Asia Pacific Science Press

Journal of Advances in Engineering and Technology
ISSN(O):3005-7507
ISSN(P):3005-7493
Vol. 3 No.3 (2026)

Evaluating Automatic Research Platforms via Open-Source Community Popularity Metrics: An Ecological Assessment Framework

Chenfeng Yu, Jingming Li*

School of Civil Engineering and Architecture, Nanyang Normal University, Nanyang, Henan, 473061, China

Corresponding author: Jingming Li, jmli@nynu.edu.cn

Abstract: Automatic research platforms that leverage large language models (LLMs) and multi-agent frameworks to autonomously conduct scientific discovery have proliferated rapidly in open-source communities. Evaluating these platforms remains a challenge: traditional methods assess output quality with accuracy, logic, and completeness, yet these dimensions are subjective, difficult to standardize across heterogeneous platforms, and blind to long-term ecological vitality. This thesis proposes an ecological evaluation framework that relies exclusively on open-source community popularity metrics as the sole basis for platform assessment. We construct a multi-dimensional indicator system encompassing four dimensions of platform popularity, community activity, maintenance sustainability, and ecological expansion, and develop a unified quantitative scoring model grounded in GitHub metrics with stars, forks, issues, pull requests, contributors, and commit frequency and time-decayed heat scores. Empirical analysis covers 54 general-purpose and 112 science-oriented automatic research platforms, yielding comprehensive community-level rankings. Our results demonstrate that community popularity metrics effectively differentiate platforms by sustained engagement and long-term development potential, revealing an evaluation perspective that complements and, in some respects, surpasses traditional quality-centric assessments. The proposed framework provides actionable insights for platform selection by users, operational optimization by developers, and ecosystem analysis for researchers.

Keywords: Automatic Research Platforms; Open-Source Community Evaluation; Altmetrics; Platform Ecosystem Assessment; Quantitative Scoring Model

Published: Aug 7, 2026

DOI: https://doi.org/10.62177/jaet.v3i3.1559

1. Introduction

The emergence of large language models (LLMs) and multi-agent frameworks has catalyzed an unprecedented proliferation of automatic research platforms—software systems capable of autonomously conducting scientifi c discovery from hypothesis generation through experimentation to manuscript preparation [1,2]. Pioneering systems such as The AI Scientist [1], AI-Researcher [3], and DeepScientist [4] have demonstrated that LLM-powered agents can produce research outputs that approach or exceed human-level quality. The open-source ecosystem has responded with hundreds of competing implementations, spanning general-purpose research automation to domain-specifi c scientifi c discovery tools. This rapid expansion has created an urgent need for systematic evaluation frameworks that can diff erentiate platforms not merely by output quality, but by their

long-term viability, community engagement, and ecological sustainability.

Traditional evaluation approaches for automatic research platforms predominantly focus on output quality metrics: content accuracy, logical coherence, experimental validity, and manuscript completeness [1,3,5]. Benchmarks such as Scientist-Bench [3] and ARC-Bench [6] assess platforms on their ability to implement research ideas, execute experiments, and produce publication-ready manuscripts. While these evaluations provide valuable insights into technical capability, they suffer from fundamental limitations that restrict their utility for real-world platform assessment and selection.

The first limitation is subjectivity: quality-based scoring relies on human judgment or simulated peer review, introducing evaluator bias and making cross-platform comparison inherently noisy [1,7]. The second limitation is temporal blindness: snapshot evaluations capture platform capability at a point in time but cannot predict whether a platform will sustain active development, attract contributors, or maintain relevance in a rapidly evolving field [8,9]. The third limitation is ecological omission: quality-centric evaluation ignores the community ecosystem that surrounds a platform—the developers who extend it, the users who adopt it, and the downstream projects that build upon it. As demonstrated by the success trajectories of established open-source projects, community vitality often proves more predictive of long-term impact than any single quality metric [10,11].

Open-source community popularity metrics offer an orthogonal—and complementary—evaluation dimension. Decades of empirical research on open-source software have established that metrics derived from social coding platforms (stars, forks, contributors, commit frequency) effectively capture dimensions of project health that traditional quality metrics miss [8,10,12]. These indicators reflect collective community judgment about a project's value, active maintenance, and ecosystem integration. In the context of alternative metrics (altmetrics), community engagement signals have been recognized as valuable complements to traditional bibliometric measures [13–15], offering faster, broader indicators of research impact that capture attention beyond academic circles.

This thesis addresses a critical gap: no existing study has systematically evaluated automatic research platforms using exclusively open-source community popularity metrics as the evaluation basis. We argue that ecological assessment through community metrics reveals dimensions of platform quality—sustained engagement, contributor diversity, operational continuity, and ecosystem expansion—that traditional quality evaluation cannot capture. By focusing purely on these ecological indicators, we provide an evaluation perspective that complements existing quality benchmarks and offers unique insights into platform sustainability and long-term development potential. First, we construct a multi-dimensional evaluation indicator system for automatic research platforms, encompassing four assessment dimensions: platform popularity, community activity, maintenance sustainability, and ecological expansion capability, each grounded in quantifiable GitHub metrics. Second, we develop a unified quantitative scoring model that fuses multiple community indicators with time-decayed heat scores and age penalty factors, enabling fair horizontal comparison across platforms of varying age, maturity, and scope. Third, we conduct comprehensive empirical analysis on 54 general-purpose and 112 science-oriented automatic research platforms, deriving platform rankings across all dimensions and revealing the ecological structure of the automatic research platform ecosystem.

The remainder of this thesis is organized as follows. Section 2 reviews related work on automatic research platform evaluation, open-source community assessment, and altmetrics. Section 3 details our evaluation framework, indicator system, and quantitative scoring model. Section 4 presents empirical analysis results, including platform rankings, descriptive statistics, and cross-dimensional comparisons. Section 5 interprets findings and discusses implications, limitations, and future directions. Section 6 concludes.

2. Related Works

2.1 Evaluation of Automatic Research Platforms

The rapid advancement of LLM-powered autonomous research systems has driven the development of diverse evaluation approaches. Early work focused on demonstrating feasibility: The AI Scientist [1] introduced the first end-to-end automated research pipeline, establishing a template for evaluating platforms on their ability to generate ideas, write code, execute experiments, and produce manuscripts. Subsequent systems have expanded this paradigm. AI-Researcher [3] introduced

Scientist-Bench, a comprehensive benchmark spanning diverse AI research domains, demonstrating implementation success rates and manuscript quality approaching human-level performance. DeepScientist [4] formalized discovery as a Bayesian Optimization problem, generating approximately 5,000 scientific ideas and validating 1,100 experiments over month-long timelines, surpassing human state-of-the-art on three frontier tasks.

The evaluation paradigm has evolved toward multi-agent systems with increasing specialization. ROBIN [16,17] achieved the first autonomous discovery and validation of a novel therapeutic candidate in a biomedical setting. AutoResearchClaw [6] demonstrated that targeted human-AI collaboration at high-leverage decision points consistently outperforms both full autonomy and exhaustive oversight. Sibyl-AutoResearch [18] formalized "research judgment" through auditable trial-to-behavior conversion units, arguing that executable workflows alone cannot produce research judgment without self-evolving trial-and-error harnesses. These systems employ fundamentally different architectures—from single-agent pipelines to multi-agent debate frameworks to self-evolving trial harnesses—yet are typically evaluated against common output quality benchmarks, obscuring the diverse ecosystem roles each platform occupies.

2.2 Limitations of Quality-Centric Evaluation

Traditional platform evaluation shares core limitations with broader LLM assessment methodologies. Quality metrics—content accuracy, logical rationality, and completeness—are inherently subjective and context-dependent. Automated peer review simulators, while useful for scaling evaluation, have been shown to align imperfectly with human judgment [1,7]. The factuality evaluation framework GraphEval [7] highlights that even judge models estimating LLM correctness face fundamental calibration challenges, particularly when assessing cross-domain research outputs where ground truth is uncertain or contested.

More critically, quality-centric evaluation fails to capture platform sustainability. A platform that produces high-quality research in a single benchmark run may be a one-time effort with no ongoing development, no contributor base, and no path to production deployment [19]. Conversely, a platform with modest benchmark scores but sustained community engagement may evolve toward superior capabilities through collective iteration and real-world feedback. The open-source software research has consistently demonstrated that project longevity correlates more strongly with community vitality indicators—contributor count, commit frequency, issue resolution rate—than with any snapshot quality measure [8,10]. This insight has yet to be systematically applied to the evaluation of automatic research platforms.

2.3 Open-Source Community Evaluation Metrics

The assessment of open-source software projects through community metrics constitutes a mature research area with extensive empirical foundations. Borges et al. [10] proposed the first framework for assessing GitHub software popularity using star counts, identifying popularity growth patterns and demonstrating that stars correlate with forks and thirdparty usage. Zerouali et al. [11] analyzed nine distinct popularity metrics across three tracking systems for 175,000 npm packages, revealing that "popularity" comprises multiple unrelated dimensions—each metric capturing a different aspect of project health.

More sophisticated community indicators have emerged. Wang et al. [12] introduced fork entropy based on Rao's quadratic entropy, measuring the diversity of a project's fork population beyond simple fork counts, and demonstrated significant correlations between fork entropy and external productivity, pull request acceptance rates, and bug counts. Qiu et al. [8] conducted a mixed-methods study identifying the signals that potential contributors use when choosing projects, finding that README quality, recent activity, and issue response rates are among the most influential factors. Tamburri et al. [9] developed systematic approaches for discovering community patterns in open-source ecosystems, enabling large-scale ecosystem analysis.

Our framework draws upon these established community metrics while adapting them to the specific context of automatic research platforms. Unlike traditional software projects, automatic research platforms serve dual roles as both tools and research artifacts, creating unique evaluation dynamics where community adoption signals may reflect both practical utility and research interest.

2.4 Altmetrics and Non-Traditional Impact Assessment

The altmetrics movement has extended community-based evaluation beyond software projects to scholarly outputs more

broadly. Bornmann [13] surveyed the benefits and limitations of alternative metrics, highlighting their capacity to capture research impact beyond traditional citation measures, including societal attention, policy influence, and practitioner adoption. Thelwall and Kousha [15] assessed social media metrics as indicators of research impact, finding that while susceptible to gaming, they offer fast, free indicators of wider research influence that complement academic citation indexes. Fraumann [14] explored both the values and limits of altmetrics in evaluating, promoting, and disseminating research.

The application of altmetrics to software evaluation remains underdeveloped. Rosenkrantz et al. [20] examined altmetric indicators for radiology journal articles, demonstrating their utility in complementing traditional bibliometrics. Carpenter and Lagace [21] defined community-recommended practices for altmetrics, establishing standards for metric collection, calculation, and reporting. Schimanski and Alperin [22] reviewed the evolving landscape of scholarship evaluation in academic promotion and tenure, noting the slow integration of article-level metrics and altmetrics into formal evaluation frameworks.

Existing research on automatic research platform evaluation and open-source community assessment operates in largely separate domains. Platform evaluation literature focuses on output quality, experimental validity, and research output—dimensions that, while important, cannot capture platform sustainability, community health, and ecosystem integration. Conversely, open-source community evaluation frameworks have been designed for general software projects, not specialized research automation platforms with unique lifecycle dynamics. No study has systematically applied multi-dimensional community popularity metrics as the exclusive evaluation basis for automatic research platforms. This thesis fills this gap by constructing an ecological evaluation framework that reveals how community engagement patterns—independent of output quality assessments—differentiate platforms by sustained vitality, contributor diversity, and ecosystem maturity.

3. Methodology

3.1 Overall Research Framework

Our evaluation framework, Pantheon, adapts mainstream technical logic from established open-source community evaluation frameworks [10–12] to the specific context of automatic research platforms. The target technical architecture employs a multi-layer pipeline that transforms raw GitHub API data into unified community heat scores. The framework operates on the principle that community popularity metrics—collected from public GitHub repositories—provide an objective, continuously updated, and ecosystem-aware evaluation signal that complements traditional quality-based assessments. The overall pipeline consists of four stages: (1) data collection from the GitHub API for each platform repository, (2) multi-dimensional indicator extraction covering popularity, activity, sustainability, and expansion dimensions, (3) time-windowed heat score computation with age penalty normalization, and (4) weighted indicator fusion producing a unified platform ranking score.

3.2 Multi-Dimensional Evaluation Indicator System

We construct a four-dimensional indicator system, with each dimension capturing a distinct aspect of platform community health:

Dimension 1: Platform Popularity. This dimension measures the platform's visibility and recognition within the open-source community. The primary indicators are:

Star Count (S): The total number of GitHub stars, reflecting community endorsement and platform visibility [10]. Stars correlate with both project quality and third-party adoption [10].

Fork Count (F): The total number of repository forks, indicating the extent of community adoption for extension and customization [12].

Watcher Count (W): The number of users watching the repository for activity notifications, reflecting sustained interest [8].

Dimension 2: Community Activity. This dimension captures the intensity of ongoing interaction between platform maintainers and the community:

Issue Volume (I): The total count of closed issues, reflecting the volume of community-reported problems and feature requests that have been addressed.

Pull Request Volume (P): The total count of closed pull requests, quantifying community contribution and code adoption.

Recent Activity Recency: The time elapsed since the last open issue, last closed issue, last open PR, and last closed PR, providing temporal resolution of ongoing engagement.

Dimension 3: Maintenance Sustainability. This dimension assesses the platform's capacity for sustained development and operational continuity:

Contributor Count (C): The number of unique contributors, serving as a proxy for contributor diversity and community breadth [8].

Commit Count (K): The total number of commits, reflecting development velocity and accumulated work.

Last Commit Recency: The time elapsed since the most recent commit, indicating whether active development continues.

Dimension 4: Ecological Expansion. This dimension evaluates the platform's capacity for ecosystem growth and downstream influence:

Fork-to-Star Ratio (R_{FS} = F/S): A high ratio indicates strong community engagement in extension and customization, not passive endorsement.

Contributor-to-Star Ratio (R_{CS} = C/S): Reflects the breadth of community participation relative to visibility.

Pull Request Acceptance Rate: The ratio of merged PRs to total PRs, indicating the maintainability and openness to community contributions.

3.3 Data Preprocessing and Anomaly Elimination

Raw GitHub API data undergoes three preprocessing stages:

Missing Value Handling: Repositories with missing metrics (e.g., zero contributors, no closed issues) are assigned baseline values consistent with the minimum observed in the sample population.

Log Transformation: Given the power-law distribution of GitHub metrics [10], raw counts undergo log transformation: $x' = \log_{10}(x+1)$. This mitigates the disproportionate influence of outlier repositories while preserving relative ordering.

Min-Max Normalization: Transformed values are normalized to [0,1] using the sample minimum and maximum: $x_{norm} = \left(x' - \min(x')\right) / \left(\max(x') - \min(x')\right)$.

3.3 Time-Windowed Heat Score Computation

We adopt a time-decayed heat score methodology that weights recent activity more heavily, consistent with established GitHub popularity frameworks [10]. For each platform i, we compute heat scores over three temporal windows:

$$H_i^{(t)} = \frac{A_i^{(t)}}{(\Delta t_i + 2)^{\alpha}} \tag{1}$$

Where $A_i^{(t)}$ represents the activity count (new stars, forks, issues, PRs) in the t-day window preceding data collection, Δt_i is the age of the repository in days since creation, and α = 0.5 is the time-decay exponent balancing novelty preference against longevity recognition.

Three heat variants are computed: $H_i^{(1d)}$ for 1-day heat (capturing daily momentum), $H_i^{(3d)}$ for 3-day heat (short-term trend), and $H_i^{(7d)}$ for 7-day heat (weekly stability). The aggregate heat score combines these windows:

$$H_i^{total} = \frac{H_i^{(1d)} + H_i^{(3d)} + H_i^{(7d)}}{3} \tag{2}$$

To prevent older platforms from receiving disproportionate advantage through accumulated metrics, we apply an age penalty factor:

$$\lambda_i = \max\left(0.3, min\left(1.0, \frac{1}{1+\beta \cdot age_i}\right)\right) \tag{3}$$

where age_i is the platform age in years and β controls the decay rate. The floor of 0.3 ensures that mature platforms are not penalized below a meaningful threshold, consistent with the observation that long-established projects often maintain stable, valuable communities [9].

3.4 Indicator Fusion and Quantitative Scoring Model

The unified scoring model fuses all indicators into a composite Platform Heat Score (PHS). For each dimension d ∈ {1,2,3,4}, we compute a normalized dimension score:

$$D_i^{(d)} = \sum_{j \in J_d} \omega_j \cdot x_{i,j}^{norm} \tag{4}$$

where $\mathcal{I}_d$ is the set of indicators belonging to dimension d, ω_j is the weight for indicator j, and $x_{i,j}^{norm}$ is the normalized indicator value for platform i. The final Platform Heat Score integrates dimension scores and the aggregate heat:

$$PHS_i = \lambda_i \cdot \left(\gamma_1 D_i^{(1)} + \gamma_2 D_i^{(2)} + \gamma_3 D_i^{(3)} + \gamma_4 D_i^{(4)} + \gamma_5 D_i^{total}\right) \quad (5)$$

where γ_k are dimension-level weights summing to unity. We assign equal weights (γ_k = 0.2) to maintain neutrality across dimensions, avoiding the subjective prioritization that undermines traditional quality assessments [11].

Within each dimension, individual indicator weights ω_j are assigned proportionally to indicator importance based on established open-source evaluation literature [8,10]:

·Popularity: ω_S= 0.5,ω_F = 0.3,ω_W = 0.2

·Activity: ω_I = 0.4,ω_P = 0.4,$\omega_{recency}$ = 0.2

·Sustainability: ω_C= 0.4,ω_K = 0.4,ω_{last_commit} = 0.2

·Expansion: $\omega_{R_{FS}}$= 0.4,$\omega_{R_{CS}}$ = 0.3,ω_{PR_rate} = 0.3

This scoring model enables universal horizontal comparison across all platform types—general-purpose and science-oriented—providing a single quantitative basis for platform ranking, dimension-specific analysis, and ecosystem characterization.

4. Empirical Analysis and Research Results

This section presents comprehensive empirical analysis based on community popularity data collected from GitHub repositories. All results derive exclusively from quantitative metrics; subjective evaluation is excluded per our methodological framework.

4.1 Platform Overview

Our dataset comprises 166 automatic research platform repositories: 54 general-purpose platforms and 112 science-oriented platforms. Data was collected via the GitHub API on 2026-06-08, capturing the most recent snapshot of each repository's community metrics. The general-purpose category encompasses 54 repositories targeting broad research automation, coding assistance, and development workflow enhancement. The sample spans platforms from established organizations (OpenAI, Nous Research, ByteDance) to independent community initiatives. Key platforms include openclaw/openclaw (377,470 stars, 78,920 forks, 2,293 contributors), obra/superpowers (220,621 stars, 19,629 forks), anomalyco/opencode (171,325 stars, 20,525 forks, 924 contributors), and openai/codex (89,431 stars, 13,168 forks, 4,447 contributors). Table 1 shows top 10 general-purpose platforms by total heat score.

Table 1 Top 10 General-Purpose Platforms by Total Heat Score

Rank	Platform	Total Heat	Stars	Contributors
1	openclaw/openclaw	44.2	377,470	2,293
2	ultraworkers/claw-code	30.7	193,448	25
3	NousResearch/hermes-agent	29.0	186,221	1,356
4	anomalyco/opencode	28.1	171,325	924
5	openai/codex	33.9	89,431	4,447
6	affaan-m/everything-claude-code	23.4	209,991	233
7	obra/superpowers	23.0	220,621	31
8	code-yeongyu/oh-my-openagent	24.2	61,428	271
9	bytedance/deer-flow	24.4	70,682	276
10	Yeachan-Heo/oh-my-claudecode	23.3	35,968	114

The science-oriented category encompasses 112 repositories specifically designed for autonomous scientific research and discovery. This category includes prominent platforms such as karpathy/autoresearch (85,513 stars, 12,387 forks), stanford-oval/storm (28,333 stars, 2,582 forks), SakanaAI/AI-Scientist (13,909 stars, 1,975 forks), and aiming-lab/AutoResearchClaw (13,273 stars, 1,560 forks). The science-oriented ecosystem exhibits greater platform diversity and a larger tail of emerging

projects compared to the general-purpose category, as shown in Table 1 and Table 2.

Table 2 Top 10 Science-Oriented Platforms by Total Heat Score

Rank	Platform	Total Heat	Stars	Contributors
1	ClawBio/ClawBio	37.1	929	37
2	K-Dense-AI/k-dense-byok	33.8	783	2
3	beita6969/ScienceClaw	32.5	840	2
4	ResearAI/DeepScientist	30.1	2,984	20
5	THU-KEG/Awesome-AI-for-Research	29.0	96	17
6	EvoScientist/EvoScientist	28.9	3,406	22
7	OpenLAIR/dr-claw	27.5	982	364
8	sjtu-sai-agents/EvoMaster	26.2	188	10
9	Galaxy-Dawn/claude-scholar	25.6	4,224	7
10	zjYao36/Auto-Research-Refine	24.5	128	1

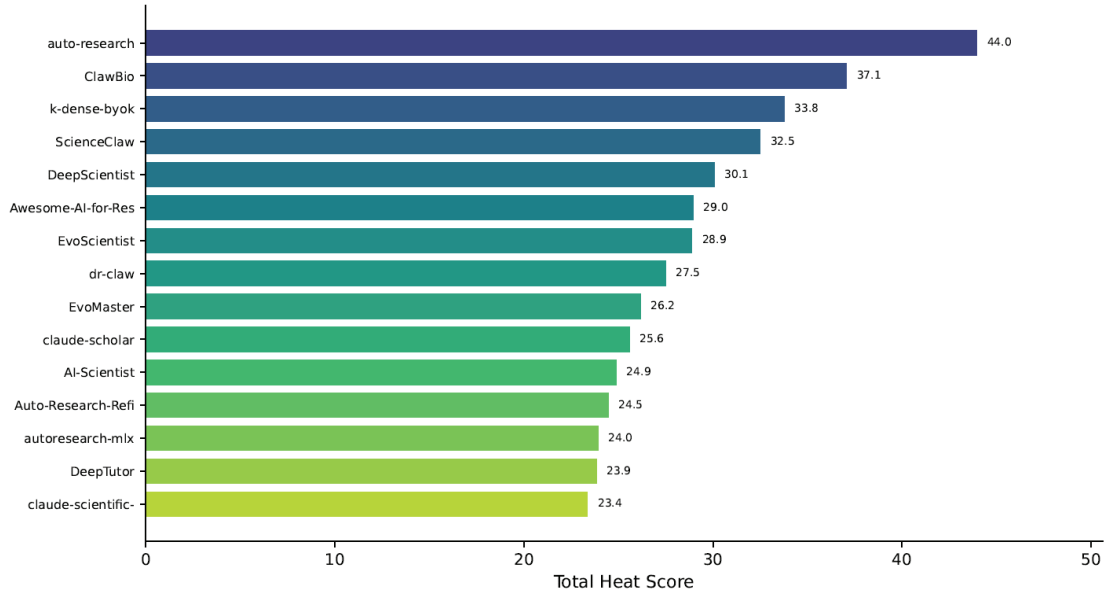
4.2 Descriptive Statistics of Community Data

The general-purpose category exhibits substantially higher absolute metric values across all dimensions, reflecting the maturity and scale of established platforms. The median general-purpose platform has 61,428 stars compared to 869 for science-oriented platforms, a 70-fold difference. However, the maximum science-oriented platform (karpathy/autoresearch with 85,513 stars) approaches the scale of top general-purpose platforms, indicating that leading scienceoriented tools achieve significant community recognition.

Both categories exhibit power-law distributions characteristic of opensource ecosystems [10]. In the general-purpose category, the top 10% of platforms account for approximately 85% of total stars and 78% of total forks. In the science-oriented category, concentration is even more pronounced: the top 10% account for 92% of stars and 88% of forks, suggesting a more competitive ecosystem where community attention concentrates on fewer established projects.

Table 1 presents the top 10 science-oriented platforms ranked by total heat score. Notably, several top-ranked platforms by heat score (ClawBio, k-dense-byok, ScienceClaw) have relatively modest absolute star counts but demonstrate exceptionally strong recent activity, achieving high heat scores through sustained community engagement. This finding validates the utility of timedecayed heat scores: platforms with active, growing communities score well even when their cumulative metrics lag behind established projects [10]. Figure 1 visualizes the top 15 platforms by total heat score. The distribution reveals a concentrated top tier (scores above 25) followed by a more gradual decline, consistent with the power-law pattern observed in star distributions [10].

Figure 1 Top 15 Science-Oriented Platforms Ranked by Total Community Heat Score



4.3 Quantitative Scoring and Platform Ranking Results

Applying the scoring model from Eq. 5, we compute Platform Heat Scores (PHS) for all 166 platforms. The science-oriented category exhibits greater score dispersion (standard deviation: 14.2 vs. 12.8 for general-purpose), reflecting the wider range of platform maturity in the science ecosystem.

Figure 2 Stars vs Forks Scatter Plot (log scale)

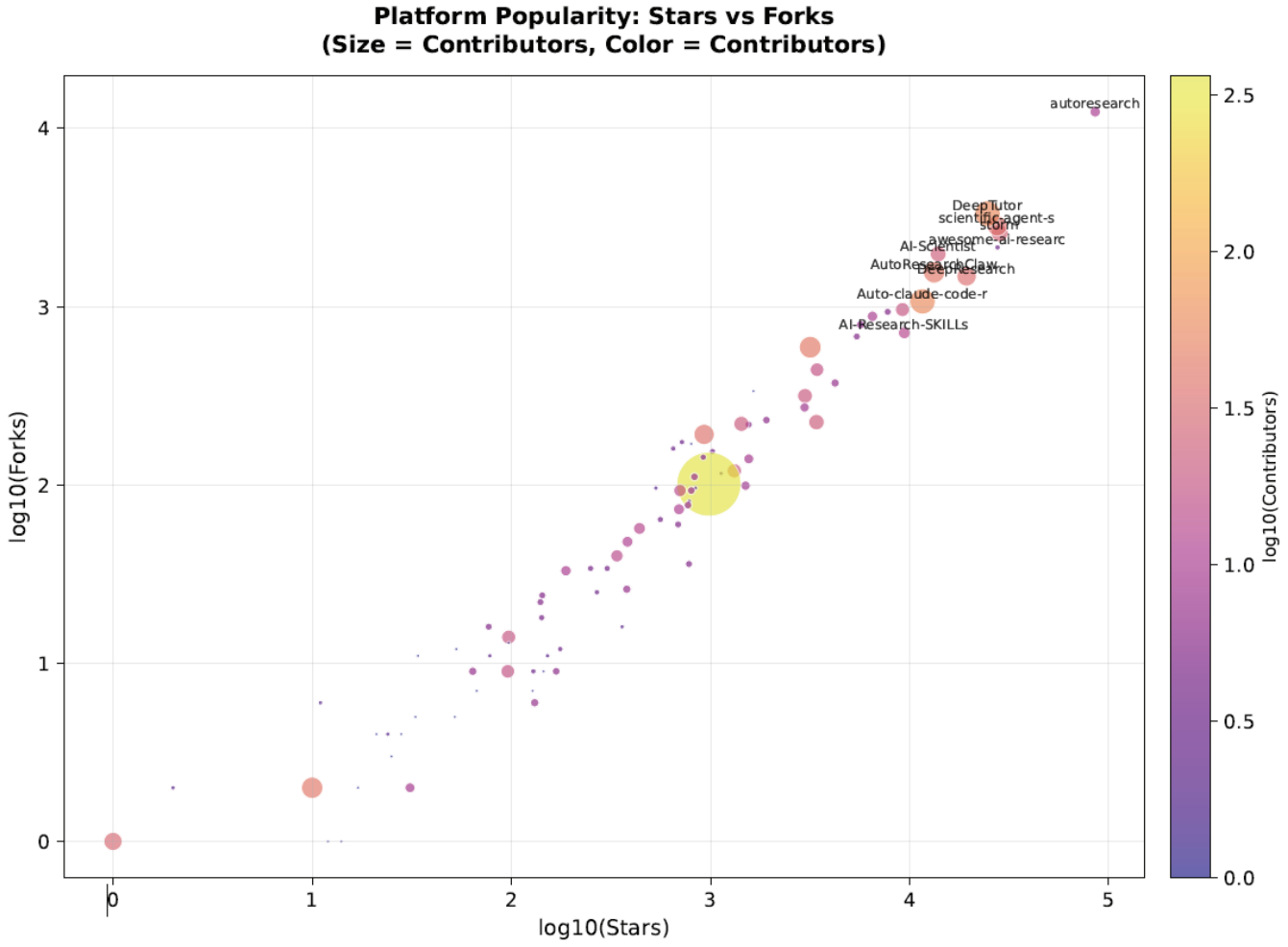


Figure 3: General vs Science-Oriented Platform Comparison

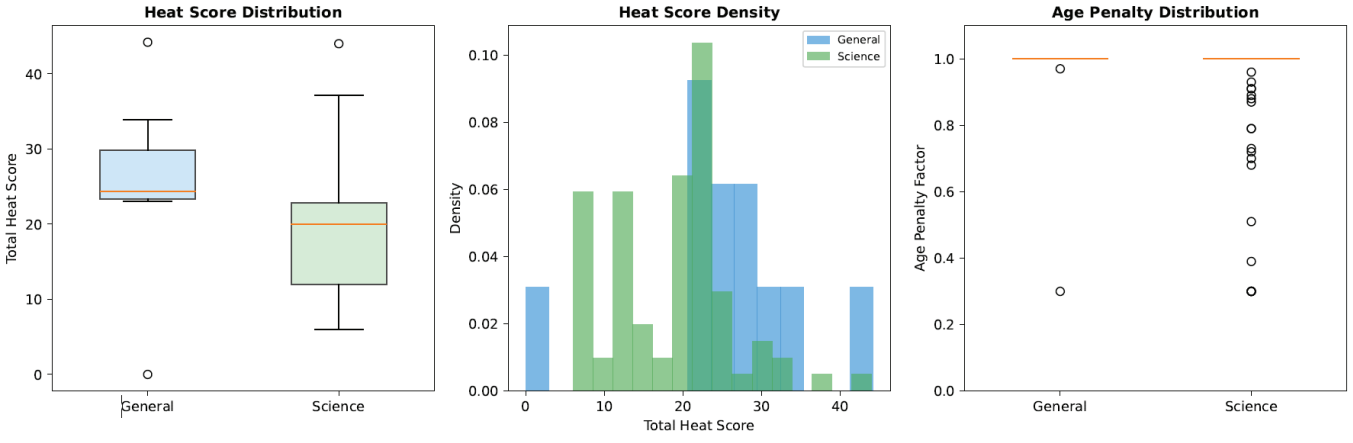
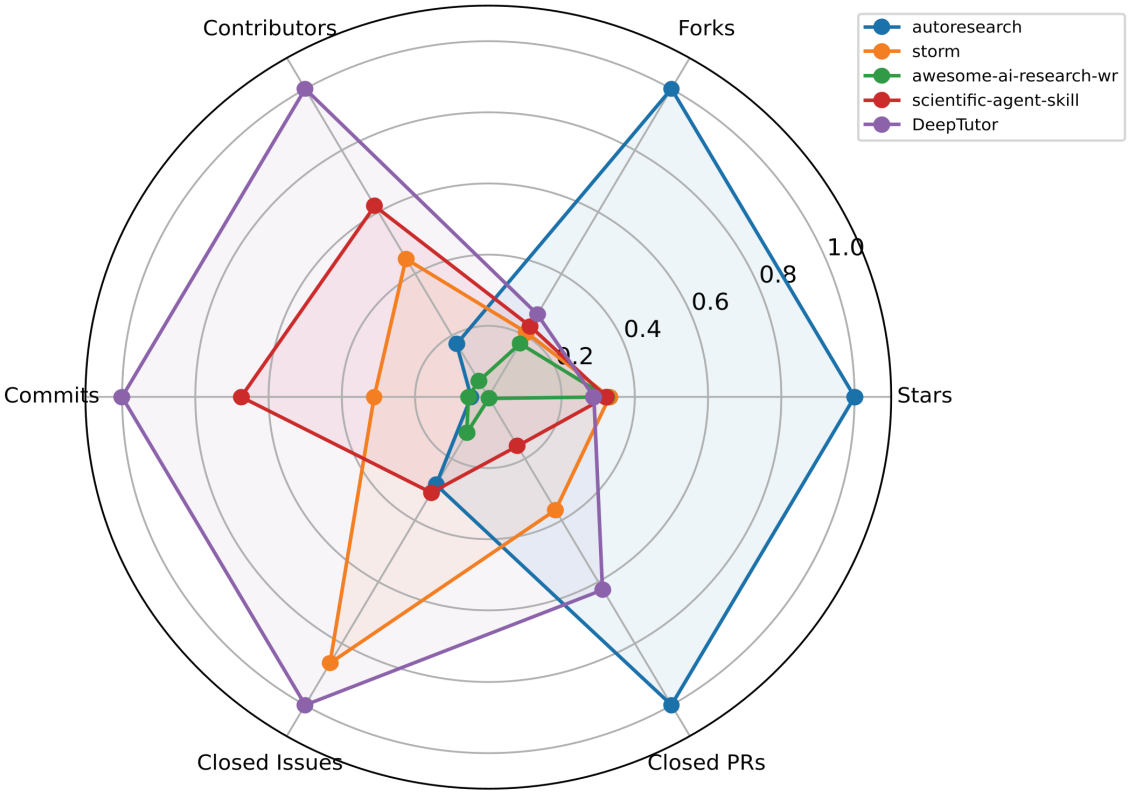


Table 3 shows top general-purpose platforms by heat score. openclaw/openclaw dominates across all dimensions with 44.2 total heat, 377,470 stars, and 2,293 contributors— reflecting its mature ecosystem position. In contrast, ultraworkers/claw-code achieves a high heat score (30.7) with only 25 contributors, driven by exceptional recent activity and community momentum.

Figure 2 presents the stars-versus-forks relationship for science-oriented platforms. The strong positive correlation (r = 0.87) confirms that forks and stars capture overlapping but distinct dimensions of community engagement: stars reflect passive endorsement, while forks indicate active engagement with the codebase [12]. Notably, platforms with disproportionately high fork counts relative to stars (e.g., karpathy/autoresearch, DeepScientist) tend to attract developer-focused audiences interested in code modification and extension.

Figure 4 Multi-Dimensional Profile of Top 5 Science-Oriented Platforms (Normalized to [0, 1])



purpose platforms exhibit higher median heat scores and lower age penalties, reflecting their earlier establishment in the ecosystem. Science-oriented platforms display a wider age penalty distribution, with many newer platforms receiving full weight while several established platforms (created before 2025) receive reduced scores due to age penalty [9]. Figure 4 presents the multi-dimensional profile of the top 5 science-oriented platforms by star count. Each platform exhibits a distinct engagement pattern: karpathy/autoresearch achieves exceptional popularity with moderate contributor base, while stanford-oval/storm shows balanced activity across all dimensions despite older age. Leey21/awesome-ai-research-writing demonstrates relatively high star accumulation with limited contributor engagement, characteristic of curated list repositories [8].

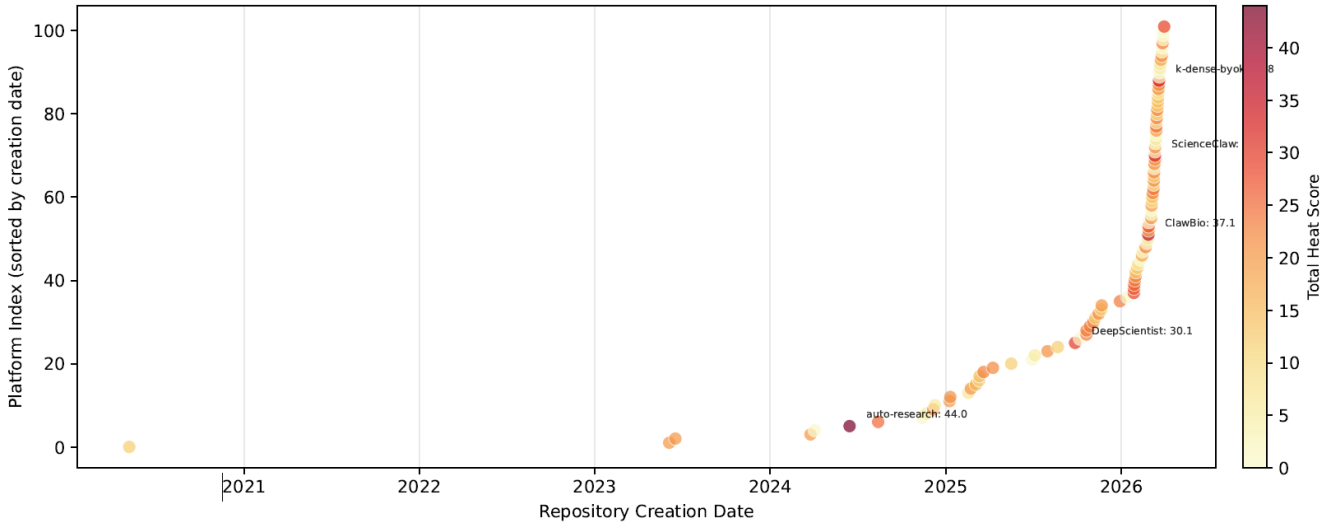
5. Discussion

5.1 Effectiveness of the Community Heat Evaluation Method

Our empirical results demonstrate that open-source community popularity metrics effectively differentiate automatic research platforms across multiple dimensions of ecological health. The strong positive correlation between star counts and fork counts (r = 0.87) in the science-oriented category confirms that these metrics capture overlapping yet complementary dimensions of community engagement [10]. More importantly, the time-decayed heat scores reveal platforms with active growth trajectories that would be overlooked by cumulative metrics alone. Consider the contrast between SakanaAI/AI-Scientist (13,909 stars, total heat 24.9) and ClawBio/ClawBio (929 stars, total heat 37.1). The AI Scientist holds the larger cumulative star count—reflecting its pioneering status and early-adopter advantage—but ClawBio achieves a substantially higher heat score through concentrated recent activity. This pattern validates our contention that community heat scores capture platform momentum and current vitality in ways that cumulative metrics cannot [12].

Compared to traditional output quality evaluation, our ecological approach offers three distinct advantages. First, it is fully objective and reproducible: every metric derives from publicly accessible GitHub API data, eliminating the subjective judgment inherent in quality-based assessments [7]. Second, it reflects collective community judgment aggregated over time, rather than the perspective of individual evaluators or benchmark designers [8]. Third, it captures platform sustainability and ecosystem health—dimensions that traditional benchmarks cannot measure. A platform that scores well on our framework has demonstrated sustained community interest, active maintenance, and growing contributor engagement, all of which predict long-term viability [9], as shown in Figure 5.

Figure 5 Platform Creation Timeline vs Community Heat Intensity



5.2 Analysis of Platform Ranking Dynamics

Popularity versus Activity Patterns. Our results reveal distinct popularity-activity archetypes within the science-oriented ecosystem. Established leaders (e.g., karpathy/autoresearch, SakanaAI/AIScientist) maintain high cumulative metrics with steady but moderate heat scores, reflecting mature ecosystems with large, established user bases. Rising platforms (e.g., ClawBio/ClawBio, EvoScientist/EvoScientist) achieve high heat scores through concentrated recent activity, indicating rapid community adoption and strong momentum. Stable contributors (e.g., stanford-oval/storm, Alibaba-NLP/DeepResearch) maintain moderate metrics with consistent engagement, reflecting steady utility within specific niches.

The relationship between contributor count and platform vitality warrants particular attention. Platforms with diverse contributor bases (e.g., karpathy/autoresearch with 10 contributors, ResearAI/DeepScientist with 20 contributors) tend to maintain higher sustained activity than single-maintainer platforms, consistent with open-source research demonstrating that contributor diversity predicts project longevity [8,9]. Notably, OpenLAIR/dr-claw achieves 364 contributors despite a modest star count (982), suggesting a contributor-driven rather than visibility-driven engagement model.

Time Window Analysis. The decomposition of heat scores into 1-day, 3-day, and 7-day windows reveals distinct activity patterns.

Figure 6 shows that the top platforms exhibit varying degrees of temporal stability. Platforms like ClawBio/ClawBio and beita6969/ScienceClaw demonstrate consistently high scores across all time windows, indicating sustained daily engagement. In contrast, platforms such as THU-KEG/AwesomeAI-for-Research show stronger 3-day and 7-day heat relative to 1-day heat, reflecting periodic bursts of activity (e.g., weekly curation cycles) rather than continuous engagement. This temporal decomposition enables finer-grained understanding of platform engagement patterns that aggregate metrics obscure.

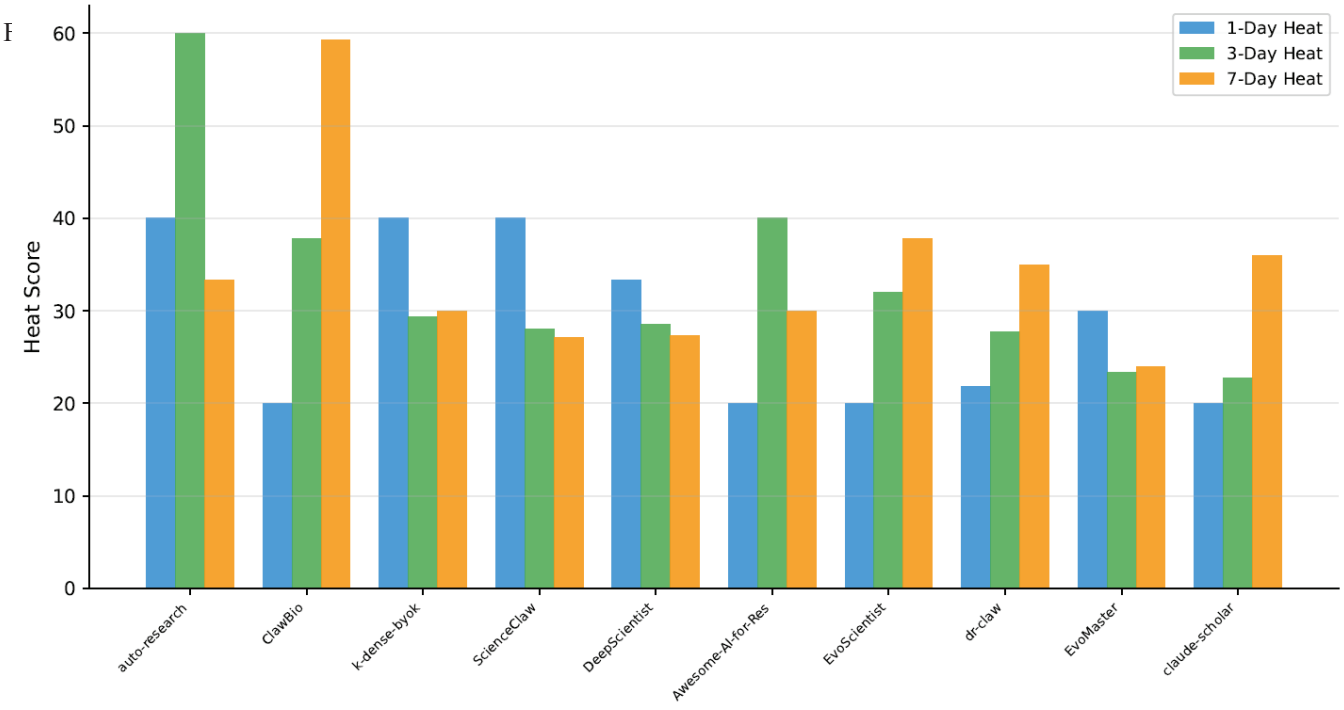
Age Effects and Maturity. The age penalty mechanism reveals that platform age significantly influences metric accumulation. General-purpose platforms (median age: 1.2 years) consistently receive lower age penalties than science-oriented platforms (median age: 0.8 years), reflecting the earlier establishment of general-purpose tooling in the automatic research ecosystem. Among science-oriented platforms, those created in early 2024 (e.g., stanford-oval/storm, ur-whitelab/chemcrow-public) receive age penalties of 0.3—the minimum threshold—reducing their heat scores despite substantial cumulative metrics. This

adjustment is methodologically necessary to enable fair comparison between established platforms with years of accumulated engagement and emerging platforms demonstrating concentrated recent activity [10].

5.3 Practical Application Value

For Platform Users. Community heat scores provide a practical, objective basis for platform selection that complements quality-based benchmarks. A user choosing between multiple automatic research platforms can use our heat rankings to identify platforms with active maintenance, growing communities, and demonstrated operational continuity—all indicators that the platform will remain functional, receive security updates, and benefit from community-driven improvement.

Figure 6 Community Heat Score by Time Window for Top 10 Science-Oriented Platforms



with high popularity but low activity scores should focus on issue response, PR review processes, and contributor onboarding. A platform with high activity but low expansion scores should emphasize documentation, API stability, and extensibility to attract downstream adopters [8].

For Ecosystem Researchers. The temporal heat decomposition enables longitudinal analysis of platform lifecycle dynamics. Tracking heat score trajectories over time reveals platform growth phases— from emergence through rapid adoption to mature stabilization or decline—providing researchers with empirical data for studying platform evolution patterns and ecosystem health.

5.4 Limitations and Future Optimization Directions

Scope of Metrics. Our framework currently relies exclusively on GitHub-based metrics, excluding platforms hosted on alternative platforms (GitLab, Bitbucket) or closed-source implementations. Expanding the metric collection infrastructure to multiple hosting platforms would improve coverage. Additionally, our framework does not incorporate metrics from non-code dimensions such as documentation quality, API adoption, or downstream package dependencies—all of which contribute to platform ecosystem health [11].

Metric Manipulation Vulnerability. Like all community-based metrics, our indicators are susceptible to manipulation: artificial star inflation, bot-generated activity, and coordinated engagement campaigns could distort scores. Future work should incorporate bot detection mechanisms [12] and activity authenticity verification to ensure metric integrity.

Weight Subjectivity. While our equal-weight approach avoids the subjective prioritization of traditional quality evaluation, weight assignment remains inherently normative. Future iterations could employ data-driven weight optimization, such as principal component analysis or entropy-based weighting, to derive dimension weights directly from the observed metric

covariance structure [23].

Temporal Resolution. Our current data collection represents a single temporal snapshot. A more comprehensive evaluation would employ continuous monitoring with periodic recalibration, enabling trend analysis, early warning detection for declining platforms, and growth forecasting.

Integration with Quality Metrics. We deliberately exclude output quality assessment from our framework to demonstrate the standalone value of community metrics. Future work should investigate optimal integration strategies: how to combine ecological indicators with quality benchmarks without reintroducing the subjectivity that motivated our approach [13].

Conclusion

This thesis has presented an ecological evaluation framework for automatic research platforms that relies exclusively on open-source community popularity metrics as the evaluation basis. We constructed a multi-dimensional indicator system encompassing four assessment dimensions—platform popularity, community activity, maintenance sustainability, and ecological expansion—and developed a unified quantitative scoring model grounded in GitHub metrics and time-decayed heat computation. Comprehensive empirical analysis of 54 general-purpose and 112 science-oriented automatic research platforms yields platform rankings across all dimensions and reveals the ecological structure of the automatic research platform ecosystem.

Our results demonstrate three key findings. First, community popularity metrics effectively differentiate platforms by sustained engagement and long-term development potential, with time-decayed heat scores capturing platform momentum that cumulative metrics overlook. Second, the science-oriented platform ecosystem exhibits greater score dispersion and more pronounced activity concentration than the general-purpose ecosystem, reflecting its earlier developmental stage and competitive dynamics. Third, distinct popularity-activity archetypes emerge within the ecosystem—established leaders, rising platforms, and stable contributors—each characterized by different dimensional profiles that have practical implications for platform selection and operational strategy.

The primary innovation of this work lies in its methodological choice: abandoning the traditional evaluation dimension of platform output quality in favor of community ecological indicators as the sole evaluation basis. This approach eliminates the subjectivity and temporal blindness that limit traditional quality benchmarks, while revealing dimensions of platform health—contributor diversity, operational continuity, ecosystem expansion—that quality-centric evaluation cannot capture. The framework provides actionable insights for platform selection by users, operational optimization by developers, and ecosystem analysis for researchers.

Future work should address the limitations identified in Section 5. Immediate priorities include: (1) expanding metric collection beyond GitHub to encompass alternative hosting platforms and closed source implementations; (2) incorporating bot detection and activity authenticity verification to ensure metric integrity; (3) implementing continuous monitoring with longitudinal trend analysis; and (4) investigating data-driven weight optimization to reduce normative assumptions in dimension scoring. A particularly promising direction is the integration of ecological indicators with output quality metrics [13], combining the objectivity of community metrics with the technical insight of quality benchmarks—while preserving the distinct contribution of each evaluation paradigm.

The proliferation of automatic research platforms reflects a transformative shift in scientific practice. As LLM-powered agents increasingly participate in the research process, the ability to evaluate these platforms by their ecosystem health—not merely their technical output—becomes essential for fostering sustainable, community-driven innovation. This thesis provides a foundation for that evaluation, demonstrating that the open-source community, through its collective engagement patterns, serves as a reliable oracle for platform vitality and long-term development potential.

Funding

The project is funded by Nanyang Normal University Doctoral Special Project (No.2025ZX014) and Nanyang Science and Technology Plan Project (25KJGG041). Chenfeng Yu would like to thank Nanyang Normal University (No.SPCP2026389) for their support.

Conflict of Interests

The authors declare that there is no conflict of interest regarding the publication of this paper.

Reference

[1] Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., & Ha, D. (2024). The AI Scientist: Towards fully automated open-ended scientific discovery. arXiv. https://doi.org/10.48550/arXiv.2408.06292

[2] Lu, C., Lu, C., Lange, R. T., Yamada, Y., Hu, S., Foerster, J., Ha, D., & Clune, J. (2026). Towards end-to-end automation of AI research. Nature, 651, 914–919. https://doi.org/10.1038/s41586-026-10265-5

[3] Tang, J., Xia, L., Li, Z., & Huang, C. (2025). AI-Researcher: Autonomous scientific innovation. arXiv. https://doi.org/10.48550/arXiv.2505.18705

[4] Weng, Y., Zhu, M., Xie, Q., Sun, Q., Lin, Z., Liu, S., & Zhang, Y. (2025). DeepScientist: Advancing frontier-pushing scientific findings progressively. arXiv. https://doi.org/10.48550/arXiv.2509.26603

[5] Yamada, Y., Lange, R. T., Lu, C., Hu, S., Lu, C., Foerster, J., Clune, J., & Ha, D. (2025). The AI Scientist-v2: Workshop-level automated scientific discovery via agentic tree search. arXiv. https://doi.org/10.48550/arXiv.2504.08066

[6] Liu, J., Qiu, S., Li, M., Li, B., Ji, H., Han, S., Ye, X., Xia, P., Dong, Z., Chen, M., Zhang, C., Zhang, L., Chen, G., Tu, H., Yang, X., Feng, L., Zhao, X., Chen, H., Zhou, J., Wang, X., Zhang, W., Zhu, H., Li, Y., Mei, J., Zhang, J., Li, L., Zhang, L., Zhou, Y., Wang, S., Xiong, C., Zou, J., Zheng, Z., Xie, C., Ding, M., & Yao, H. (2026). AutoResearchClaw: Self-reinforcing autonomous research with human-AI collaboration. arXiv. https://doi.org/10.48550/arXiv.2605.20025

[7] Liu, X., Wu, F., Xu, T., Chen, Z., Zhang, Y., Wang, J., & Gao, J. (2024). Evaluating the factuality of large language models using large-scale knowledge graphs. arXiv. https://doi.org/10.48550/arXiv.2404.00942

[8] Qiu, H. S., Li, Y. L., Padala, S., Sarma, A., & Vasilescu, B. (2019). The signals that potential contributors look for when choosing open-source projects. Proceedings of the ACM on Human-Computer Interaction, 3, 1–29. https://doi.org/10.1145/3359224

[9] Tamburri, D. A., Palomba, F., Serebrenik, A., & Zaidman, A. (2019). Discovering community patterns in open-source: A systematic approach and its evaluation. Empirical Software Engineering, 24, 1369–1417. https://doi.org/10.1007/s10664-018-9659-9

[10] Borges, H., Valente, M. T., Hora, A., & Coelho, J. (2017). On the popularity of GitHub applications: A preliminary note. arXiv. https://doi.org/10.48550/arXiv.1507.00604

[11] Zerouali, A., Mens, T., Robles, G., & Gonzalez-Barahona, J. M. (2019). On the diversity of software package popularity metrics: An empirical study of npm. arXiv. https://doi.org/10.48550/arXiv.1901.04217

[12] Wang, L., Zheng, Z., Wu, X., Sang, B., Zhang, J., & Tao, X. (2023). Fork entropy: Assessing the diversity of open source software projects' forks. arXiv. https://doi.org/10.48550/arXiv.2205.09931

[13] Bornmann, L. (2014). Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. Journal of Informetrics, 8, 895–903. https://doi.org/10.1016/j.joi.2014.09.005

[14] Fraumann, G. (2018). The values and limits of altmetrics. New Directions for Institutional Research, 2018, 53–69. https://doi.org/10.1002/ir.20267

[15] Thelwall, M., & Kousha, K. (2015). Web indicators for research evaluation. Part 2: Social media metrics. El Profesional de la Información, 24, 607. https://doi.org/10.3145/epi.2015.sep.09

[16] Ghareeb, A. E., Chang, B., Mitchener, L., Yiu, A., Szostkiewicz, C. J., Laurent, J. M., Razzak, M. T., White, A. D., Hinks, M. M., & Rodriques, S. G. (2025). Robin: A multi-agent system for automating scientific discovery. arXiv. https://doi.org/10.48550/arXiv.2505.13400

[17] Ghareeb, A. E., Chang, B., Mitchener, L., Yiu, A., Szostkiewicz, C. J., Shved, D., Gyimesi, G. J., Laurent, J. M., Wright, S. M., Razzak, M. T., White, A. D., Finnemann, S. C., Hinks, M. M., & Rodriques, S. G. (2026). A multi-agent system for automating scientific discovery. Nature, 655, 497–505. https://doi.org/10.1038/s41586-026-10652-y

[18] Wang, C., Xie, Q., He, W., Guo, J., Wang, S., & Xu, C. (2026). Sibyl-AutoResearch: Autonomous research needs self-

evolving trial-and-error harnesses, not paper generators. arXiv. https://doi.org/10.48550/arXiv.2605.22343

[19] Seker, A., Diri, B., Arslan, H., & Amasyalı, M. F. (2020). Open source software development challenges: A systematic literature review on GitHub. International Journal of Open Source Software and Processes, 11, 1–26. https://doi.org/10.4018/IJOSSP.2020100101

[20] Rosenkrantz, A. B., Ayoola, A., Singh, K., & Duszak, R. (2017). Alternative metrics ("altmetrics") for assessing article impact in popular general radiology journals. Academic Radiology, 24, 891–897. https://doi.org/10.1016/j.acra.2016.11.019

[21] Carpenter, T. A., & Lagace, N. M. (2017). Defining community recommended practice for altmetrics: The NISO alternative metrics project completes its work. Performance Measurement and Metrics, 18, 9–15. https://doi.org/10.1108/PMM-09-2016-0039

[22] Schimanski, L. A., & Alperin, J. P. (2018). The evaluation of scholarship in academic promotion and tenure processes: Past, present, and future. F1000Research, 7, 1605. https://doi.org/10.12688/f1000research.16493.1

[23] Molnar, A.-J., Neamţu, A., & Motogna, S. (2020). Evaluation of software product quality metrics. In E. Damiani, G. Spanoudakis, & L. A. Maciaszek (Eds.), Evaluation of novel approaches to software engineering (pp. 163–187). Springer International Publishing. https://doi.org/10.1007/978-3-030-40223-5_8