiable public repository or archival code artifact for the reported system could be identified from the paper or via a structured external search. E_4 Missing indispensable artifacts required to rerun the primary experiment could not be identified, accessed, or reconstructed under the study protocol. E_5 Lacks quantitative evaluation. E_6 Outside the scope of neuro-symbolic methods. E_7 Duplicate, superseded, or version-of-record already retained. E_8 No full-text access (pay-walled or retracted). To avoid conflating sample selection with audit outcomes, we treat eligibility exclusions and post-entry rerun results as analytically distinct. Papers failing E_1–E_8 were removed during title and abstract screening (or eligibility assessment) and therefore did not enter the attempted rerun pool. For papers that entered the attempted rerun pool, we assigned outcome labels indicating the result of the same-artifact reproducibility audit rather than making additional exclusion decisions. E_1–E_8 are eligibility exclusions only and were applied during full-text eligibility assessment. Failures observed after entry into the attempted rerun pool were recorded as audit outcomes, not as exclusions. Counts for screening and eligibility exclusions are reported in

3) Reproduction Procedure: Post-entry rerun outcomes are reported in Figure 2. Every study with a publicly accessible repository was evaluated under a five-stage protocol that compressed the workflow into discrete, auditable checkpoints [2]. All annotators began with a one-hour onboarding workshop[3] that introduced the replication workflow and replication template. This was followed by two 4-hour live sessions (covering environment builds and dependency management, metric verification, and licensing constraints) and a 3-hour supervised drop-in lab in which each participant could reproduce exemplar studies end-to-end in a supervised environment. All replication attempts and artifact assessments were conducted by a team of eight trained graduate student annotators over a nine-month period, who completed the study's standardized onboarding workshop and calibration protocol. On average, each annotator conducted 64 replication attempts, and reproduction workload assignments were allocated randomly throughout the team. A ten-paper calibration pilot produced Cohen's $\kappa = 0.82$ (a standard measure of inter-rater agreement, where 1.0 is perfect agreement and values above 0.80 are considered strong), and outstanding disagreements were reconciled in group discussion. During the nine-month replication phase, the team met every second week to review edge cases and realign on the reproduction protocol. Outcome labels were assigned after entry into the attempted rerun pool as follows: **O1) Fully reproduced.** The primary experiment was executed successfully, and the reproduced result satisfied the study's fidelity criterion. **O2) Partially reproduced.** The core pipeline executed, and the paper's main qualitative claim was preserved,

[2]NSAI survey index (posted 4 Apr 2026)

[3]REMOVEDFORREVIEW

but one or more quantitative results fell outside the acceptance band, or only a subset of the headline experiments could be rerun. **O3) Executed but did not reproduce within tolerance.** The system ran to completion, but the reproduced results materially exceeded the acceptance threshold or contradicted the paper's main quantitative claim. **O4) Not executable due to missing or inaccessible artifacts.** The attempted rerun could not proceed because one or more indispensable artifacts were unavailable, inaccessible, or not reconstructable under the study protocol. **O5) Not executable due to environment or code failure.** Attempted rerun failed because the released code or environment could not be built or executed under the protocol despite the permitted minimal fixes. For this study, an artifact was classified as missing only when it was an indispensable input to the primary experiment and could not be reconstructed from the paper and released materials under the study protocol. Indispensable inputs included fixed datasets or splits, preprocessing scripts that alter data semantics, checkpoints or weights when evaluation depended on a fixed trained model state, rule sets or knowledge bases, configuration files, environment descriptors, and evaluation assets. All reproduction logs and extracted data are version-controlled, and a consolidated record is available in the public spreadsheet.

C. Evaluation Design

We define *reproduction success* only for papers that entered the attempted rerun pool. A successful rerun required satisfying the O1 or O2 criteria, and papers failing earlier eligibility checks were not part of this evaluation set. Specifically, a repository must execute under the author-supplied environment (or a minimally updated equivalent), with a max 7 day wall-clock time and on any amount of compute required as specified by the author, and reproduce the paper's primary metric to within $\pm 5\,\%$ absolute error (or inside the authors' $95\,\%$ confidence interval), yield outputs consistent with the paper's headline claims, and require no correction of bugs intrinsic to the model architecture. Annotators were authorized to (i) update deprecated package versions or apply path fixes, provided it is minimal, (ii) adjust file paths, (iii) supply a lightweight evaluation harness when none was provided, and (iv) patch minor scripting errors. Malformed or undocumented environments, missing indispensable post-entry artifacts, or architecture-level defects were recorded as O4 or O5 audit outcomes, not as exclusions.

IV. Results

A. System-Level Reproducibility

Figure 2 reports outcomes for the attempted rerun set only ($n = 455$). Within this set, **48** studies ($10.55\,\%$) were fully reproduced and **37** ($8.13\,\%$) were partially reproduced, yielding **85** successful reruns overall ($18.68\,\%$). Failures within the attempted set comprised **321** papers ($70.55\,\%$) blocked by missing non-code artifacts, **42** ($9.23\,\%$) blocked by unavailable or unusable code repositories, and **7** ($1.54\,\%$) that had the nominal artifact set but failed because of environment or code defects. The **61** records that failed late full-text eligibility checks are reported separately in Figure 1 and are not counted as rerun outcomes. Therefore, the rerun success rate was $\frac{85}{455} \times 100 = 18.68\%$. and relative to the eligible accessible-full-text NSAI audit corpus ($n = 1304$), the corresponding success rate is $\frac{85}{1304} \times 100 = 6.52\%$. The primary obstacle to successful same-artifact reruns was incomplete artifact availability. Within the attempted set, **321** papers were blocked by missing non-code artifacts, and **42** by missing or unusable code repositories. A residual **7** studies shared the nominal artifact set yet still failed because of environment or code defects. By contrast, when code, data, and weights were all available, reruns succeeded in **85** of **92** cases ($92.4\,\%$). As shown in Figure 2, missing artifacts concentrated in a few recurrent components which included absent model checkpoints or weights affected **89** papers ($19.56\,\%$ of attempted reruns), environment or specification issues affected **79** ($17.36\,\%$), and incomplete codebases affected **67** ($14.73\,\%$). These three categories account for **235** of **321** missing-artifact failures ($73.2\,\%$), leaving **86** failures distributed across missing datasets (**36**, $7.91\,\%$), missing or incomplete documentation (**25**, $5.49\,\%$), and other missing items (**25**, $5.49\,\%$). The skew implies that most failures arose from predictable artifact-release gaps.

B. Year-Wise Reproducibility Patterns

Figure 3 tracks NSAI reproduction outcomes by publication year and shows that, despite a steep rise in paper volume from single digits before 2019 to 140 publications in 2024 and 92 in the partial-year 2025 cohort. The proportion of studies reproduced each year has remained essentially flat, fluctuating between $11.1\,\%$ (2019) and $22.0\,\%$ (2025) with intermediate values of $10.0\,\%$ (2020), $16.7\,\%$ (2021), $18.2\,\%$ (2022), and $17.9\,\%$ (2024). Because the 2025 cohort covers only January–May, its denominator is still evolving; we therefore treat the $22.0\,\%$ figure as provisional and exclude 2025 from formal trend tests. Across the entire time span, the proportion of studies reproduced (fully or partially) remains in a narrow band of roughly between $10\,\%$ to $22\,\%$ of the yearly output, and the stacked-bar profiles reveal no sustained upward trajectory. In practical terms, the field is publishing substantially more work each year, but the likelihood that any given paper can be reproduced under our protocol has not materially improved, suggesting that the surge in publications has not, by itself, improved real-world reproducibility under our protocol.

C. Venue-Level Impacts on Reproducibility

Figure 4 compares reproduction outcomes across twenty-seven venue families and shows no systematic advantage for journals, conferences, or preprint servers once sampling variation is taken into account. Large outlets illustrate this point as arXiv achieves **16/108** successes ($14.8\,\%$), ACL/EMNLP conference proceedings **11/55** ($20.0\,\%$), IEEE journals **4/33** ($12.1\,\%$), and NeurIPS **5/28** ($17.9\,\%$). Comparable rates are observed for AAAI family journals/conferences ($20.8\,\%$) and for ICLR ($23.8\,\%$) despite differing review models, while domain-specific journals, vision conferences, and publisher platforms all cluster in the low-to-mid-teens. The only apparent

outlier is ICML at **7/15** (46.7 %) however, its small denominator limits generality. Aggregating by broad class confirms the overlap as conferences overall reproduce at ≈ 18 %, journals at ≈ 15 %, and preprints at ≈ 13 %. In short, publication venue alone is not a reliable predictor of NSAI reproducibility, reinforcing the conclusion that artifact completeness, rather than outlet type, drives successful replication.

D. Failure modes behind missing artifacts

Figure 5 assigns each of the $n = 363$ papers that were non-reproducible due to incomplete artifact provision to a missing artifact type and a primary cause. The largest cell is missing model weights with files missing with no reason provided (79 papers), and the second is missing environment specifications linked to build or install failure from unpinned dependencies (61 papers). For model weights, restricted access and dead links are rare (1 and 6 papers, respectively), so the bottleneck is usually not permissioning but non-release or decay, and the claim that a repository enables rerunning the reported pipeline is falsified by inspection in the majority of weight-missing cases. Environment failures concentrate on dependency drift, which is consistent with repositories that omit a locked descriptor or ship one that no longer resolves under current tooling. Code and repository issues are also common, with files missing with no reason provided (36) and dead links or removed artifacts (27), plus a substantial remainder attributed to other causes (26), so "code available" is an unreliable proxy for an executable codebase. Data failures split between restricted access (11) and unexplained absence (22), while documentation failures cluster in missing or unclear instructions (18) and build failures (14).

E. Citation impact vs reproducibility

Figure 6 demonstrates whether citation impact is a useful proxy for practical reproducibility by comparing DOI-resolved citation counts across the same reproduction outcome and exclusion buckets shown in figures 2 and 5 above. The citations of $n = 450$ manuscripts were obtained via bibliographic API-based lookups (resolved via manuscript DOI), and we omitted 25 papers that lacked DOIs, as well as papers that were excluded for precursor reasons (such as background articles, etc.). The distribution of citations from the remaining 450 papers, as shown in Figure 6, indicates substantial overlap across all buckets, with a large mass of zero and low citation papers in every outcome and a long right tail that inflates the average number of citations per bucket. The fully reproducible bucket has the highest mean citation count at 14.1 but a median of 2, and comparably high means occur in dominant non-reproducible buckets such as missing model checkpoints or weights with mean 10.0, median 1, max 234 and environment or specification issues with mean 10.5, median 2, max 233, showing that incomplete artifact release does not correlate at all with number of citations that a paper may receive.

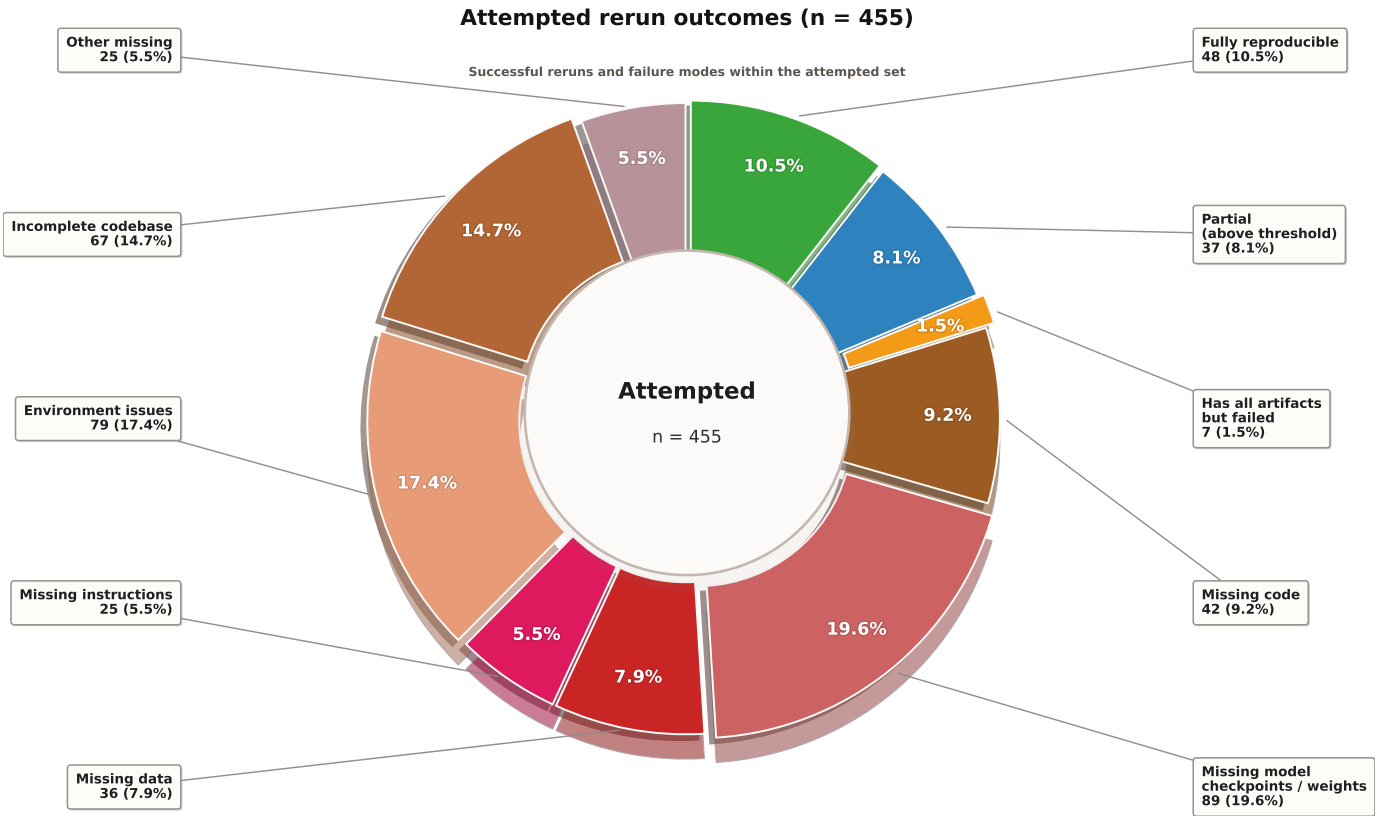

Fig. 2: Rerun outcomes for the attempted same-artifact audit set ($n = 455$).

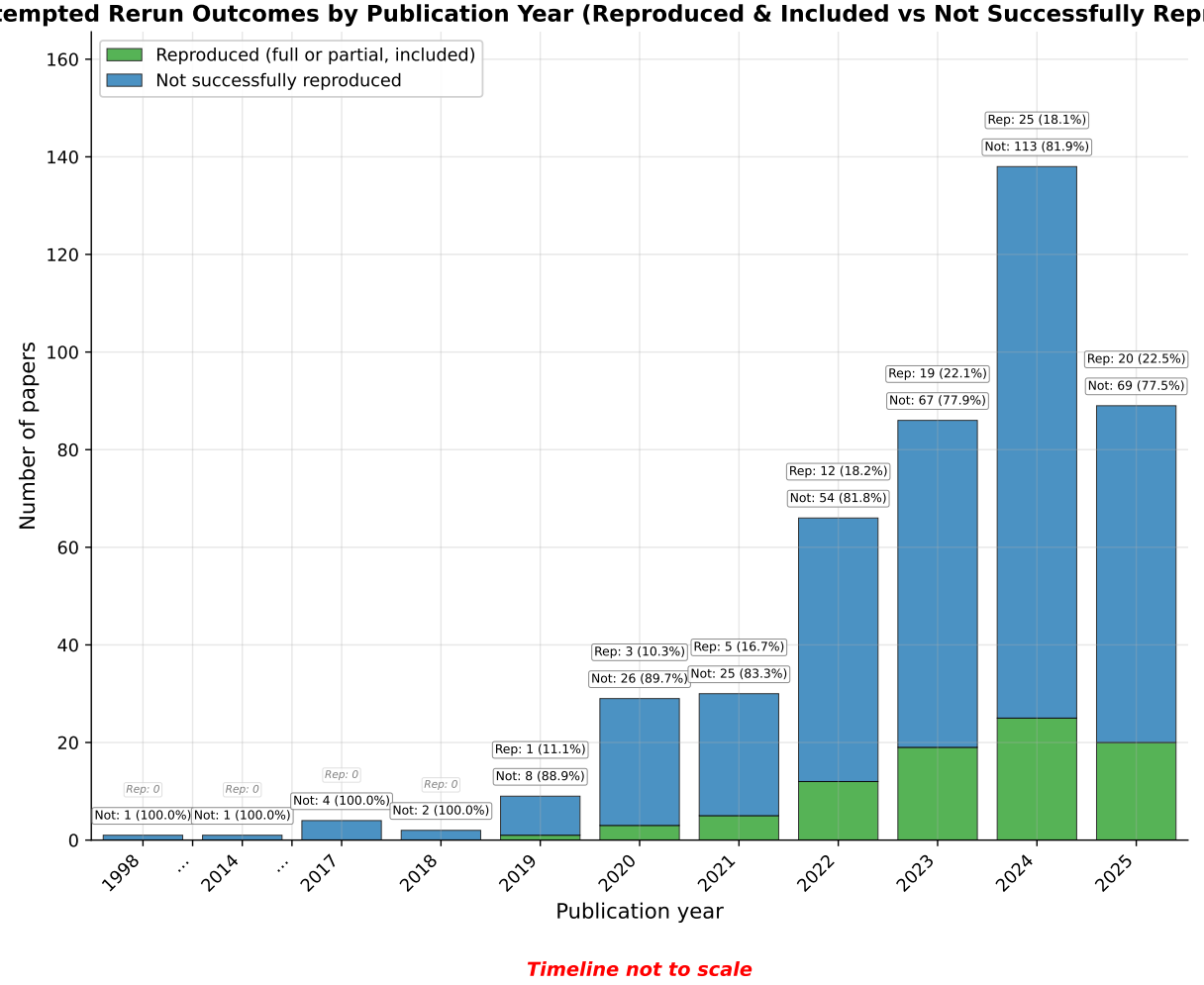

Fig. 3: Reproduction outcomes over time for the attempted NSAI papers. Stacked bars show, for each publication year, the number of papers that were successfully reproduced (full or partial) versus not reproduced under the study protocol, illustrating both the growth of the field and the modest improvements in reproducibility across years.

F. Answers to Research Questions RQ1 – RQ3

For **RQ1** [proportion of NSAI papers that can be reproduced], within the attempted rerun set ($n = 455$), 48 studies were fully reproduced and 37 partially reproduced, giving 85 successful reruns (18.68%). Relative to the full eligible corpus ($n = 1304$), this falls to 6.52%.

For **RQ2** [effect of artifact completeness on reproduction success], artifact completeness is the dominant factor. When code, data, and weights were all present, reproduction succeeded in 85 of 92 cases (92.4%). When artifacts were incomplete, 321 of 455 reruns (70.55%) were blocked, with missing model weights ($n = 89$), environment issues ($n = 79$), and incomplete codebases ($n = 67$) accounting for 73.2% of those failures.

For **RQ3** [variation in outcomes by publication year and venue family], neither publication year nor venue predicts reproducibility. Success rates remained flat between 10% and 22% from 2019 to 2025 despite rapid growth in paper volume, and large outlets cluster in a similar narrow band, with arXiv at 14.8%, ACL/EMNLP at 20.0%, IEEE at 12.1%, and NeurIPS at 17.9%, and aggregate rates of 18% for conferences, 15% for journals, and 13% for preprints.

V. Discussion

A. Summary of Principal Findings in Context

Empirical same-artifact rerun reproducibility across the NSAI audit corpus remains the exception rather than the norm. Of the 455 attempted reruns, 85 were fully or partially reproduced (18.68 %). Relative to the eligible accessible-full-text NSAI audit corpus of 1304 records, this corresponds to 6.52 %. Late full-text eligibility failures are reported separately and are not counted as rerun outcomes. Every study in the attempted set had a verifiable public code artifact at screening, but many repositories were incomplete, decayed, or unusable at audit time.

B. Trends and Patterns

Despite adopting a replication protocol that permits limited dependency updates, path corrections, and lightweight evaluation harnesses to be generated by replicators, the proportion

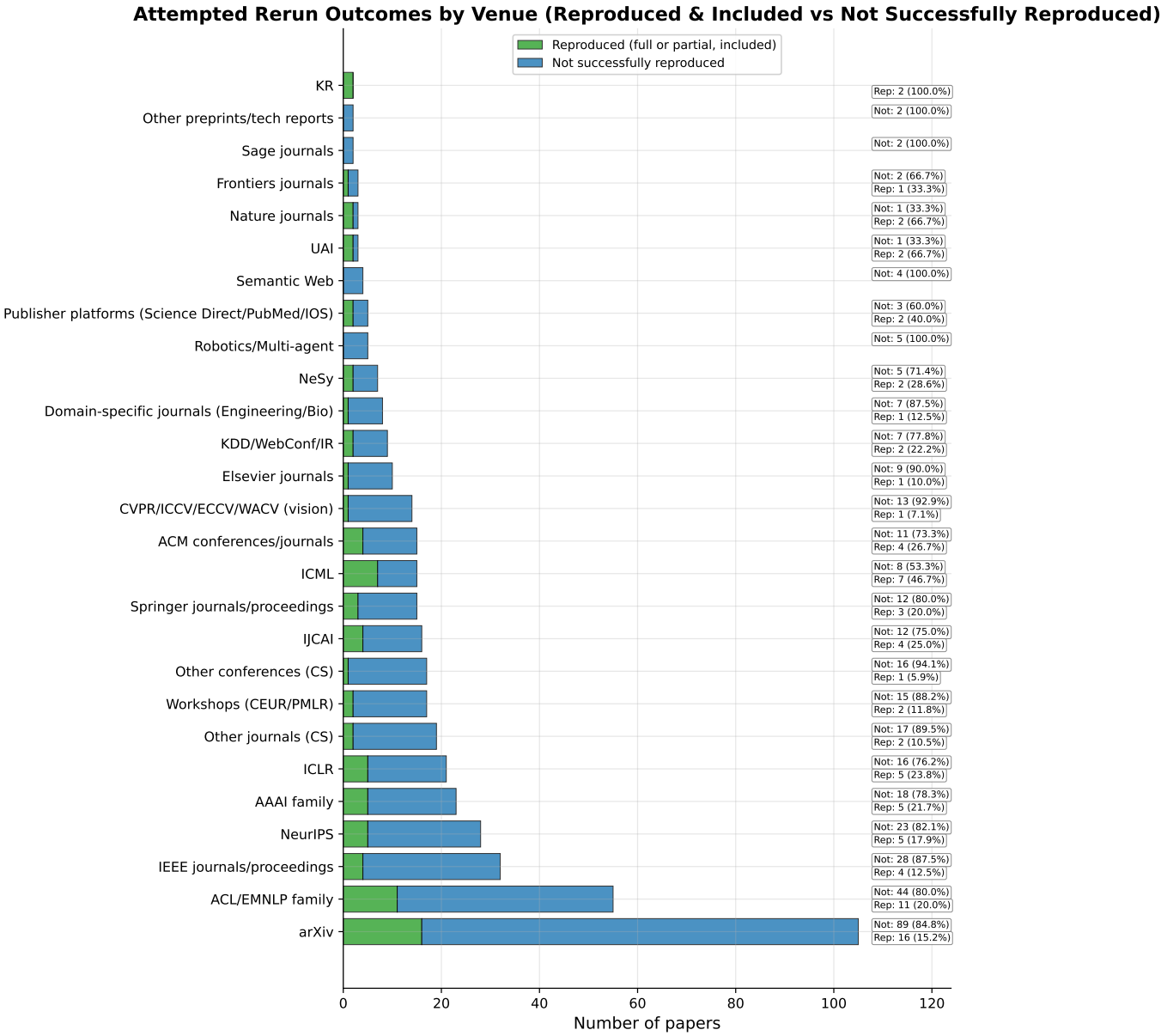

Fig. 4: Reproduction outcomes by venue group for the attempted NSAI papers. Stacked bars show, for each venue family, the number of papers that were successfully reproduced (full or partial) versus not reproduced. Venues are ordered to highlight substantial heterogeneity in reproducibility across conferences, journals, and preprint servers.

of NSAI papers that reproduced remained stubbornly flat, oscillating between $10\,\%$ and $22\,\%$ per year from 2019 through 2025. The funnel diagram in Fig. 2 attributes most failures to missing non-code artifacts ($70.55\,\%$) and, to a lesser extent, to absent code ($9.23\,\%$), yet when the full range of artifacts required for reproduction is present, reproduction succeeds in $93.4\,\%$ of cases. Perhaps most striking is the insignificance of venues as a determinant for whether a manuscript will provide the resources required to reproduce the work from a manuscript as we found that conferences, journals, and preprints cluster at reproduction rates of roughly $18\,\%$, $15\,\%$, and $13\,\%$, respectively, with no statistically meaningful separation (Fig. 4). The evidence points to artifact availability, rather than venue type or methodological novelty, as the main bottleneck to reproducibility in current NSAI research. The analysis, therefore, shifts the conversation from improving experimental technique to enforcing comprehensive artifact release across all publication outlets.

C. The Curious Case of the Missing Codebases

Despite passing the initial NSAI relevance screen and subsequent repository-identification step, 42 of the in-scope papers ultimately fell into the Missing Code category because the referenced repository was dead, private, empty, unrelated,

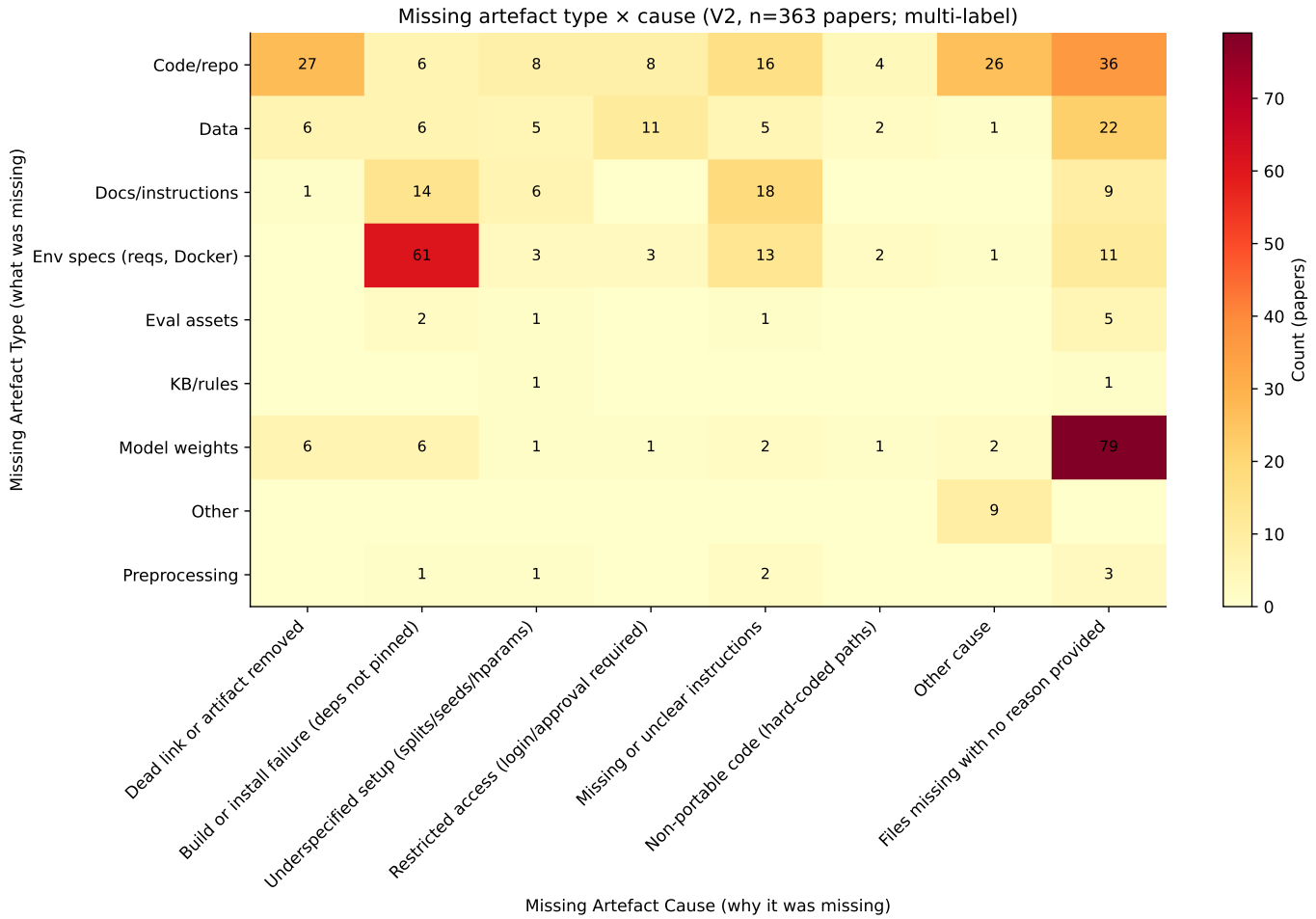

Fig. 5: Rows indicate *what* was missing (e.g., data, model weights, environment specifications, documentation), while columns indicate *why* it was missing (e.g., dead links, restricted access, an under-specified setup, or files absent with no explanation)

or otherwise unusable at audit time. In most instances, the hyperlink supplied in the manuscript was (i) dead or resolving to a private or removed repository, (ii) redirected to a project homepage that described the system but held no source files, (iii) pointed to a repository that never held any source code, or (iv) led to code that was unrelated or insufficient to implement the reported method. These false-positive disclosures illustrate that nominal compliance with "code available" guidelines is not enough, and that persistent, content-verified repositories and explicit version tags are essential if claims of openness are to translate into practical reproducibility.

D. Reproducibility requirements for the domain of Neuro-Symbolic AI as a science

1) Practical implications.: Systematic examination of the AAAI Author Reproducibility Checklist[4], the NeurIPS Paper-Checklist Guidelines[5], the IJCAI Reproducibility Guidelines[6], the NeurIPS report [15], and the AI Magazine survey of reproducibility barriers and drivers [16] shows that the five documents converge on a common baseline of artifacts that an empirical paper must provide: *(i)* runnable source code; *(ii)* the exact datasets or immutable links to them; *(iii)* pre-trained checkpoints; *(iv)* a single command or script that reproduces the reported metrics (an evaluation harness) *(v)* a machine-readable environment file (`environment.yml`, `requirements.txt`, Dockerfile, etc.); and *(vi)* documentation in the form of instructions that integrate items (i)–(v) into one executable workflow. All checklists emphasize persistent identifiers, explicit version locking, and permissive licensing, underscoring that transparency, not methodological novelty, is the primary coin of scientific credibility.

2) Policy levers: Existing venue initiatives fall into two categories. ACM artifact badges [7] include an external review of the repository, whereas the AAAI, NeurIPS, and IJCAI checklists are self-attested by the authors. Each submission to a peer-reviewed conference, journal, and even pre-print venues should undergo an automated repository audit before

[4]https://aaai.org/conference/aaai/aaai-26/reproducibility-checklist/
[5]https://neurips.cc/public/guides/PaperChecklist
[6]https://www.ijcai.org/reproducibility

[7]https://www.acm.org/publications/policies/artifact-review-and-badging-current

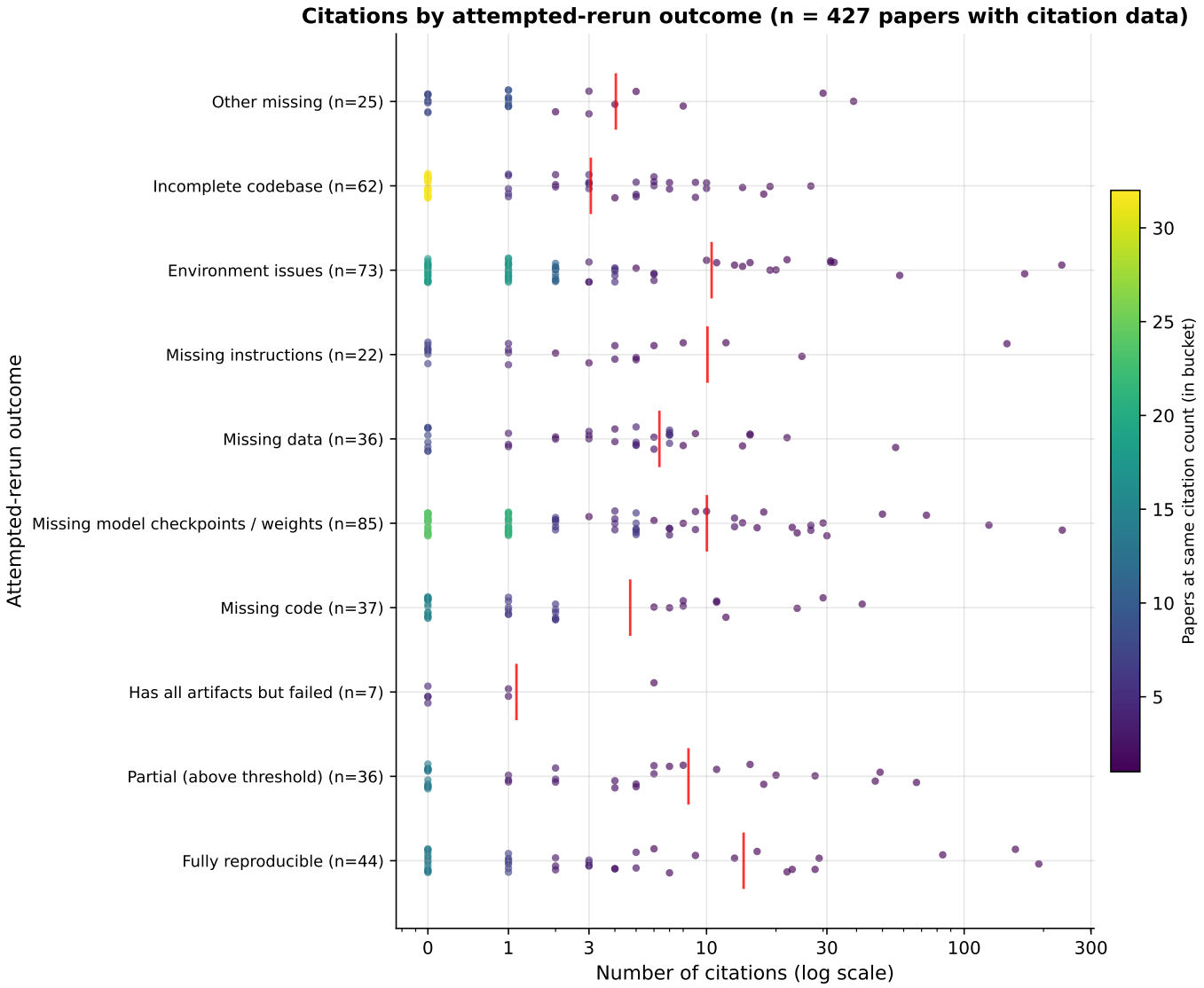

Fig. 6: Citation counts by reproduction outcome/exclusion reason for NSAI papers with available citation data ($n = 450$) shown on a log scale. The red marker indicates the mean citation count for each bucket.

peer review, confirming that all six artifacts are present and executable. The required computation is modest relative to the effort of peer review and the expense of rectifying non-reproducible work after publication.

3) Future directions for NSAI Research: Neuro-Symbolic models integrate multiple components in complicated systems to deliver functional systems, so a single missing artifact can invalidate the entire pipeline. An absent checkpoint severs logic bindings, and undocumented rule sets could compromise evaluation and as such, authors must archive the complete artifact bundle (codebase, datasets, model weights, evaluation scripts, environment files and documentation) in a DOI-minting repository (e.g., Zenodo, HuggingFace-Hub, Figshare, OSF, or Mendeley Data)[8] Computer-science research is only science when its results withstand independent execution. A manuscript and a transient GitHub link do not satisfy this condition. Requiring the six-item artifact bundle, bound by immutable URIs for all publication venues (including pre-publication), is the most immediate, low-cost intervention available to the community. The good news is that the barrier to improvement is low. The included papers from this study (as well as the three case studies, explored further in the appendix) demonstrate that full reproducibility is achievable at any publication venue when authors commit to a complete artifact bundle from the outset. The community can act now by depositing code, data, model weights, environment files, and a single-command evaluation script in a DOI-minting repository such as Zenodo or Hugging Face Hub at submission time, not as an afterthought. Venues can reinforce this by requiring automated artifact checks before peer review begins, a low-cost intervention relative to the effort

[8]These services issue persistent identifiers, maintain long-term storage, and provide metadata suitable for citation.

wasted reproducing or discarding non-reproducible work.

E. Case Studies

1) Case Study 1: Scallop-A Badge-Awarded Benchmark for Reproducibility: The Scallop PLDI 2023 paper [17] earned both the ACM *Artifacts Available* and *Artifacts Evaluated—Reusable* badges, and its release exemplifies full-stack reproducibility. A Zenodo snapshot (DOI 7804200) freezes the exact commit, dataset splits, pretrained weights, and SHA-256 hashes. Docker and Conda manifests plus a Rust nightly lockfile reconstitute the software stack, while a single script (`run_all.sh`) rebuilds, trains, and evaluates the eight-task benchmark. GitHub CI recompiles the core library, runs unit tests, and executes a smoke benchmark on every push, preventing configuration drift. Our audit reproduced all eight tasks on a single A100 GPU with a median absolute deviation ≤ 3 % from the reported metrics and required no manual intervention. Scallop, therefore, is a fantastic demonstration that badge-level artifact curation and automated validation can turn a complex NSAI system into a reliably re-runnable research object.

2) Case Study 2: LogiCity-Conference-Level Reproducibility without Formal Badging: Unlike Scallop, LogiCity [18] earned no external badge, yet its NeurIPS 2024 Datasets and Benchmarks package replicated on the first attempt. The authors froze code, data, and checkpoints in a version-tagged release, shipped Docker and Conda manifests, and wired a one-command launcher that rebuilds the simulator, trains agents, and scores results. Continuous-integration runs the full smoke benchmark on every commit, and SHA-256 checks ensure downloads match the snapshot. On our A100 test node, the Safe-Path-Following and Visual-Action-Prediction tasks reproduced within $\pm$ 2 pp of the reported scores and retained the original baseline ordering. LogiCity therefore demonstrates that rigorous curation and automated checks can deliver badge-level reproducibility at a premier conference—even when no formal badging program exists.

3) Case Study 3: MARS-Preprint-Level Reproducibility through Curated artifacts: The *Mechanism-of-Action Retrieval System* (MARS) [19] appears only as an arXiv preprint posted in March 2025, yet its artifact package met every criterion for a gold standard reproducible package. A version-tagged GitHub repository provides Docker and Conda descriptors, immutable data archives, and pretrained checkpoints, and a single `run.sh` script rebuilds the biomedical-graph pipeline, trains the models, and computes evaluation metrics. Re-execution on one NVIDIA A100 GPU reproduced the MoA-Net benchmark with MRR 0.315 versus the reported 0.318 and Hits@10 0.672 versus 0.685 (absolute deviation ≤ 2 percentage points), preserving the published ranking of baselines. All code, data, and weights are protected by SHA-256 hashes, permitting table verification without retraining. Although the release has not yet undergone an external artifact audit, this example indicates that a fully versioned, one-command artifact stack can deliver robust reproducibility irrespective of publication venue.

4) Cross-Case Comparison and Venue Implications: The three case studies span the principal publication strata in computer science, as a badge-audited archival proceedings (Scallop), a top-tier conference track without external certification (LogiCity), and an open-access arXiv preprint (MARS), yet each enabled low-effort, reliable replication by adhering to the same practical principles. Each artifact package (i) freezes the software stack with Docker or Conda, (ii) archives immutable data and pretrained checkpoints under permanent identifiers, (iii) offers a one-command script that rebuilds and evaluates the full pipeline, (iv) discloses every configuration parameter, and (v) offers a good level of documentation with instructions on environment build requirements. With these ingredients in place, our audit reproduced all reported metrics with an absolute error no greater than three percentage points for all three frameworks. Scallop achieves this through a formal ACM badge review, LogiCity relies on conference-driven community norms and continuous-integration tests, and MARS depends on author-integrity. These results indicate that enforceable artifact standards, not venue prestige, are the decisive factor in achieving reliable reproducibility.

VI. Conclusion

This study provides, to the best of our knowledge, the first longitudinal, large-scale assessment of same-artifact rerun reproducibility within neuro-symbolic AI. Across 1365 self-identified NSAI records identified by our search, 1304 met full-text eligibility and 455 entered the attempted same-artifact rerun audit and 85 were fully or partially reproduced. Failure analysis shows that reproducibility typically collapses when essential artifacts such as full codebases, datasets, pretrained model weights, evaluation harness scripts, environmental files and / or documentation are absent at release. Computer-science research attains scientific legitimacy only when independent investigators can verify its empirical claims. To that end, we recommend that *all* empirical submissions, regardless of venue, be required to pass an automated artifact audit *prior* to peer review. The audit should confirm the availability of six items including (i) full source code, (ii) immutable datasets, (iii) pretrained model checkpoints, (iv) an executable evaluation script, (v) a machine-readable environment descriptor, and (vi) concise documentation sufficient to invoke the pipeline. Persistent DOI-minting repositories such as Zenodo, Hugging Face Hub, Figshare, OSF, or Mendeley Data already provide the necessary infrastructure to host these artifacts at scale and with minimal cost to authors. Adopting this minimal standard would align computer-science publication practice with the norms of experimental science and render future advances in Neuro-Symbolic AI transparent, verifiable, and readily extensible. The community must therefore institutionalize immutable artifact disclosure, requiring and archiving scientific works prior to publication, and thereby ensure that progress is measurable and cumulative by lowering the barrier for full pipeline reproduction of Neuro-Symbolic systems.

References

[1] M. Baker, "1,500 scientists lift the lid on reproducibility," *Nature*, vol. 533, no. 7604, pp. 452–454, May 2016.

[2] R. López-Nicolás, J. A. López-López, M. Rubio-Aparicio, and J. Sánchez-Meca, "A meta-review of transparency and reproducibility-related reporting practices in published meta-analyses on clinical psychological interventions (2000-2020)," *Behav. Res. Methods*, vol. 54, no. 1, pp. 334–349, Feb. 2022.

[3] M. Miłkowski, W. M. Hensel, and M. Hohol, "Replicability or reproducibility? on the replication crisis in computational neuroscience and sharing only relevant detail," *J. Comput. Neurosci.*, vol. 45, no. 3, pp. 163–172, Dec. 2018.

[4] W. M. Hensel, "Double trouble? the communication dimension of the reproducibility crisis in experimental psychology and neuroscience," *Eur. J. Philos. Sci.*, vol. 10, no. 3, Oct. 2020.

[5] M. Rubin, "The costs of HARKing," *Br. J. Philos. Sci.*, vol. 73, no. 2, pp. 535–560, Jun. 2022.

[6] "Research, reuse, repeat," *Nat. Mach. Intell.*, vol. 2, no. 12, pp. 729–729, Dec. 2020.

[7] Open Science Collaboration, "PSYCHOLOGY. estimating the reproducibility of psychological science," *Science*, vol. 349, no. 6251, p. aac4716, Aug. 2015.

[8] J. P. A. Ioannidis, "Why most published research findings are false," *PLoS Med.*, vol. 2, no. 8, p. e124, Aug. 2005.

[9] J. Vanderdonckt and R.-D. Vatavu, "Context is key for reproducibility of empirical studies in human-computer interaction," in *Proceedings of the 3rd ACM Conference on Reproducibility and Replicability*, ser. ACM REP '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 41–50. [Online]. Available: https://doi.org/10.1145/3736731.3746142

[10] M. Ferrari Dacrema, S. Boglio, P. Cremonesi, and D. Jannach, "A troubling analysis of reproducibility and progress in recommender systems research," *ACM Trans. Inf. Syst.*, vol. 39, no. 2, Jan. 2021. [Online]. Available: https://doi.org/10.1145/3434185

[11] A. Belz and C. Thomson, "The 2023 ReproNLP shared task on reproducibility of evaluations in NLP: Overview and results," in *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, A. Belz, M. Popović, E. Reiter, C. Thomson, and J. Sedoc, Eds. Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria, Sep. 2023, pp. 35–48. [Online]. Available: https://aclanthology.org/2023.humeval-1.4/

[12] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep reinforcement learning that matters," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018.

[13] M. Pawlik, T. Hütter, D. Kocher, W. Mann, and N. Augsten, "A link is not enough – reproducibility of data," *Datenbank-Spektrum*, vol. 19, 06 2019.

[14] J. McGowan, M. Sampson, D. M. Salzwedel, E. Cogo, V. Foerster, and C. Lefebvre, "PRESS peer review of electronic search strategies: 2015 guideline statement," *J. Clin. Epidemiol.*, vol. 75, pp. 40–46, Jul. 2016.

[15] J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché Buc, E. Fox, and H. Larochelle, "Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program)," *J. Mach. Learn. Res.*, vol. 22, no. 1, Jan. 2021.

[16] H. Semmelrock, T. Ross-Hellauer, S. Kopeinik, D. Theiler, A. Haberl, S. Thalmann, and D. Kowald, "Reproducibility in machine-learning-based research: Overview, barriers, and drivers," *AI Mag.*, vol. 46, no. 2, Apr. 2025. [Online]. Available: https://doi.org/10.1002/aaai.70002

[17] Z. Li, J. Huang, and M. Naik, "Scallop: A language for neurosymbolic programming," *ACM*, 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/3591280

[18] B. Li, Z. Li, Q. Du, J. Luo, W. Wang, Y. Xie, S. Stepputtis, C. Wang, P. S. Katia, P. K. Ravikumar, A. G. Gray, X. Si, and S. Scherer, "Logicity: Advancing neuro-symbolic ai with abstract urban simulation," *Neurips*, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/8196be81e68289d7a9ece21ed7f5750a-Paper-Datasets_and_Benchmarks_Track.pdf

[19] L. N. DeLong, Y. Gadiya, P. Galdi, and J. D. Fleuriot, "Mars: A neurosymbolic approach for interpretable drug discovery," *arXiv*, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2410.05289