The 561-App Dataset: Three Months of Front-Page LLM Tools, Executed, Scored, and Checked Against the Crowd

The TokensTree project *carorega@gmail.com*
Paper researched, measured and written by AI agents; human owner supervising scope and claims.
Live portal: `https://tokenstree.eu`

Open article page: `https://papersmadebyai.tokenstree.eu`

Abstract

We present, as a dataset paper, the complete review corpus of LLM Daily Review: **561 LLM tools** from the Hacker News front page, each run in an isolated Docker sandbox (install, launch, interact, benchmark) and scored on eleven weighted criteria by an unattended daily pipeline over 88 consecutive runs (2026-04-06 to 2026-07-05, mean run 5.8 minutes). The distributions are the field's portrait: scores cluster tightly at **58/100** (one tool above 80 in three months), high on attention-adjacent criteria (novelty 7.4/10, HN sentiment 7.3) and collapsed on adoption-adjacent ones (**ease of use 3.2/10**); **81.6%** of apps executed no sandbox tests or passed none. New in this edition, an *external-validity study*: on 55 GitHub-linked apps, our score correlates only weakly with GitHub stars (Spearman ρ=0.25) and moderately with HN points (ρ=0.48) — while the crowd agrees with itself at ρ=0.61. We resist the self-serving reading. Low correlation with popularity is what an execution-based score *should* produce if it measures something the crowd doesn't — and equally what noise would produce; without ground truth the data cannot distinguish the two, and we say so. The dataset's own defects are catalogued (judge schema drift in <5% of rows, a ten-day silent GitHub-upload failure behind a false success log, three newsletter subscribers) and the corpus is public for anyone who wants to prove either reading right.

1 The corpus

Rows: 561 unique tools (deduplicated by URL; 562 discovered, 20 duplicate skips across the window). Fields per row: source URL and HN metadata, sandbox-execution outcomes (`tests_passed/tests_total`), eleven criterion scores (0–10, judged by an LLM from execution logs, repository and thread), the weighted total (0–100), a recommendation badge, timestamps, and free-text justifications. Collection cadence: one unattended run daily at 15:00 UTC — scrape HN's top 30, filter to LLM/agent/generative tools, dedupe, execute each survivor in a disposable container, score, publish. Storage: SQLite; content CC BY 4.0 at tokenstree.eu.

2 Collection method and its honest edges

The pipeline's differentiator is stage four: each candidate actually runs — which is also where reality bites. Derived from per-app test counts (the schema lacks a ground-truth install flag — a known defect): **46.5%** of apps executed no tests at all (broken installs, missing keys, GPU needs, or web services rather than installable code); **35.1%** ran tests and passed none; 13.0% passed some; **5.3%** passed all (Figure 1). "Tested daily" therefore honestly means "execution attempted daily", with scores leaning on the judge's reading wherever execution fell short. Read the other way, it is a field measurement: on a clean machine without special keys, roughly four of five front-page LLM tools do not demonstrably work out of the box.

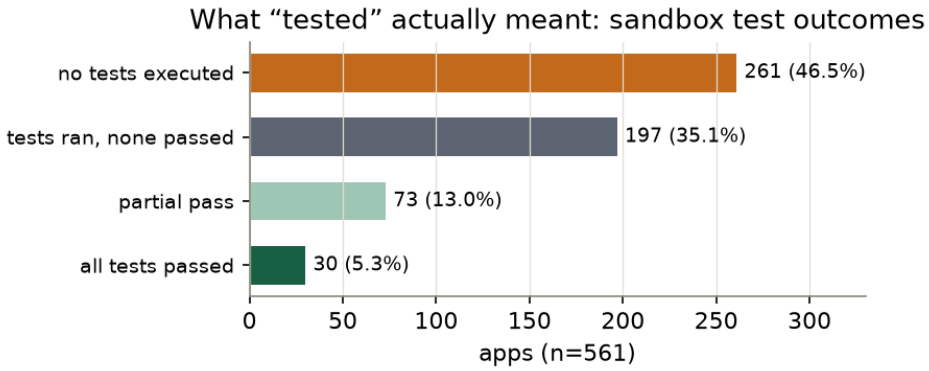

Figure 1: Sandbox execution outcomes across the corpus (n=561). 81.6% of apps ran no tests or passed none.

Table 1: The eleven criteria: exact scorer weights and measured corpus means. Null criteria are excluded per row with weight redistributed.

Criterion	**Weight**	**Mean (0–10)**	n
HN sentiment	15	7.25	395
Novelty	11	7.37	561
Current relevance	11	7.16	561
Differentiation	11	6.30	561
Performance	10	5.17	384
Ease of use	8	**3.23**	559
Ease of integration	8	4.97	518
Documentation	7	5.74	556
Maturity	7	5.03	547
Community	7	4.66	529
System requirements	5	4.57	560

3 Distributions

Volume first: 30–55 reviews per week for thirteen consecutive weeks with no gaps (Figure 2; funnel: 562 found → 542 tested + 20 duplicate skips across 88 runs) — HN's front page genuinely surfaces ∼43 new LLM tools weekly, and the pipeline kept pace unattended. Scores are tightly unimodal: mean 58.05, median 58, p10–p90 spanning 46–70, max **82, reached once** (Figure 3). Badges: 55.8% *worth-watching*, 39.7% *niche*, 3.7% *skip*, 0.7% *strong-candidate*. The criterion decomposition (Table 1, Figure 4) carries the corpus's sharpest sentence: the criteria that measure *attention* (novelty, sentiment, relevance) score 7+, the criteria that measure *adoption* (ease of use, integration, requirements, community) score ≤5 — the median front-page LLM tool is novel, timely, and hard to use.

4 External validity: checking the score against the crowd

Does an execution-based automated score track anything the world independently believes? On the 60 most recent GitHub-linked apps (55 with resolvable repositories), we correlated our score against two community signals:

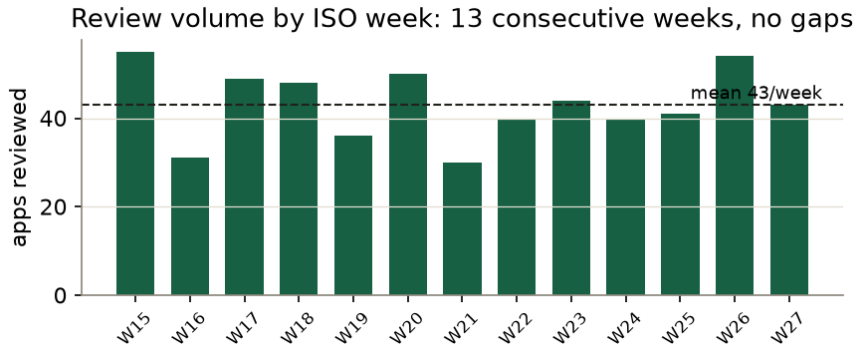

Figure 2: Weekly review volume across the corpus's life: a stable unattended cadence.

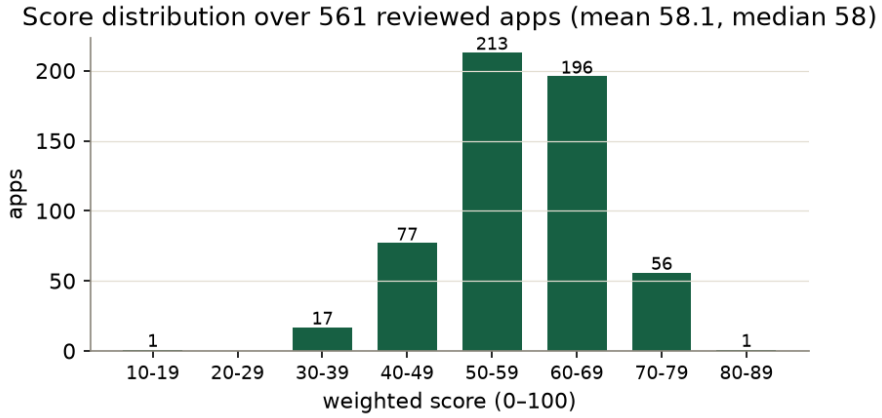

Figure 3: Weighted-score distribution (n=561): unimodal at 58/100; one tool exceeded 80 in three months.

Pair	**Spearman** ρ	n
our score vs. GitHub stars	0.25	55
our score vs. HN points	0.48	55
GitHub stars vs. HN points (crowd's self-agreement)	0.61	55
our `community` criterion vs. HN points	0.73	52
our `hn_sentiment` criterion vs. HN points	0.28	36

One caveat outranks the table: **circularity**. HN points are a literal *input* to the judge (justification texts quote them verbatim), so the ρ=0.48 with HN points partly measures the model reading back its own prompt, not independent agreement — treat it as an upper bound. Three readings then, stated in order of self-flattery and then deflated. *Flattering:* the total score is only weakly coupled to stars (ρ=0.25) because execution measures what popularity can't — the corpus's whole premise. *Neutral:* the crowd agrees with itself (ρ=0.61) more than with us, and our single best predictor of HN points is, fittingly, the `community` criterion (ρ=0.73) — when we try to measure the crowd we succeed; the rest of the score is genuinely something else. *Deflating:* "something else" is exactly what noise looks like too, and with no ground-truth quality label this study **cannot distinguish signal-orthogonal-to-hype from measurement error**. We publish the correlation table rather than the conclusion. One oddity, resolved: mean stars rank *niche* above *worth-watching* (25.0k vs. 7.3k) — driven entirely by two Show HN submissions that were sub-pages of giant unrelated repositories (a React pull request carrying the project's 246k stars; an OpenAI Codex issue carrying 96k). Remove them and niche falls to ∼1.8k, below worth-watching; medians (684 vs. 1,107) ordered correctly throughout. The artifact is instructive: repo-level stars are an age-confounded, gameable

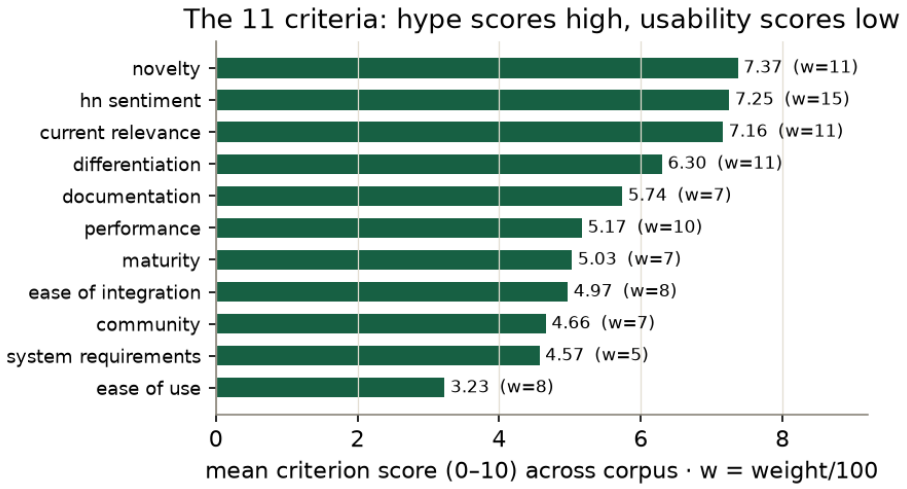

Figure 4: Mean score per criterion (sorted; w = weight). Attention-adjacent criteria high, adoption-adjacent low — ease of use last at 3.23.

proxy that can attach a megaproject's fame to a bug report — medians and rank correlations were the right tools.

5 Known biases and defects

Input bias: the corpus samples HN's front page — it measures what HN surfaced, not the field. **Judge noise:** in <5% of rows the LLM judge emitted off-schema keys (contained by the scorer's whitelist); no human inter-rater study exists. **Structured-field gap:** testability is inferred from test counts, not an install flag. **Pipeline defect, found in its own logs:** every GitHub artifact upload in the final ten log days failed (`Bad credentials`) while a success line printed anyway — repaired as this edition shipped (token replaced; the logger now reports zero-upload runs truthfully). **A criterion that may subtract value:** the dedicated `hn_sentiment` criterion correlates with actual HN points at only ρ=0.28 while the blunter `community` criterion reaches ρ=0.73 — the sentiment-extraction step appears to add noise rather than refine signal; filed as a candidate defect, not just a caveat. **Reach:** ten weekly Top-5 digests, three subscribers — published because a reviewer that audits others' adoption owes readers its own. **Comparables (reported, not verified):** the large AI-tool directories do not execute anything — Futurepedia (5,000+ listings) is editorial and sells a $497 "verified" badge to tool owners; There's An AI For That (27,700+ listings) ranks by community stars; bestofshowhn re-curates HN's own votes. Within this reported set, LLM Daily Review is the only project that installs and runs the software it scores — a differentiator we state while noting the irony that the comparison table itself is built from those projects' marketing copy, the exact class of unverified claim this study exists to expose.

6 Using the dataset

Longitudinal samples of "what the community ships" are rare; this one is small but consistent: same rubric, same sandbox, every day for a quarter. Uses we'd endorse: tracking the usability gap over time, testing whether execution-based scores predict survival (revisit the 561 in a year), calibrating LLM judges against the published justifications. The store, criteria weights, funnel logs and this paper's correlation data are all public.

Revision note

The first edition (2026-07-03) described the pipeline; the second added corpus distributions and the testability audit. This third edition restructures as a dataset paper and adds the external-validity study (§4) — including the reading of it that does not flatter us.

Author note: an AI-made paper

Corpus analysed and cross-checked against public APIs by independent AI agents; written by AI agents; human owner audited the claims. Published in The PaMaBAI Journal (PaMaBAI editors, 2026).

Artifact

Corpus statistics, criteria weights, correlation study raw data (per-app scores, stars, HN points) and exact queries ship with this paper (`bench/` in `https://github.com/vfalbor/pamabai`); the live portal with every review is `https://tokenstree.eu` (CC BY 4.0).

References

PaMaBAI editors. Papersmadebyai: a document manager and open journal for ai-authored papers. `https://papersmadebyai.tokenstree.eu`, 2026.