Lessons Learned from Five Years of Artifact Evaluations at EuroSys

Daniele Cono D'Elia
Sapienza University of Rome
Rome, Italy
delia@diag.uniroma1.it

Thaleia Dimitra Doudali
IMDEA Software Institute
Madrid, Spain
thaleia.doudali@imdea.org

Cristiano Giuffrida
VU Amsterdam
Amsterdam, Netherlands
c.giuffrida@vu.nl

Miguel Matos
IST Lisbon & INESC-ID
Lisbon, Portugal
miguel.marques.matos@tecnico.ulisboa.pt

Mathias Payer
EPFL
Lausanne, Switzerland
mathias.payer@nebelwelt.net

Solal Pirelli
Independent Researcher
Lausanne, Switzerland
solal.pirelli@gmail.com

Georgios Portokalidis
IMDEA Software Institute
Madrid, Spain
georgios.portokalidis@imdea.org

Valerio Schiavoni
University of Neuchâtel
Neuchâtel, Switzerland
valerio.schiavoni@unine.ch

Salvatore Signorello
NOVA University Lisbon
Lisbon, Portugal
s.signorello@fct.unl.pt

Anjo Vahldiek-Oberwagner
Intel Labs
Berlin, Germany
anjovahldiek@gmail.com

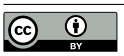
Abstract

Artifact Evaluation ("AE") is now an accepted practice in the systems community. However, AE processes are inconsistent across venues and even across different editions of the same venue. AE processes regularly encounter the same problems across venues and years. Based on our collective experience in chairing various and heterogeneous AE committees for five consecutive editions of EuroSys, a large systems conference, we present the challenges we believe most pressing. We propose concrete steps to address these challenges in future AEs, serving as guidelines for future chairs and AE committees.

Keywords

Artifact Evaluation, Reproducibility, Conference-scale artifact evaluation experiences, Conference-scale artifact evaluation practices

ACM Reference Format:
Daniele Cono D'Elia, Thaleia Dimitra Doudali, Cristiano Giuffrida, Miguel Matos, Mathias Payer, Solal Pirelli, Georgios Portokalidis, Valerio Schiavoni, Salvatore Signorello, and Anjo Vahldiek-Oberwagner. 2025. Lessons Learned from Five Years of Artifact Evaluations at EuroSys. In *ACM Conference on Reproducibility and Replicability (ACM REP '25), July 29–31, 2025, Vancouver, BC, Canada.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3736731.3746152



ACM REP '25, Vancouver, BC, Canada

ACM ISBN 979-8-4007-1958-5/25/07
https://doi.org/10.1145/3736731.3746152

1 Introduction

Artifact Evaluation ("AE") processes help make science trustworthy. Independent verification of data is a cornerstone of science, thus a lack of reproducibility in scientific research hampers scientific progress. Any scientific community that performs empirical experiments but does not take trustworthiness seriously undermines public trust in all scientific research. Reproducibility must be a first-class citizen in science, not an afterthought. One common overarching goal of these considerations is scaling up AE practices to increase their long-term impact. This mindset sparked the creation of various initiatives in CS research, such as the ACM Emerging Interest Group for Reproducibility and Replicability [11], the SIGSOFT Artifact Evaluation Working Group [24], the ACM SIGMOD ARI [13], and various other AE processes [23, 25].

AE is the conceptually simple process of checking whether the artifacts published alongside a paper, such as code and data, correspond to what the paper describes. In practice, this leads to many questions and challenges. The very first AE process we are aware of, at ESEC/FSE 2011 [2], awarded a badge to papers that passed an artifact evaluation executed by a dedicated committee. Since then, other venues have adopted various forms of AE, using multiple badges with varying definitions and requirements. However, the exact meaning of each badge and the corresponding granting criteria are ambiguous across CS research fields, as we describe in §5.

Even the definition of "reproducibility" lacks consensus. We identify several challenges in any CS-related AE process, including the difficulty of evaluating non-code artifacts, the subjectivity of deciding whether a set of specific results are sufficient evidence for a claim, and the need for specialized hardware for systems artifacts. The process of evaluating artifacts is more complex in practice than

it appears in theory, which also makes it more interesting to study and improve. We describe AE and its challenges further in §2.

We report on our collective experiences chairing the AE process for EuroSys over the last 5 years. EuroSys is a well-known venue for computer systems research, an area that includes specific challenges for artifact evaluation. The AE process has stayed relatively similar after its introduction in the venue. It consists of three phases: (1) artifact submission; (2) an early "kick the tires" phase to identify problems early; and (3) the full evaluation. Evaluators are mostly graduate students and early-career researchers, who are assigned a few artifacts to evaluate and must reach a consensus with their fellow reviewers for each artifact. In addition to standard AE features, EuroSys has introduced *checklists* provided to both authors and evaluators to define expectations, and *artifact appendices* required from authors to describe their artifact. After an initial excitement in artifact submission followed by a slump, successful evaluation of artifacts in EuroSys has stabilized with around 50% of accepted papers being considered reproducible. We present the AE process of EuroSys in §3, including details on submissions and evaluations.

In our experience, certain challenges within the process are recurring. Each AE instance, as well as the conference itself, has limited decision-making power and resources to address these issues. We believe that tackling these obstacles requires broader community efforts, which benefits from targeted discussion. Review deadlines are tight, and the various constraints dictating them cannot be easily changed. There is little interaction between the program and artifact committees, which means the major claims of the paper as seen by the paper reviewers may not match what the artifact evaluators expect to reproduce. Some artifacts may require specific hardware or may even *be* new hardware, which leads to headaches in trying to provide remote access or even physical access through shipping. The AE committee, including chairs, is appointed for both rounds of submission of a single edition with little handover between successive committees, which makes it hard to build institutional knowledge. Ensuring artifacts remain available in the long term, including any software dependencies, is nontrivial, despite encouraging efforts such as the Software Heritage initiative [3]. The badge system introduced by the ACM is more descriptive than prescriptive, which leads to open interpretations and confusion when authors and evaluators disagree on what criteria an artifact must meet. To address these recurring challenges, we provide both short- and long-term proposals. Our short-term proposals should be usable by an individual AE committee wishing to improve the process. Our long-term proposals require collaboration with the program committee and possibly the steering committees. We present the challenges in further detail and our proposals in §4.

In summary, we provide a qualitative and quantitative look at the last 5 years of EuroSys artifact evaluation. We present recurring challenges and propose concrete steps to address them. We hope to foster broader debate and inspire the next 5 years of AE, for computer science research in general and for system researchers at EuroSys in particular.

2 Background

We begin by presenting common AE terms used to describe artifacts and evaluation results, and the reasons that make the evaluation of CS artifacts more than reading and executing code.

2.1 State of the world

Science is built one result at a time, each relying on previous results. This model fundamentally relies on the trustworthiness of published results. However, there is a broad consensus across scientific fields that this assumption is not always justified [29]. Peer review is traditionally not intended to check whether the data presented in a paper is real, only whether the conclusions drawn by the paper make sense. This leaves open the possibility of human error [41] and even scientific fraud [47].

In computer science, the trustworthiness of research results was brought to mainstream attention in the 2010s. Collberg and Proebsting [34] attempted to find and use the software artifacts accompanying papers in computer science conferences with decidedly mixed results, often being unable to find, compile, or execute the software given reasonable time and effort. Few computer science papers are dedicated to evaluating previous results, in part because scientific venues usually do not explicitly encourage such work. One such paper [30] found that while the original results held in part, they were also overly optimistic on more realistic inputs. Another reason why evaluating computer science results is hard is that such results can take forms other than code, such as hardware [35].

The first "artifact evaluation" in computer science was run as part of ESEC/FSE 2011 [39], awarding an "artifact evaluated" badge to papers that passed evaluation. ACM SIGMOD started a similar process in 2008 [1, 31] under the name "experimental repeatability requirements." More recent efforts have awarded multiple badges for different criteria [28], and required author input such as an "artifact appendix" at the end of published papers or an "artifact description" along with submissions [40].

In the broader scientific community, *registered reports* are seen as a way forward [32]. Instead of submitting results and their associated data for review once research is done, authors submit the methods and analyses they plan to use. Venues then commit to accepting whatever results come out of a report whose registration they accept, so long as the plan is followed. The model of registered reports is applicable to computer science, as has been suggested already [33], and successfully used in the Fuzzing workshop [21].

2.2 Goals and vocabulary

The goal of artifact evaluation is to make science more trustworthy by ensuring published results are accurate. Scientists should be able to build upon others' results with confidence.

Concretely, research results should be *reproducible* and *reusable*. The exact definitions of these terms, and even which terms to use, are unfortunately not subject to a consensus. Different authors and organizations have proposed different terminologies, notably Feitelson [36], the National Information Standards Organization (NISO) [43], and the ACM [28]. ACM inverted "replicability" and "reproducibility" in the second edition of its badging standard, current as of June 2025, exemplifying the lack of consensus.

In the remainder of this paper, we refer to the subset of the ACM badges used in computer systems research. These are three levels of badges related to artifact review: Artifacts Available, Artifacts Evaluated, and Results Validated. The latter two have two sub-levels, but computer systems venues only use one of each:

- **Artifacts Available**: Artifacts are permanently available.

- **Artifacts Evaluated – Functional**: Artifacts are documented, consistent, complete, exercisable, and include appropriate evidence of verification and validation.
- **Results Validated – Reproduced**: The main results of the paper have been obtained in a subsequent study by a person or team other than the authors, using, in part, artifacts provided by the authors.

These badges are independent, meaning that any combination of one, two, or all three badges can be applied to a given paper. The descriptions of the badges do not include specific details about the review process established by a particular venue.

2.3 Complexity in Computer Science

Evaluating CS artifacts is not as simple as running through an objective checklist, though we argue in this paper that it is also not as complex as some may think. There are four main obstacles.

First, because not all CS artifacts are code, some level of trust in the authors is often required. Evaluating a piece of hardware given remote access fundamentally cannot yield the same level of trust as evaluating code that can be downloaded and used in isolation. Some artifacts are unreproducible by design, such as survey results, though raw data should be provided in full so it can be checked for errors, and results can be replicated by running a new survey.

Second, even code artifacts often have dependencies beyond what the average scientist has access to. Distributed systems need to be distributed across many nodes to be tested properly. Some software requires lots of computing power, or specific kinds of computing resources such as hardware accelerators. While emulation can solve some problems, such as running a cluster of virtual machines on a single machine, it carries the inherent risk of not evaluating what the authors have actually done.

Third, preserving entire artifacts is nontrivial. Beyond where to permanently store code, software dependencies are often implicit and not permanently archived. Guilloteau et al. [37] found widespread longevity concerns in computer systems conferences.

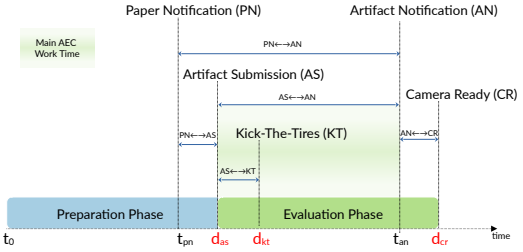
Finally, some subjectivity is unavoidable. "Zero tolerance" cannot work in practice: even two runs of the same software on the same hardware yield slightly different results due to environmental factors like room temperature affecting CPU clock speed. Artifact evaluation requires subjective decisions for how close results must be, and for which results must be reproduced for a claim to hold.

2.4 Problem Statement

Overall, artifact evaluation represents progress towards reusability and reproducibility, but is currently stagnating. While there is consensus on the goal, current implementations are too different to naturally converge. In this paper, we expose current challenges and propose concrete steps to meet them, based on our experience running artifact evaluation processes for five editions of EuroSys.

3 Artifact Evaluation at EuroSys

The artifact evaluation process was first introduced at the 16th edition of EuroSys, inspired by similar initiatives at leading systems conferences like OSDI and SOSP [4]. As part of this process, authors of accepted or conditionally accepted papers were given the opportunity to submit artifacts—such as software, data, and documentation—that support the research presented in their papers for evaluation. The goal of the conference organization during that edition was to encourage reproducibility and facilitate the reuse of the work presented. Since then, the artifact evaluation process has become an optional, regular feature of the conference. This section outlines the organization of the overall artifact evaluation process, its main phases, and key features. Additionally, it documents the progress of this process over time by providing important statistics that help analyze its current impact and future evolution.



(a) Timeline and main deadlines (d_i) of the AE process.

Year	PN-AS	AS-KT	AS-AN	PN-AN	AN-CR
2021	4	-	60	64	11
2021 Rev.	4	-	25	29	4
2022	13	10	47	60	1
2023 Spring	14	14	44	58	2
2023 Fall	14	14	42	56	2
2024 Spring	18	5	43	61	3
2024 Fall	12	3	18	30	2
2025 Spring	13	7	19	32	4
2025 Fall	13	7	20	33	4

(b) Duration in days of the AE phases starting from paper notification (PN), '-' means phase not defined in that edition.

Figure 1: Figure 1a illustrates the timeline of the AE process, while Table 1b provides information about the duration of some of its phases over the various editions.

3.1 Description

Figure 1a illustrates the timeline adopted for the AE process at EuroSys, along with its various phases, whose exact durations are reported in Table 1b. During the main review phase of the conference, a Call for Artifacts is posted on the conference website. Authors of accepted papers have the opportunity to submit their artifacts for evaluation shortly after they receive notification of their paper's acceptance. They can apply for one to three different ACM badges [28]: Artifacts Available, Artifacts Functional, and Results Reproduced. Each submitted artifact is evaluated by the conference's AEC between the paper notification date and the camera-ready deadline. At the conclusion of this process, papers whose artifacts have passed evaluation include an appendix in their camera-ready version to describe the artifacts submitted, along with one or more badges displayed on the first page of the paper to acknowledge the results of the artifact evaluation. The process is chaired by a variable number of chairs, between two and four, who collect artifact submissions from authors and reviews and scores from the evaluators through ad-hoc instances of the HotCRP submission system.

3.1.1 Preparation Phase. There exist three main preparatory steps to setting up the artifact evaluation process, to be completed prior to the artifact submission deadline.

Forming the Artifact Evaluation Committee. Membership in the Artifact Evaluation Committee (AEC) is particularly appealing for early-career researchers, such as senior graduate students and postdoctoral researchers, who are actively working in computer science areas relevant to EuroSys. Traditionally, AEC members are recruited through a self-nomination process. In this process, the AE chairs invite nominations for potential candidates and request information about the nominee's research domain (academic or industrial), role, areas of interest and expertise, as well as their previous experience with Artifact Evaluation (AE) processes. The target size of the committee is determined by the anticipated number of accepted papers, which is usually estimated in consultation with the Technical Program Committee (TPC) chairs of the conference. This target size aims to ensure that each artifact receives multiple independent reviews—typically between three and four—while keeping the overall workload for each AEC member reasonable, generally limited to one to three artifact reviews during the evaluation phase.

Instructing Authors and Evaluators. Artifact evaluation is a comparably younger and less established process compared to traditional peer review of scientific papers. As a result, both authors and reviewers may be less familiar with the common steps, tools, and objectives of this process. To assist both parties with artifact evaluation, EuroSys provides and maintains a set of supplementary materials through an ad-hoc website [25]. While the terminology has evolved slightly across different editions, the supplementary materials cover three main topics: instructions for packaging artifacts, best practices for preparing quality artifacts, and a guide for evaluators. The first two resources offer essential information for authors to prepare their artifacts for submission and provide guidelines to help present them in the most effective way for evaluation and future reuse. The third resource is designed to guide evaluators on their objectives, setup, and tools for the evaluation process.

Preparing Infrastructure for Evaluation. The AEC can evaluate artifacts using various setups, depending on the software and hardware requirements of the specific artifact being assessed. AEC members have the option to use their own commodity hardware or any specialized hardware that is available to them through their affiliations. Additionally, for artifacts that necessitate more hardware than they can provide for evaluation, they can utilize public research and commercial cloud services. The AE chairs are responsible for requesting and allocating quotas on these cloud platforms for the evaluators if the evaluation of an artifact requires it. Lastly, the AEC can also use the machines of the artifact authors, accessed through secure and anonymized remote connections, for artifacts that cannot be run elsewhere due to hardware dependencies or other restrictions.

3.1.2 Review Phase. This phase is similar to a paper review process. Evaluators are assigned artifacts for independent evaluation, and are expected to submit their scores on the requested badges and reviews by a given deadline. The review phase is designed to be collaborative rather than adversarial. Evaluators can discuss their evaluations of the artifacts with the authors anonymously through the submission system, as well as their reviews among themselves and with the chairs.

Artifact Submission. After receiving notification, authors of accepted papers who wish to apply for artifact evaluation should express their interest through the submission platform. This involves submitting their paper as a starting point and indicating the badges for which the artifact is applying. The final submission must include a packaged artifact, a URL to an external public platform where the artifact is hosted, a short description of the artifact and any special information and instructions for evaluation, the version of the accepted paper and the artifact appendix. The artifact appendix is a self-contained document that serves as a roadmap for evaluators. Authors must follow a template provided by the AE process when writing their appendix. This appendix must include a description of the hardware, software, and configuration requirements, as well as a list of the main claims in the paper that the artifact supports. It should also explain how to reproduce the main experiments described in the associated paper and compare the reproduced results with those reported in the paper.

Kick-The-Tires. The purpose of this phase is to identify at an early stage various problems that may prevent a complete and thorough evaluation of an artifact too close to the evaluation deadline. Common issues to look for include evaluators who may not have the necessary hardware or software environments for the evaluation, difficulties with remote access to the authors' testbed, poorly organized or incomplete artifact appendices, and a lack of essential instructions for replicating the experiments or components included in the packaged artifact. During this phase, evaluators are responsible for ensuring that all conditions for an artifact evaluation are met. They should discuss and resolve any issues with the authors and fellow evaluators as needed. If any issues remain unresolved, the evaluators must inform the AE chairs. The chairs may reassign reviews or request missing materials from authors based on the outcomes of this initial phase.

Evaluation. For each artifact they are assigned to, evaluators are asked to write a review explaining which badges they think the artifact should or should not receive, justifying their binary score for each badge requested by the authors. For artifacts applying for the "Results Reproduced" badge, they are also asked to check that the experiments and results presented in the appendix for the corresponding claims are validated by their evaluation. Unlike the typical peer review process for scientific papers, the reviews and scores from artifact evaluations are sent to authors as soon as they are submitted by the evaluators. Thus, authors are promptly informed of specific issues with their artifacts throughout the evaluation process, enabling them to address important concerns regarding the requested badges before the artifact review deadline. Authors can discuss their artifacts with evaluators and make updates to both the artifact and the appendix in order to potentially change the evaluators' opinions by the deadline.

After the reviews are completed, the evaluators may need to come to a consensus regarding which badges to award. This discussion process may result in additional specific requests to authors by the AEC to confirm certain badges for their artifacts. In their reviews, evaluators can also flag artifacts that they consider to be of exceptional quality.

Results. At the end of the evaluation phase, authors are notified of the final results of the artifact evaluation. The decision may award zero or more badges to the artifact, depending on what was requested. This final decision could be contingent upon pending changes to the artifact and/or appendix requested by the AEC during the evaluation phase. By the camera-ready deadline, the AEC requires authors to create and provide a Digital Object Identifier (DOI) for their artifact. Typically, authors store their artifacts in open science persistent repositories, such as Zenodo [27], Figshare [20], or Dryad [18]. These platforms offer easy integration with popular open-source software development platforms (like GitHub) and allow authors to generate a DOI for their artifact with just a few simple steps. Artifacts themselves are not stored directly in the ACM Digital Library; instead, their DOIs and badge metadata are linked to the corresponding paper entries in the digital library. Badges assigned to an artifact are also displayed on the first page of the PDF version of the paper available from the digital library.

Between the results announcement and the conference, a subcommittee of the AEC further analyses one or a few artifacts among those recommended as outstanding during the evaluation phase. This process is aimed at selecting a few artifacts for the *Gilles Muller Best Artifact Award*, which is announced live at the conference.

3.1.3 Distinctive Features. This section focuses on specific phases and mechanisms—referred to as features—that have been consistently or occasionally integrated into the AE process at EuroSys, which we believe have enhanced the overall organization of the process and contributed to achieving its original objectives. We refer to them as distinctive because they are not strictly present in AE installations at similar conferences.

Features that assist the process. The possible misinterpretation of the official ACM badge definitions [28] by evaluators and authors can lead to unintended review procedures and outcomes. In response to this problem, the AE organization at EuroSys has provided both parties with ***Badge Checklists*** that translate the ACM badge definitions into a set of concrete criteria that an artifact should tick for being awarded the corresponding badges. According to a survey conducted with the AE committee at EuroSys 2022 [5], the badge checklists made evaluation easier.

The evaluation of artifacts is a relatively new process compared to the traditional review of scientific papers, and as a result, the roles and responsibilities of artifact evaluators are often not well understood. Additionally, if authors do not properly prepare their artifacts for evaluation, it can complicate the evaluators' task. Therefore, it is crucial to provide clear and explicit guidance to both the committee and the authors and to closely monitor their initial interactions. The ***Kick-the-Tires*** phase serves as an initial effort aimed at addressing these issues, offering the dual benefit of identifying potential problems with either the artifact or the evaluators assigned to it at an early stage. Ultimately, this approach reduces the likelihood of incomplete or missing evaluations.

Features that help assess reproducibility. To effectively assess reproducibility, it is crucial to connect the claims made in a paper to the corresponding artifact. This connection allows evaluators to reproduce the results that are essential for supporting the paper's main claims. The ***Artifact Appendix*** serves as a document where authors must clearly outline and connect the primary results and claims of their paper. It explicitly lists elements such as results, plots, and tables from the paper, cross-referencing them with the experiments that need to be reproduced using the artifact. This document is particularly important when the claims may be challenging for evaluators to deduce from the paper or may differ from the expectations set by the document itself.

According to the 2022 EuroSys chairs' report [5], the use of *academic clouds*, such as Chameleon and CloudLab, was useful for evaluating the reproducibility of several submitted artifacts. In fact, some authors provided specific instructions for running their experiments on one or both of these clouds. Overall, the use of academic clouds facilitates the reproducibility of experiments. Additionally, the AEC did not incur costs associated with commercial cloud services for this purpose.

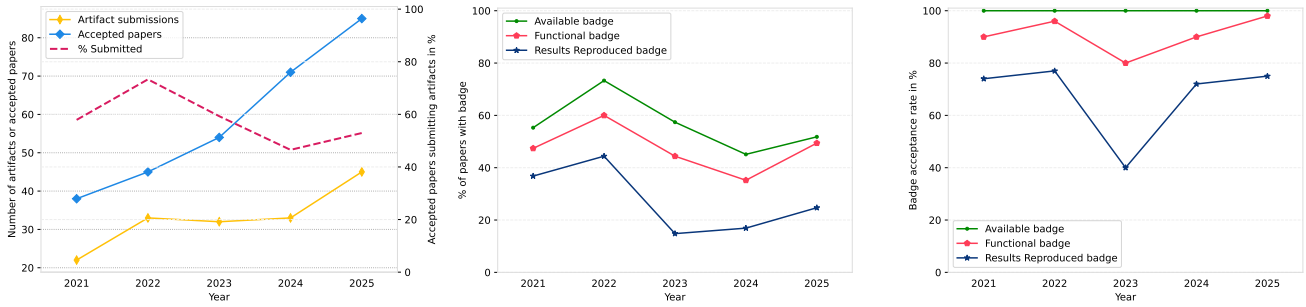
3.2 Statistics from 2021-2025 EuroSys AEs

In this section we present statistics from the past 5 installments of AE at EuroSys (2021–2025). In total, 161 artifacts available badges, 136 artifact functional badges, and 75 results reproduced badges were awarded over the past 5 years. Due to space limitations, we provide additional data in Appendix A. All presented data should be taken cautiously and further research is needed to expand this analysis, since 5 years of data collection is only a start and does not provide a large data set.

Data Sources. We collect data from sysartifacts.github.io [25] and the proceedings front matter [4, 6, 7, 10, 19]. In addition, we built tooling to scrape sysartifacts.github.io [14]. Since EuroSys 2022, artifact badging results and storage locations of each artifact repository or DOI are reported in YAML format in the sources of sysartifacts.github.io. The tool offers the ability to extract and parse the YAML data and offers additional scripts to validate the availability of storage locations like Zenodo, FigShare and Github, and present statistics.

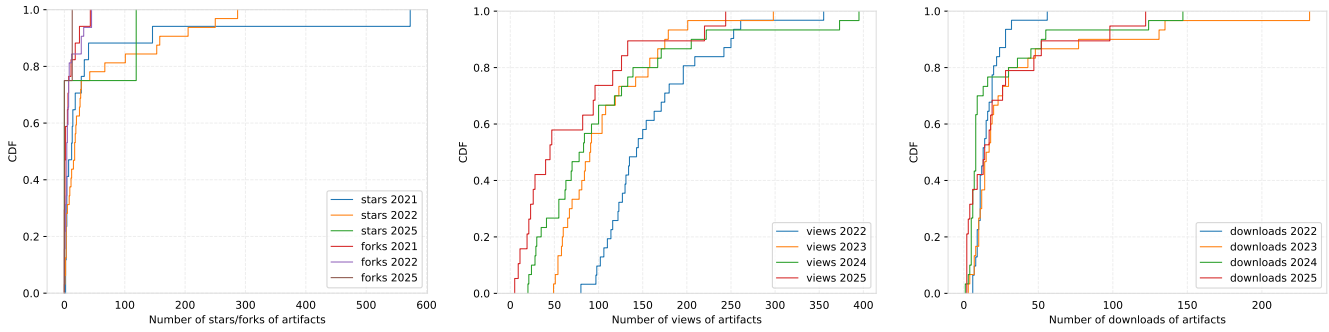
Number of submissions. Figure 2a shows the evolution of number of papers accepted to the program and participating in the AE process, with a clear increasing trend in both. More specifically, it is a 2× increase in the number of submissions to the artifact evaluation process, from 22 in 2021 to 45 submissions in 2025 over both Spring and Fall submission cycles. However, this growth has not led to a significant rise in the proportion of submissions relative to the total number of papers accepted into the EuroSys program. Figure 2a also shows that, on average, **58% of accepted papers** have participated in the artifact evaluation process each year, showing no clear trend. Also, there was no visible distinction in the number of submissions between the Fall and Spring cycles. In 2023, submissions were evenly distributed across both cycles. However, in 2024, the Spring cycle had twice as many submissions as the Fall cycle, while in 2025, the pattern reversed, with Fall submissions being twice as numerous as those in Spring.

Badges awarded. Figure 2b shows the percentage of papers accepted into the EuroSys program that were awarded the 3 different types of badges. As shown before, on average, 58% of the accepted papers participated in AE and indeed received the "artifacts available" badge. However, this percentage reduces by 10% for the "artifacts functional" badge, and even more for the "results reproduced"



(a) (b) (c)

Figure 2: Number of accepted papers and artifact submissions and the percentage of accepted papers that submitted an artifact in (a), percentage of accepted papers being awarded a badge in (b), acceptance rates of badges in (c) over the five years.



(a) (b) (c)

Figure 3: Cumulative distribution functions (CDFs) of repository stars and forks on Github in (a), of artifact views on Zenodo & FigShare in (b), of artifact downloads on Zenodo & FigShare in (c) over the five years.

badge which only around 28% of the accepted papers receive. Next, we look into these trends in more detail.

Figure 2c shows the percentage of papers that were awarded a badge they applied for. First, almost all papers involved in artifact evaluation applied for the "artifacts available" badge and all were successfully got it. We validated that these artifacts exist and are available to this date, as discussed later. Second, the "artifacts functional" badge was requested by an average of 82% of the evaluated papers over the years. 2025 reports the highest rate, with 93.3% of the papers applying for the badge. Of these, 90% were awarded the badge. 2025 also reports the highest acceptance rate, 98%.

Regarding the "results reproduced" badge, throughout the years, on average, only 47% of the papers applied for this badge, which represents only 28% of the papers accepted into the program. There is also a great difference in the distribution of the values over the years. In 2021-2022 62% of the papers applied for the badge. In 2023 there was a significant dip, down to 25% of the papers. This was due to various reasons, such as overly long experiments, limited access to specialized hardware needed for the evaluation and artifacts poorly packaged. In 2024 and 2025, the percentage did increase, reaching 46.7% in 2025. Out of the papers applying for the "results reproduced" badge, if we exclude 2023 as an outlier, we observe that 75% of the papers were awarded the badge.

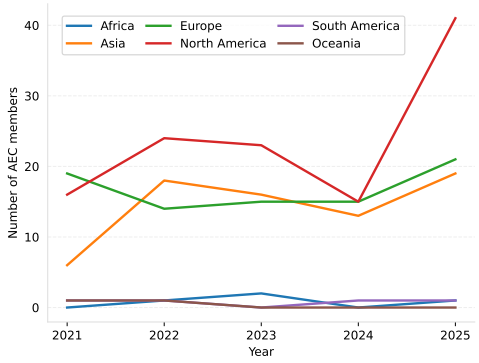
Artifacts availability. We rely on our tooling to validate that repositories and DOIs are still available. We assume that a storage location is still available, if it returns an HTTP 200 response. The collected data in available in Table 3 of the Appendix A. In 2021 and 2022 artifacts with available badges needed to provide a code repository and could voluntarily provide a DOI (e.g., Zenodo). Only 1 repository from 2021 is no longer available. From 2022, 1 repository is no longer accessible and 2 artifacts do not provide a DOI, but all artifacts are either available via the repository or the DOI. In the following years, repositories are no longer collected, and only DOIs are recorded in sysartifacts.github.io. Besides 1 DOI, all DOIs specified on sysartifacts.github.io are still available. One artifact was removed upon request by the uploading user and is hence no longer available even though it was stored in Zenodo. The DOI now links to a tombstone. Overall, artifacts that have received the available badge, indeed remain available to this day.

Visibility and downloads. We use the tooling to further examine the use of the artifact based on usage metrics of the corresponding storage service. Almost all artifacts are stored either on Github, Zenodo or Figshare (72, 107, and 3 respectively over all 5 years). At the time of collection, artifacts from 2025 are only days old and, hence, the numbers are close to zero and likely reflect mostly traffic during AE. Figures 3a, 3b, 3c, show the distribution of stars, forks, views, and downloads for the years 2021 to 2025, whenever the

respective repository entries were available for those years. As expected, artifacts from older iterations have received more traffic because they have been available for a longer time. What is interesting to observe is that certain years have had statistics with long tails, meaning the existence of few artifacts that received lots more attention (stars/forks, views and downloads) than others. It would be interesting to verify if those artifacts have received community contributions or if they are still actively used and maintained. There is room for much more exploration of such relevant aspects.

Evaluation Committees. The size of the artifact evaluation committee has been quite substantial throughout the year, with 64 members on average, to accommodate the load. Every artifact received 3 reviews, on average. Every year, a call for self-nominations for reviewers was issued, from which the chairs selected the final committee members. In all years, the majority of the self-nominated reviewers were selected. In 2025, we got an astonishing amount of self-nominations, resulting in a committee size of 98 members. Due to the much larger number of reviewers vs. artifact submissions, half of the reviewers served in the Spring cycle, and the other half in the Fall cycle of this year. In the previous years, almost all reviewers served in both cycles, with few exceptions due to availability or poor review performance. Figure 4a shows the number of committee members affiliated with institutions across the different continents. There is a strong participation of US-based institutions, that increased significantly in 2025, which aligns with the increase in the overall size of the committee. Finally, over the years, there has been notable representation from countries in Europe and Asia.

In Table 4b, we highlight the return of AEC members in subsequent years. With the AEC sizes growing over time, the number of returning AEC members has also increased. Some of the early members are continuing to serve since as early as 2022. Overall, the number of returning members remains relatively small, suggesting that most AEC members are new.



(a) Geographical distribution of AEC members.

Year	2021	2022	2023	2024	2025
2021	-	4	1	0	0
2022	4	-	6	2	3
2023	1	6	-	5	3
2024	0	2	5	-	11
2025	0	3	3	11	-

(b) AEC members retention.

Figure 4: AEC distribution and retention over the years.

4 Lessons Learned

Artifact evaluation has become an integral part of major conferences and is expected to remain standard practice. In five years of artifact evaluation at EuroSys, we have gained insights and identified challenges that we believe will continue to show in future editions and, most likely, also in related efforts in systems research and neighboring communities such as security. In this section, we distill key open challenges and elaborate on potential directions for addressing them in the short and longer term, hoping that this will improve the artifact evaluation process.

4.1 Open Challenges

The open challenges we discuss below are both structural and procedural in nature. For some, we will later provide our suggestions on how artifact evaluation efforts can be improved to tackle them more effectively. For others, we call on the community and hope, in the meantime, to stimulate discourse by sharing the insights we gained about their nature and manifestations.

Challenge 1: Tight review deadlines. The review deadlines for the artifact evaluation remain extremely tight and must traditionally fit between the paper notification and the camera-ready deadline. This means that the process *starts after paper notification* but *must complete before the camera ready deadline* that finalizes the papers. This dependency on the earlier deadlines makes the timing of the artifact evaluation challenging and requires careful control from the artifact evaluation chairs to ensure smooth progress with little to no slack. The artifact evaluation process consists of two main phases: the preparation phase where authors prepare the artifact and the evaluation phase for the actual artifact evaluation (which includes the kick-the-tires phase with extensive interaction between the reviewers and the authors, Figure 1a). Extending the second phase is challenging because the camera ready deadline is constrained by publishing requirements, while moving the paper notification earlier seems even less feasible due to the increased program committee workload from rising submission numbers. For the first phase, authors decide at their discretion whether to allocate any effort in the time frame between paper submission and notification to artifact preparation, and at submission time they are only optionally required to express an interest in artifact evaluation. Artifact evaluation committees are forced set their pace around the second phase, whereas shortening the first phase has not yet been explored.

Challenge 2: Lack of interaction between paper and artifact review. One key challenge we observed during the artifact evaluation is that, so far, artifact evaluation remains disjoint from paper review. During the paper review, the scientific merits are assessed and reviewers have a clear understanding of how they would expect the artifact to back them up. These expectations are not codified, and the artifact evaluation process revolves around a list of major claims proposed by the authors to the artifact reviewers. In most cases, the current badges are assigned after evaluating only some of the experiments of the paper. While this may overlap with some of the paper's claims, there is no guarantee of completeness. Even worse, the badge description generally demands a validation of the major claims but does not specify that all claims have to be

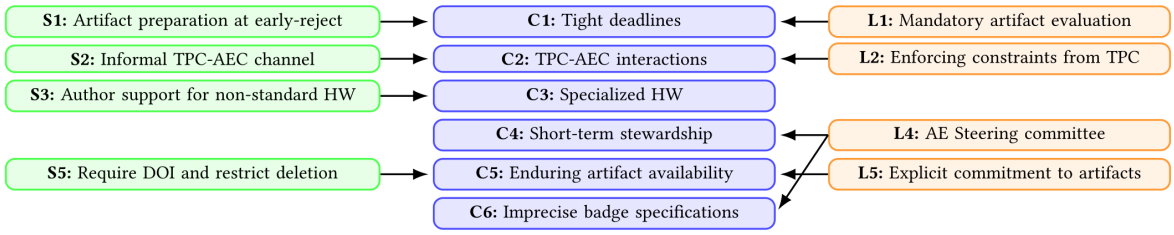


Figure 5: Different challenges (C1-C6) and our short term (S1-3, S5) and long term (L1-2, L4-5) proposals to address them.

validated. Additionally, in multiple occasions, we witnessed committee members asking the authors why parts that they believed to be important in the paper were left out of the artifact evaluation materials. Judging the legitimacy or significance of such concerns is challenging, and chairs can only mediate these discussions and avoid friction, both due to their role and the lack of deep technical insights about the paper. We find that all these issues ultimately root in artifact evaluation operating disjointedly from the official paper review process.

Challenge 3: Specialized hardware or infrastructure. Reproduction becomes harder when an artifact requires extremely recent, highly performant, or niche hardware. Many paper evaluations are nowadays conducted on such specialized hardware: this leads to the baffling outcome that, for several artifacts, reproduction may be feasible only on the authors' machines. Similar circumstances arise when artifacts depend on complex setups or custom test-beds. This introduces several challenges. First, it complicates the evaluation process, as reviewers need remote access, which can introduce delays. This issue is particularly pronounced with the growing prevalence of ML/AI research, where many artifacts depend on high-performance GPU-based systems. Access to such hardware is highly competitive, both within research labs, where resources are often shared, and in cloud environments, as for many research groups it may not be financially sustainable to rent equipment to have their artifact evaluated. This challenge is likely to become even more pressing in the future. Second, in some cases, granting remote access to authors' facilities is simply not an option due to institutional IT policies that prohibit external access. Finally, requiring reviewers to use systems controlled by the authors raises concerns about the objectivity and validity of the evaluation. On a related note, AE processes should protect the identity of reviewers during these interactions, and in some cases (for example, when export control regulations restrict the pool of countries for eligible AEC reviewers) additional measures should be taken.

Challenge 4: Short-term stewardship. In many conferences, including EuroSys, artifact evaluation chairs are appointed for a single edition. Newly appointed chairs are typically encouraged to reach out to their predecessors for insights on running the process effectively and navigating key challenges. However, these interactions often focus on broad guidance, leaving out specific nuances and recurring circumstances that each edition's chairs must handle using their best judgment. As a result, important knowledge is lost in transition. Unlike artifact submission and reviewing guidelines, which are refined with each edition, no written record is left to inform future artifact evaluation organizers. Some conferences, such as ISOC NDSS and USENIX Security, address this issue by appointing chairs for two years with overlapping terms. In this model, the newly appointed chair spends one year learning from the senior chair while serving in a junior capacity before stepping into the senior role. This approach mitigates knowledge drain and promotes continuity in committee practices. However, it also increases the workload for chairs, limiting the pool of candidates willing to serve in a role that, in larger conferences, carries a workload second only to that of the general and program chairs.

Challenge 5: Enduring artifact availability. The Artifact Available badge requires artifacts to be permanently accessible. In our experience, most authors choose GitHub—which does not guarantee permanent availability—or services like Zenodo for hosting. Our validation of storage services for EuroSys artifacts showed that artifacts remained available on both GitHub-based repositories and long-term storage platforms. We found only one exception: an artifact was removed from Zenodo at the author's request, leaving a tombstone. So far, the community considered services like Zenodo as long-term storage without exception and the gold standard. However, if these services allow removals, we may need to reconsider their use. Additionally, the community must decide whether a paper should lose its badge if the artifact is later removed. On a related note, if the upload to permanent storage is requested only at the end of the evaluation process, which is often the case to ease the initial submission and accommodate improvements, additional work is required for the AEC. We observed multiple cases of authors inadvertently removing key scripts or data when polishing the artifact, and thorough reviewers were the key to spot any missing key materials compared to what the committee had evaluated.

Challenge 6: Imprecise badge definitions. ACM established a policy for artifact badging to accommodate the needs of multiple communities, and conferences edited with other publishers later created analogous versions of such badges. At a closer look, the ACM badge definitions are more descriptive than prescriptive. This naturally introduces some variation into the practical implementation of an artifact evaluation process, by allowing chairs, authors, and reviewers to exercise discretion. This leaves room for potential inconsistencies across different venues or even editions of the same conference. For the availability badge, the permanent nature of the storage location (Challenge 5) and the suitability of the chosen license remain unclear. For the functionality badge, the ACM policy hints at consistency (i.e., contributing to the paper results' generation) and completeness, but many venues, including EuroSys, require only a minimal working example to be runnable. For the documental part, the "sufficient description" for the artifacts to be exercised

opens to discretion in determining whether, for example, high-level documentation for running scripts and software components is adequate to compensate for completely undocumented source code. For the reproducible badge, the ACM policy generically states that the "main results" could be obtained in a subsequent study by a third party "using, in part, the authors' artifacts." Summarizing, the current badging system contains a *hidden curriculum* of implicit criteria that are likely only known by authors and reviewers who have already been through multiple artifact evaluations.

4.2 Proposed Directions

Based on the challenges in the previous section, we now derive short term and long term suggestions on how to continuously improve and solidify the artifact review process. The main challenge remains the tight review timeframe and we suggest to start artifact evaluation earlier. Another challenge is the loose coupling that results in information loss between the paper review and the artifact evaluation. We therefore suggest some information exchange between the committees and some standardization of the information.

Figure 5 visually depicts the connections between the proposed directions and the main challenges they address, a criterion we used also to numerically identify directions in a compacted form.

4.2.1 Short-Term Proposals.

Short-term Proposal S1: Start artifact preparation after early-reject deadlines. The tight race between paper notification, artifact evaluation, and camera-ready deadline can be lessened if we start the artifact preparation process earlier. The early-reject deadline is an intermediate step in reviewing that is well defined and adopted in most major conferences. We suggest that authors express a commitment to artifact evaluation when they submit the paper, and that TPC chairs send out, alongside the notification of round-2 paper advancement, a note about the upcoming start of the AE process and instructions to submit artifacts no later than the day after paper notification. This gives the authors time to prepare the artifact (usually a few weeks) and would allow the kick-the-tires to start right after the paper notification. This would alleviate the tight constraints of the artifact evaluation and give the reviewers more time to reproduce the claims, reducing the pressure of Challenge 1. At the same time, as many round-2 papers are ultimately rejected, authors may decide to opportunistically wait, for example, until the rebuttal phase before they actively work on the artifact. Nevertheless, there is a risk of putting effort on artifacts that will eventually undergo changes due to the paper improvements needed for a resubmission, and this may frustrate authors. We acknowledge that our proposal is imperfect, and the only feasible long-term solution we foresee is to fully decouple the artifact evaluation timeline from the camera-ready deadline, which in turn mandates cooperation with publishers or a different model to promote paper badge visibility.

Short-term Proposal S2: Informal TPC-AE channel. An immediate direction for further improvement is introducing an, initially informal, channel between the TPC and the AE. Using an open-text field in the reviews, the paper reviewers can encode what claims and core experimental results they expect the artifact to satisfy. Even if forwarding the complete reviews is not feasible, forwarding this compartmentalized field would already help the AE reviewers to better assess the quality and completeness of the artifact. Similarly, the TPC should extend the paper submission form to include a section that lists the main claims of the artifact. The reviewers could then assess these claims and comment on them in their reviews. The claims from the paper and the section of the review about the artifact claims would then be forwarded to the artifact evaluation. This would partially mitigate Challenge 2.

Short-term Proposal S3: Define available hardware or require author support. We suggest that artifact evaluations clearly define the available hardware that authors can expect reviewers to have. Authors must then ensure that their artifacts run on this available hardware (and knowing the specifics may help authors design scaled-down experiments when applicable) or must alternatively provide access to special hardware on an as-needed basis to support artifact evaluation. This would partially mitigate Challenge 3. Chairs can still survey AEC members for special hardware, but only with the purpose of conducting independent experiments that add to those to be conducted on authors-supplied resources.

Short-term Proposal S5: Artifact availability. We observed that artifacts may be removed after the artifact evaluation process and cease to be available. A simple first step is to require that all artifacts be available through DOI-backed platforms. In neighboring communities, this policy has been enforced, for example, by NDSS and later also by USENIX Security. Nevertheless, while platforms like Zenodo offering DOI-backed storage are a great asset to artifact evaluation initiatives, we noted that authors may occasionally be able to delete uploaded artifacts. The process should be regulated so that artifacts are locked to a paper and can only be removed with approval from the chairs or the steering committee. Deleting artifacts arbitrarily should cause a paper to lose its badges or be redacted. This addresses Challenge 5 and would require collaboration with service providers. However, questions remain about the long-term availability of artifacts [37]. For ACM conferences, a natural option would be to host them in the Digital Library, but similar provisions would also need to be made by other publishers.

4.2.2 Long-Term Proposalss.

Long-term Proposal L1: Making AE mandatory. Requiring some form of AE for all accepted papers will tighten the ties between academic papers and implementation prototypes along with their evaluation. Several conferences have already started enforcing artifact evaluation and we believe this will be the path forward to address Challenge 1. Authors that commit to making artifacts available, know during paper submission that they will have to prepare the artifact as the review of their paper progresses, thereby starting the artifact evaluation process in a more open mind. We are aware this requirement may be controversial, and conferences will have to offer an opt-out for experience papers and for papers with industry authors or partners behind the research. Even longer term, the paper acceptance could be tied to successful artifact evaluation. For academic papers, we should expect artifact availability for the sake of open science, and later on functionality, whereas reproducibility is too an elusive goal and also unrealistic for general feasibility.

Long-term Proposal L2: Enforcing constraints from the paper review. Following up on Short-term Proposal S2, we suggest making the information channel from paper review to artifact evaluation

explicit. Using dedicated fields in their reviews, the paper reviewers would make the expectations of the artifact explicit. While this can initially be free text, we could develop a more structured approach as well. Ideally, the reviewers would encode a set of constraints that are then shared with the authors, similarly to the review summary that is now appended to papers at IEEE Security and Privacy where reviewers encode their noteworthy concerns for threats to validity. An analogous summary could contain expectations of the artifact that are made public in the final paper. The artifact evaluation could then evaluate these expectations and provide post-assessment remarks that become integral to the summary before publication. This would satisfy and mitigate Challenge 2. For papers whose technical contributions heavily rely on artifacts (e.g., tool papers), acceptance could even be tied to successful artifact evaluation according to the author-submitted and TPC-validated claims.

Long-term Proposal L4: AE Steering Committee. We propose to establish a steering committee for artifact evaluation whose mission includes maintaining guides for chairs, evaluators, and authors; as well as collecting and archiving statistics and author feedback to monitor the health of artifact evaluation efforts. One possible path would be to provide a formal version of sysartifacts.github.io [25] and discussions between the steering committee and the AE chairs of individual conferences. This will address Challenge 4 and 6. The steering committee may also orchestrate initiatives to reward participants, starting by awarding particularly impactful artifacts in the long term—where impact can be demonstrated, for example, by use in several publications from other research groups.

Long-term Proposal L5: Explicit commitment to artifacts in papers. The submission instructions could be adjusted to include an explicit section about artifact and data availability, similar to the Open-Science policy recently initiated at USENIX Security [8]. Similar to the Ethics sections that are now being required for several computer security conferences, we could enforce a section on, say, "Data availability" or "Artifact availability" that encodes which claims and experimental results will be reproducible through the released artifacts. The TPC can use this section as part of reviews, i.e., accepting the paper means they believe it is enough, or they can conditionally accept if authors agree to provide more. The artifact evaluation committee then uses this section as the bar the artifact must meet. If artifact evaluation for that bar fails, the TPC reviewers for the paper must decide whether the paper is still accepted. This requires little effort from the TPC as the burden of listing claims is on the authors. It preserves the non-mandatory aspect of AE, but empowers the TPC. It also provides a strong incentive to authors to not drop out of artifact evaluation. This proposal directly addresses Challenge 5, but also contributes positively to the currently missing TPC-AEC interaction (Challenge 2) and forces authors to think hard how others can independently reproduce their results (Challenge 3).

5 AE Across Computer Science

Within the computer systems community, artifact evaluation is now widely accepted, with major venues consistently running an artifact evaluation program after paper acceptance [25]. However, the process is not standardized [44]. While computer systems venues use badges similar to the ACM's, not all venues impose the same requirements for each badge [46].

The perception of artifact evaluation by systems researchers is mostly positive [48], though with more focus on *reusability* than *reproducibility*. There is also a fear that badges could backfire by discouraging researchers from working on topics that do not lend themselves to artifact evaluation. Some community members fear that if not getting AE badges becomes synonymous with not having a good paper in the mind of the public, empirical research that is hard to reproduce such as new hardware will be less attractive.

In addition, systems venues frequently publish papers from large organizations that are based on extensive, long-term system deployments. This leads to resistance against the idea of implementing mandatory artifact evaluation, as sharing the associated artifacts in these papers often faces various constraints related to privacy, business interests, and scalability [42, 48].

Other fields of computer science have initiatives to help reproducibility. The high-level goal of making scientific results reusable is widely accepted. Plale et al. [45] surveyed the high-performance computing community and found that only 15% think reproducibility concerns are exaggerated. However, there are wide variations in the kinds of processes used to achieve this goal. Sedghpour et al. [46] found wide differences in requirements from artifacts and evaluators across the many computer science conferences that accept work in distributed systems. Hermann et al. [38] surveyed software engineering and programming language artifact evaluators, finding that while most evaluators agree the process is useful, there is little consensus on what exactly that process should be. Security venues evaluate artifacts in a similar fashion to systems ones, but recently moved further with USENIX Security requiring artifact availability unless authors can provide a compelling reason [23]. Software engineering venues evaluate artifacts but for availability and reusability, additionally requiring data availability statements in submissions [9, 16, 17]. Formal verification venues also evaluate artifacts for availability and reusability, with the additional requirement that tool papers must pass the artifact evaluation process [15, 26]. Mobile systems venues focus on artifact availability and generally discourage reproducibility badges during the AE process [12]. Machine learning venues do not have AE processes, but other initiatives such as the ML Reproducibility Challenge [22].

6 Conclusion

Artifact evaluation is a process aimed at verifying whether the artifacts published alongside a research paper—such as software, data, and documentation—adequately support the research described in the paper. The primary goal of artifact evaluation is to ensure the reproducibility of the published results and to promote the reuse of scientific findings. In this paper, we share data and insights from our experience conducting this process over five consecutive editions of a major systems conference. In our experience, we faced recurring challenges within the artifact evaluation process and noted that each evaluation instance, as well as the conference itself, had limited decision-making power and resources to address them. We describe those main challenges and propose both short-term and long-term solutions that may help tackle these problems in future artifact evaluation initiatives. Our primary objective is to provide practical guidance for the next five years of artifact evaluation at EuroSys. Additionally, we hope our report will encourage broader discussions on the key goals and essential practices necessary for

establishing effective artifact evaluation processes for computer science research in general.

Acknowledgments

We are grateful to our reviewers for their thorough comments and constructive feedback on this paper. We would also like to thank the AEC members from the past five EuroSys conferences for their significant contributions to the artifact evaluation process, and the entire EuroSys community for their commitment to promoting and establishing reproducibility practices at the conference.

References

[1] 2008. *ACM SIGMOD Experimental Repeatability Requirements.* http://www.sigmod08.org/sigmod_research.shtml
[2] 2011. *ESEC/FSE 2011 Call for Artifact Evaluation.* http://2011.esec-fse.org/cfp-artifact-evaluation
[3] 2011. *Software Heritage.* http://www.softwareheritage.org
[4] 2021. *EuroSys '21: Proceedings of the Sixteenth European Conference on Computer Systems* (Online Event, United Kingdom). Association for Computing Machinery, New York, NY, USA.
[5] 2022. *Artifact Evaluation report for EuroSys 2022.* https://sysartifacts.github.io/eurosys2022/report
[6] 2022. *EuroSys '22: Proceedings of the Seventeenth European Conference on Computer Systems* (Rennes, France). Association for Computing Machinery, New York, NY, USA.
[7] 2023. *EuroSys '23: Proceedings of the Eighteenth European Conference on Computer Systems* (Rome, Italy). Association for Computing Machinery, New York, NY, USA.
[8] 2024. *34th Usenix Security Symposium - Announcement and Preliminary Call for Papers.* https://www.usenix.org/sites/default/files/sec25_cfp_061824.pdf
[9] 2024. *Artifact Evaluation – ASE 2024.* https://conf.researchr.org/track/ase-2024/ase-2024-artifact-evaluation-track
[10] 2024. *EuroSys '24: Proceedings of the Nineteenth European Conference on Computer Systems* (Athens, Greece). Association for Computing Machinery, New York, NY, USA.
[11] 2025. *ACM Emerging Interest Group for Reproducibility and Replicability (EIGREP).* https://reproducibility.acm.org/
[12] 2025. *ACM SIGMOBILE Research Papers Artifact Evaluation Guidelines.* https://www.sigmobile.org/grav/about/artifact-guidelines
[13] 2025. *ACM SIGMOD ARI (Availability & Reproducibility Initiative).* https://reproducibility.sigmod.org/
[14] 2025. *Artifact Analysis Scripts.* https://github.com/secartifacts/artifact_analysis
[15] 2025. *Artifact Evaluation – CAV 2025.* https://conferences.i-cav.org/2025/artifact/
[16] 2025. *Artifact Evaluation – FSE 2025.* https://conf.researchr.org/track/fse-2025/fse-2025-artifacts
[17] 2025. *Artifact Evaluation – ICSE 2025.* https://conf.researchr.org/track/icse-2025/icse-2025-artifact-evaluation
[18] 2025. *Dryad's website.* https://datadryad.org/
[19] 2025. *EuroSys '25: Proceedings of the Twentieth European Conference on Computer Systems* (Rotterdam, Netherlands). Association for Computing Machinery, New York, NY, USA.
[20] 2025. *Figshare's website.* https://figshare.com/
[21] 2025. *Fuzzing workshop 2025.* https://fuzzingworkshop.github.io/
[22] 2025. *ML Reproducibility Challenge.* https://reproml.org/
[23] 2025. *Security Research Artifacts.* https://secartifacts.github.io/
[24] 2025. *SIGSOFT Artifact Evaluation Working Group.* https://github.com/acmsigsoft/artifact-evaluation
[25] 2025. *Systems Research Artifacts.* https://sysartifacts.github.io/
[26] 2025. *TACAS 2025.* https://etaps.org/2025/conferences/tacas/
[27] 2025. *Zenodo's website.* https://zenodo.org/
[28] Association for Computing Machinery. 2020. *Artifact Review and Badging Version 1.1.* https://www.acm.org/publications/policies/artifact-review-and-badging-current
[29] Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* 533, 7604 (May 2016), 452–454. doi:10.1038/533452a
[30] Bachir Bendrissou, Rahul Gopinath, and Andreas Zeller. 2022. "Synthesizing input grammars": a replication study *(PLDI 2022).* Association for Computing Machinery, New York, NY, USA, 260–268. doi:10.1145/3519939.3523716
[31] Philippe Bonnet, Stefan Manegold, Matias Bjørling, Wei Cao, Javier Gonzalez, Joel Granados, Nancy Hall, Stratos Idreos, Milena Ivanova, Ryan Johnson, David Koop, Tim Kraska, René Müller, Dan Olteanu, Paolo Papotti, Christine Reilly, Dimitris Tsirogiannis, Cong Yu, Juliana Freire, and Dennis Shasha. 2011. Repeatability and workability evaluation of SIGMOD 2011. *SIGMOD Rec.* 40, 2 (Sept. 2011), 45–48. doi:10.1145/2034863.2034873
[32] Christopher D. Chambers and Loukia Tzavella. 2022. The past, present and future of Registered Reports. *Nature Human Behaviour* 6, 1 (Jan. 2022), 29–42. doi:10.1038/s41562-021-01193-7
[33] Andy Cockburn, Pierre Dragicevic, Lonni Besançon, and Carl Gutwin. 2020. Threats of a replication crisis in empirical computer science. *Commun. ACM* 63, 8 (July 2020), 70–79. doi:10.1145/3360311
[34] Christian Collberg and Todd A. Proebsting. 2016. Repeatability in computer systems research. *Commun. ACM* 59, 3 (Feb. 2016), 62–69. doi:10.1145/2812803
[35] Enzian project. 2022. *ASPLOS Artifact Evaluation badges.* https://enzian.systems/updates/2022/02/04/Badges/
[36] Dror G. Feitelson. 2015. From Repeatability to Reproducibility and Corroboration. *SIGOPS Oper. Syst. Rev.* 49, 1 (Jan. 2015), 3–11. doi:10.1145/2723872.2723875
[37] Quentin Guilloteau, Florina Ciorba, Millian Poquet, Dorian Goepp, and Olivier Richard. 2024. Longevity of Artifacts in Leading Parallel and Distributed Systems Conferences: a Review of the State of the Practice in 2023 *(ACM REP '24).* Association for Computing Machinery, New York, NY, USA, 121–133. doi:10.1145/3641525.3663631
[38] Ben Hermann, Stefan Winter, and Janet Siegmund. 2020. Community expectations for research artifacts and evaluation processes *(ESEC/FSE 2020).* Association for Computing Machinery, New York, NY, USA, 469–480. doi:10.1145/3368089.3409767
[39] Shriram Krishnamurthi. 2013. Artifact evaluation for software conferences. *SIGPLAN Not.* 48, 4S (July 2013), 17–21. doi:10.1145/2502508.2502518
[40] Tanu Malik, Anjo Vahldiek-Oberwagner, Ivo Jimenez, and Carlos Maltzahn. 2022. Expanding the Scope of Artifact Evaluation at HPC Conferences: Experience of SC21 *(P-RECS '22).* Association for Computing Machinery, New York, NY, USA, 3–9. doi:10.1145/3526062.3536354
[41] Adam Marcus. 2013. *Influential Reinhart-Rogoff economics paper suffers spreadsheet error.* https://retractionwatch.com/2013/04/18/influential-reinhart-rogoff-economics-paper-suffers-database-error/
[42] Jeffrey C. Mogul, Priya Mahadevan, Christophe Diot, John Wilkes, Phillipa Gill, and Amin Vahdat. 2021. Data-driven networking research: models for academic collaboration with industry (a Google point of view). 51, 4 (Dec. 2021), 47–49. doi:10.1145/3503954.3503960
[43] National Information Standards Organization. 2021. Reproducibility Badging and Definitions. (2021). doi:10.3789/niso-rp-31-2021
[44] Solal Pirelli. 2022. *Artifact evaluation, present and future.* https://www.sigops.org/2022/artifact-evaluation-present-and-future/
[45] Beth A. Plale, Tanu Malik, and Line C. Pouchard. 2021. Reproducibility Practice in High-Performance Computing: Community Survey Results . *Computing in Science & Engineering* 23, 05 (Sept. 2021), 55–60. doi:10.1109/MCSE.2021.3096678
[46] Mohammad Reza Saleh Sedghpour, Alessandro Vittorio Papadopoulos, Cristian Klein, and Johan Tordsson. 2024. Artifact Evaluation for Distributed Systems: Current Practices and Beyond. arXiv:2406.13045 [cs.DC] https://arxiv.org/abs/2406.13045
[47] Uri Simonsohn, Leif Nelson, and Joe Simmons. 2023. *Data Falsificada (Part 1): "Clusterfake"'.* https://datacolada.org/109
[48] Roberta De Viti, Solal Pirelli, and Vaastav Anand. 2023. HotOS XIX Panel Report: Panel on Future of Reproduction and Replication of Systems Research. arXiv:2308.05762 [cs.OS] https://arxiv.org/abs/2308.05762

A 5 Years of AE at EuroSys in Numbers

In this section, we provide additional raw data that was used to generate the figures and statistics reported in Section 3.2.

A.1 Artifact Analysis Scripts

We have developed a set of simple Python scripts to collect and analyze the data and statistics presented in this paper. These scripts are available at https://github.com/secartifacts/artifact_analysis/tree/eurosys25 for others to reproduce our results. They perform the following tasks: scraping data about artifact evaluation results and Artifact Evaluation Committees, verifying the existence of repositories and DOIs, and computing simple statistics on both the artifacts, badges and evaluation committees.

A.2 Badging Results

We collected statistics on AE badging results for each year's sysartifacts.github.io and the EuroSys proceedings' front matters, which are displayed in Table 1 for each year. For the years 2023 to 2025, we considered two submission cycles: Spring and Fall. The following data points were collected:

Papers accepted: Accepted Papers at EuroSys in this year.

Artifact submissions: Voluntary artifact submissions from all accepted papers.

% artifact submissions: Percentage of artifact submissions from accepted papers.

Spring submissions: Artifacts submitted in the spring cycle.

Fall submissions: Artifacts submitted in the fall cycle.

AEC size: Size of the artifact evaluation committee.

Artifact Available: Number of the awarded artifact available badges.

Available acceptance rate: Rate of successful available evaluations.

% available AE sub.: Percentage of successful available badges per AE submissions.

% available Paper: Percentage of successful available badges per accepted papers.

Artifact Functional: Number of the awarded artifact functional badges.

Functional acceptance rate: Rate of successful functional evaluations.

% functional AE sub.: Percentage of successful functional badges per AE submissions.

% functional Paper: Percentage of successful functional badges per accepted papers.

Results Reproduced: Number of the awarded reproduced available badges.

Reproduced acceptance rate: Rate of successful reproduced evaluations.

% reproduced AE sub.: Percentage of successful reproduced badges per AE submissions.

% reproduced Paper: Percentage of successful reproduced badges per accepted papers.

A.3 AE Committes

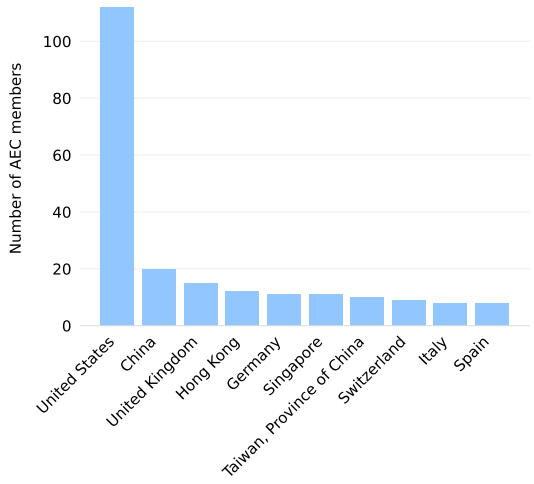
In Section 3.2, we reported on the geographic distribution of the AEC members by continent. Here, we provide information about the number of AEC members in each country. Figure 6a and Figure 6b illustrate respectively the total membership numbers for the top ten countries and their geographical distribution over the analyzed years.

A.4 Storage Services

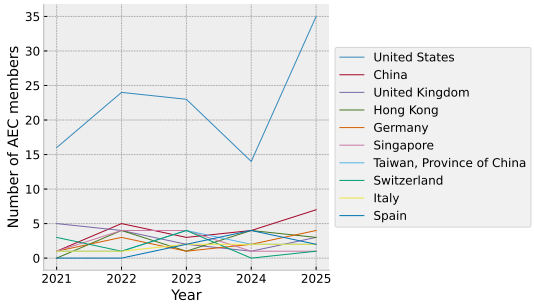
Table 2 reports the number of artifacts collected each year and the corresponding storage services used. In 2021 and 2022, both Git repositories and DOI-based storage were accepted, and at least one of these had to be provided to qualify for the available badge. In 2023 and 2024, only DOI-based storage services were allowed and collected for the available badge. Most recently, in 2025, both Git repositories and DOI-based storage services were allowed, resulting

Year	2021	2022	2023	2024	2025
Papers accepted	38	45	54	71	85
Artifact submissions	22	33	32	33	45
% artifact submissions	58%	73%	59%	47%	53%
Spring submissions	SD	SD	15	23	15
Fall submissions	SD	SD	17	10	30
AEC size	50	65	64	49	93
Artifact Available	21	33	31	32	44
Available acceptance rate	100%	100%	100%	100%	100%
% available AE sub.	96%	100%	97%	97%	98%
% available Papers	55%	73%	57%	45%	52%
Artifact Functional	18	27	24	25	42
Functional acceptance rate	90%	96%	80%	90%%	98%
% functional AE sub.	81%	82%	75%	76%	93%
% functional Papers	47%	60%	44%	35%	49%
Results Reproduced	14	20	8	12	21
Reproduced acceptance rate	74%	77%	40%	72%	75%
% reproduced AE sub.	64%	61%	25%	36%	47%
% reproduced Paper	37%	44%	15%	17%	25%

Table 1: AE badging data from EuroSys 2021 to 2025. Percentages rounded to the nearest integer. 'SD': single deadline.



(a) Total membership numbers.



(b) Geographical distribution of AEC members over time.

Figure 6: AEC distribution per top ten countries.

Year	2021	2022	2023	2024	2025
FigShare	-	-	2	3	-
Github	20	33	-	-	24
Zenodo	1	31	29	29	20

Table 2: Number of artifacts per storage service.

Year	2021	2022	2023	2024	2025
Inaccessible Repositories	1	1	-	-	-
Inaccessible DOI	0	2	1	0	0
Inaccessible Artifacts %	~5%	0%	~3%	0%	0%

Table 3: Inaccessible artifacts on March 17th 2025.

Year	2021	2022	2023	2024	2025
Figshare views	-	-	122 [89:156]	336 [55:345]	-
Figshare downloads	-	-	21 [13:29]	43 [27:48]	-
Github stars	12 [2:572]	17 [1:287]	-	-	0 [0:81]
Github forks	2 [0:44]	5 [0:44]	-	-	0 [0:8]
Zenodo views	498*	138 [79:341]	82 [44:286]	66 [20:197]	30 [3:159]
Zenodo downloads	133*	12 [5:55]	14 [2:189]	6 [1:116]	7 [0:82]

Table 4: Median views, downloads, stars and forks of artifacts from EuroSys 2021 - 2025. Min/Max in brackets, except for single-entry cells marked with *.

in a mixed collection for the available badge.

Table 3 presents the number of artifacts that were inaccessible at the time of writing this paper. For the artifacts from 2021, one GitHub repository is no longer available and the associated artifact is no longer accessible. For the artifacts from 2022, one GitHub repository is no longer available, and two artifacts did not provide a DOI-accessible version for AE. In all three cases, the alternative was still accessible ensuring availability of the artifacts. For the artifacts from 2023, one Zenodo repository was removed upon request by the user. In all other years, we have not found inaccessible artifacts.

Table 4 holds detailed data for artifact storage service statistics for all artifacts that are still accessible at the time of writing this paper by each year.