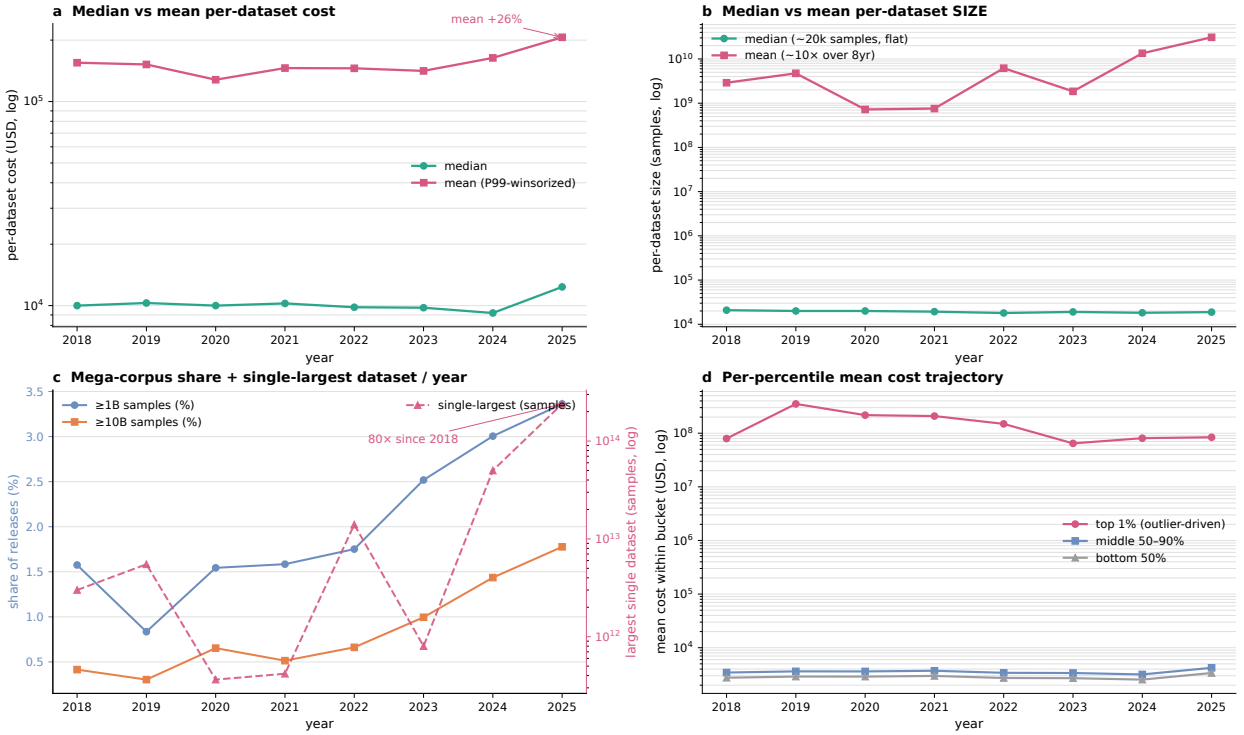
Supplementary Information for "Cost Does Not Buy Impact: A 173K-Dataset Audit of the AI Data Economy"

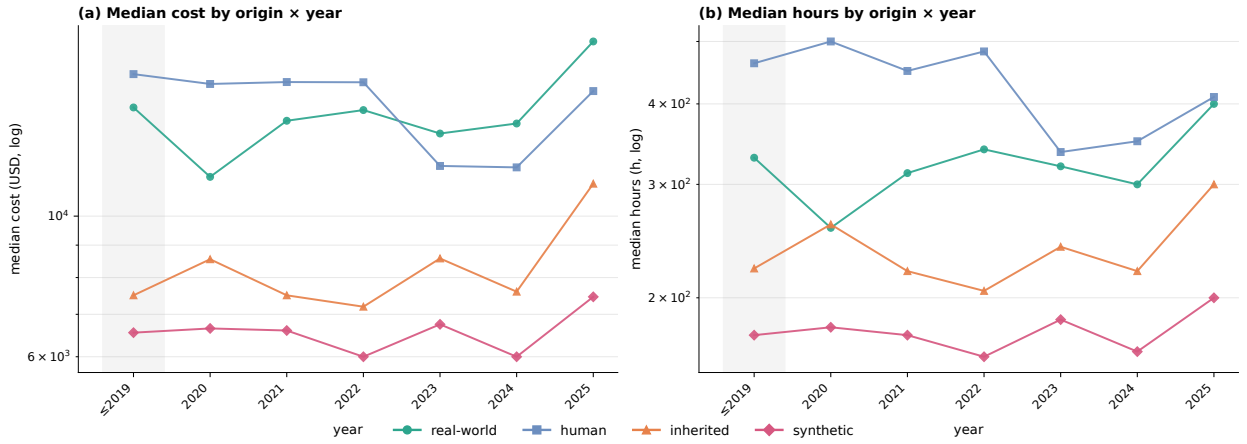
S1 Supplementary figures and tables

S1.1 Volume-growth decomposition (RQ1.1 supplementary)

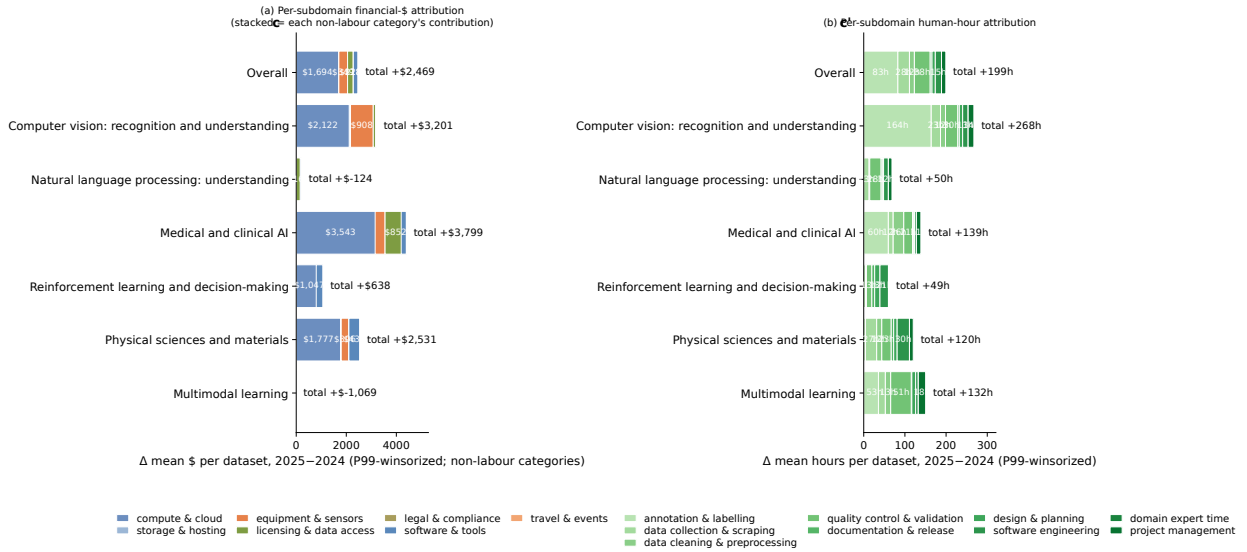
The volume axis (Fig. S1, Table S21) shows the median release essentially flat 2018–2025 while the mean grows roughly an order of magnitude entirely from the upper tail. We do not treat size as primary evidence: `num_instances` is too sparsely reported in ChatPD and our extraction to support a population claim, so only the order-of-magnitude tail story survives.



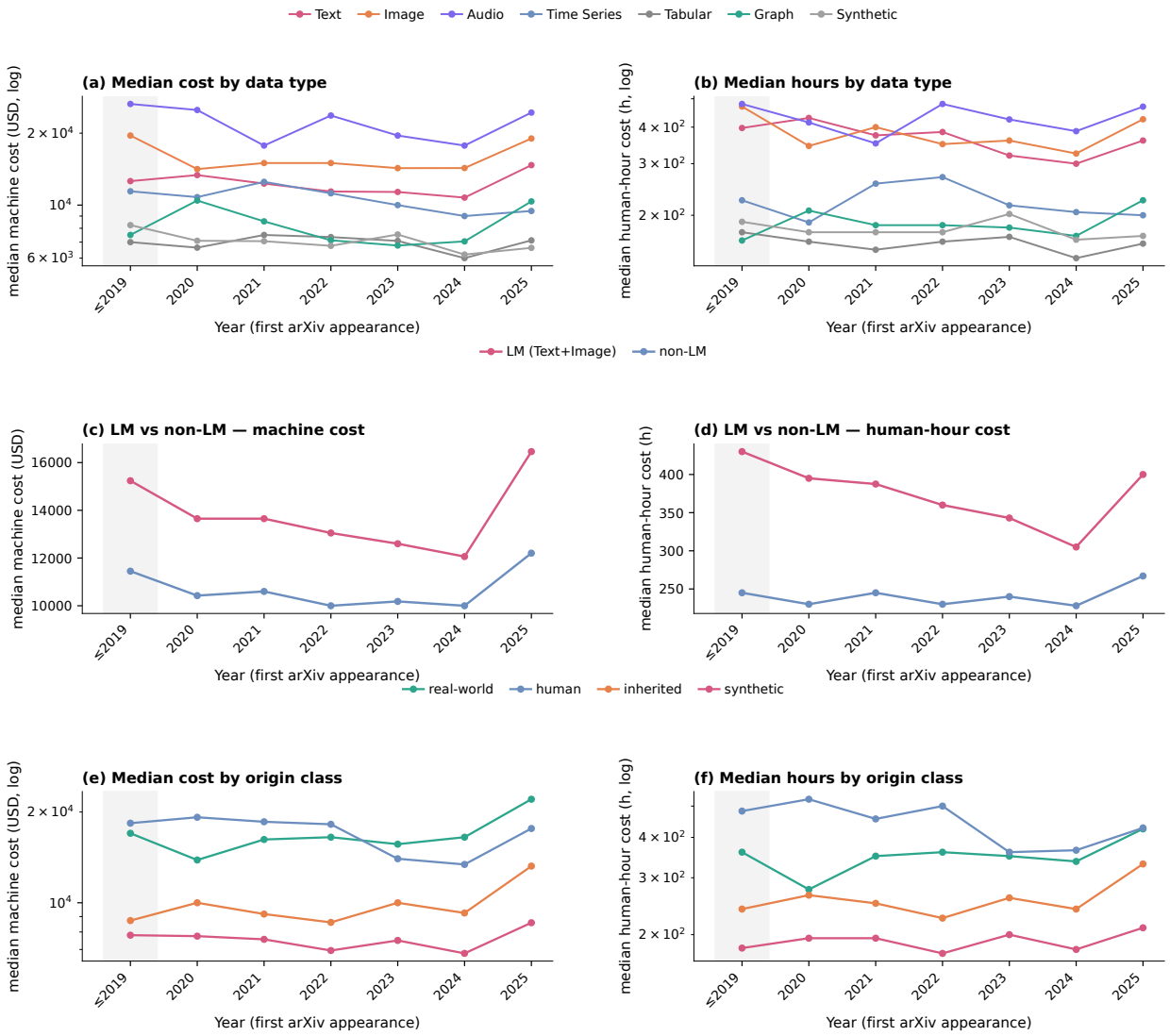
Supplementary Figure S1: **Decomposing 2018→2025 investment growth: median is flat, mean is in the tail, and the volume story is bimodal. (a) Median vs P99-winsorized mean per-dataset cost.** Both are flat 2018–2024 and jump +30% in 2025 only. The cost increase in 2025 is broad-based rather than concentrated in mega-projects; the per-percentile decomposition in panel (d) makes this explicit by showing the middle and bottom segments both rising ∼+24% while the top-1% segment is outlier-driven and chaotic. **(b) Median vs mean per-dataset size (samples).** Median is flat at ∼20K samples for 8 years; mean grows ∼ 10× (2.98B→31.6B). The mean trajectory is the upper-tail-only signal that median cannot see. **(c) Mega-corpus share + single-largest annual dataset.** The ≥1B-sample share doubles (1.6%→3.4%), the ≥10B share grows ∼6×, and the single largest released dataset of the year grows ∼80× between 2018 and 2025 (3T→240T samples): the frontier-corpus arms race written on the y-axis. **(d) Per-percentile mean cost trajectory.** Top 1% mean is outlier-driven and chaotic (no trend); middle 50–90th and bottom 50% both rise steadily ∼+24% over 2018–2025; the mean-cost trend is carried by the broad distribution rather than the upper tail. **Implication**: total-investment growth has two independent components, count 18× (democratization: many more typical-cost typical-size datasets) and tail size 10× (frontier-corpus inflation), with $18 \times 10 \approx 190\times$ total-sample growth as their product.



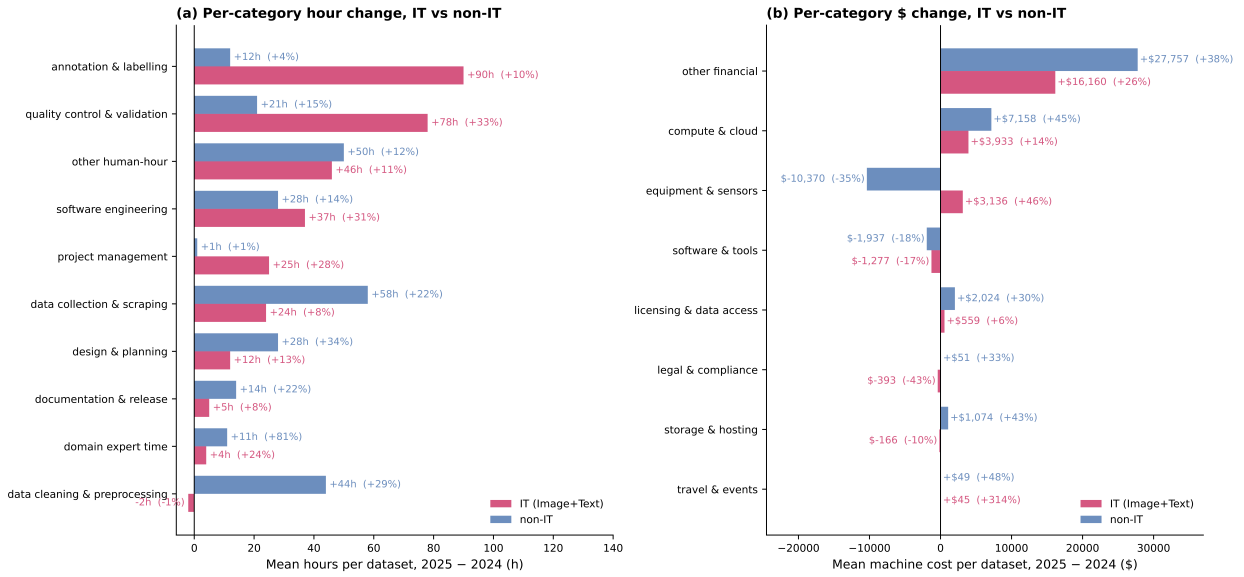
Supplementary Figure S2: **Within-origin machine cost and human-hour cost trajectory, 2010–2025.** Pre-2020 cohorts are pooled into a single "≤2019" anchor on the left. **(a)** Median machine cost (USD, log) by origin × year. **(b)** Median human-hour cost (h, log) by origin × year. The 2022→2025 within-origin deltas: **inherited rises fastest** (+56% cost / +46%h) as the 2025 inherited pool is increasingly populated by derivatives of frontier pretraining corpora; **real-world** rebounds (+28% / +18%h) after a deep 2024 trough; **synthetic** rises modestly (+24% / +23%h) but remains the cheapest origin in absolute terms; **human**-origin is the only class with *flat* cost (−3%) and *declining* hours (−15%), the candidate LLM labor-substitution signature on the origin axis.



Supplementary Figure S3: **2024 → 2025 cost-rebound attribution by research domain (top-6).** Whole-sample overall +$2,469/dataset. Medical and clinical AI (+$3,799/dataset) is equipment-dominant; CV-recognition (+$3,201) and physical sciences (+$2,531) are compute-dominant; NLP-understanding (−$124) and multimodal learning (−$1,069) *decline* on the cost axis, diluted by an inflow of cheap derivative and synthetic releases. The "+35% rebound" in §2.1 (1) is therefore a whole-sample statement that disguises sharp research-domain heterogeneity.



Supplementary Figure S4: **Per-data-type and per-origin decomposition of the per-dataset cost trajectory.** Companion to Fig. 1**a**. The trajectory is split across seven `data_type` classes, an LM-vs-non-LM aggregate view, and four origin classes (real-world, human, inherited, synthetic). The 2025 rebound is concentrated in Text and Image data types and in the human and real-world origin classes; `data_type=Synthetic` rebounds least, consistent with synthetic-data construction being largely LLM-driven rather than human-labor-driven.

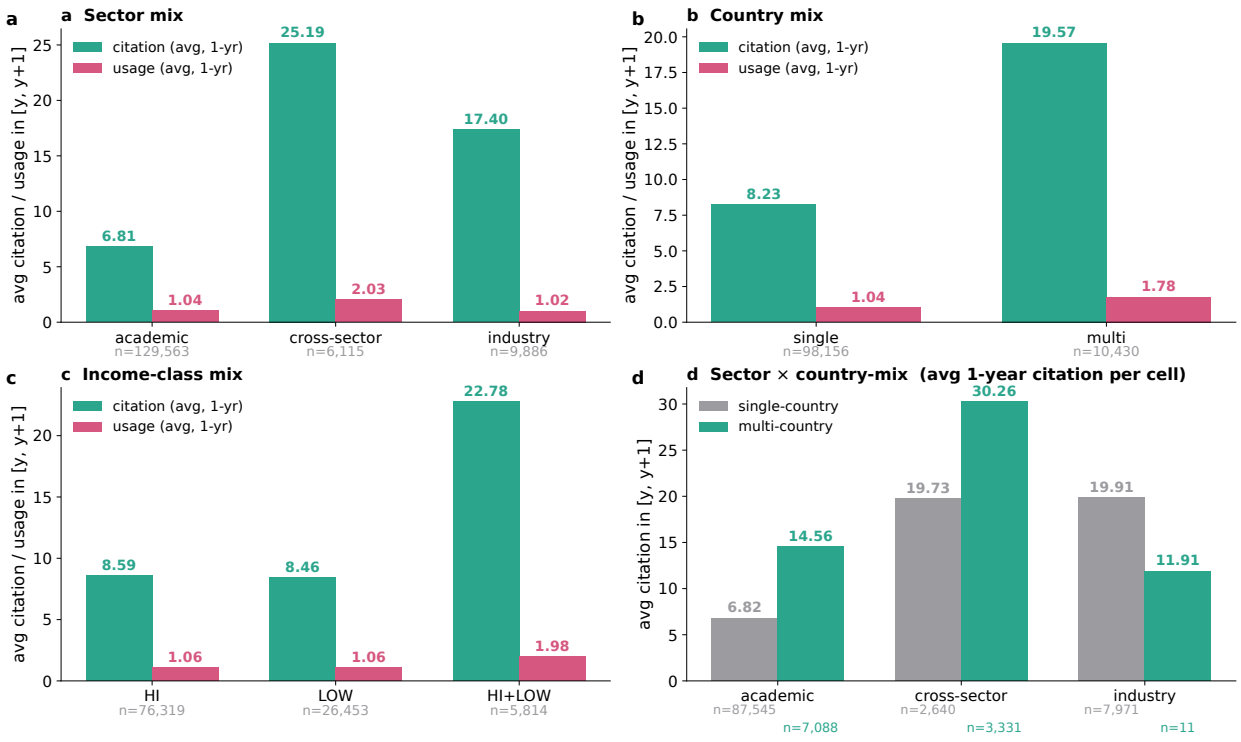


Supplementary Figure S5: **Per-category labor decomposition for IT (Image+Text) versus non-IT modalities, 2024→2025.** The same human-hour categories as Fig. 1**d** (annotation, QC, data collection, software engineering, data cleaning, project management) split into IT and non-IT panels. The headline +76% combined annotation + QC + collection share of the hour increment is concentrated in the IT subset; non-IT categories contribute a smaller and more diffuse share, consistent with the LM-data 2025 rebound being predominantly an annotation/curation event for language and vision models rather than a uniform labor-input rise.

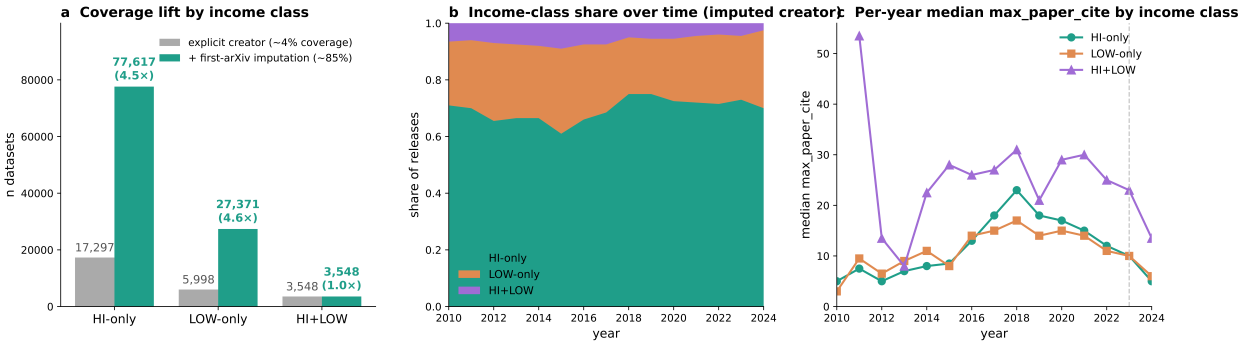
S1.2 Regression diagnostics: robustness for the cost null

Choice of headline impact metric. We adopt `max_cited_paper.citationCount` (henceforth "`max_paper_cite`"), the citation count of the *most-cited downstream paper that uses the dataset*, as the headline impact metric, for three reasons. **(i)** It captures the *ceiling* of impact a dataset enabled, i.e. the most visible piece of downstream research it made possible, which is what readers usually have in mind when they ask "did this dataset matter?". **(ii)** The cross-class ordering it produces (industry > mixed > academic) is the most pronounced and is robust across alternate aggregations of the same field (winsorized mean and geometric mean of `max_paper_cite`) and across alternate outcome families (mean and median `cite_1yr`, mean `use_1yr`); see Fig. S11 below, where every metric preserves the same sector ordering. We move the median/mean `cite_1yr` / `use_1yr` trajectories to that figure to avoid clutter in the main panels. **(iii)** `max_paper_cite` is *not* synchronized with usage (`use_1yr`): a dataset can be widely *adopted* without producing a single high-citation paper, and a dataset can spawn one viral paper without broad reuse. Reporting `max_paper_cite` alongside `use_1yr` therefore gives us two complementary views of impact rather than two views of the same thing.

Why citation and usage as the impact proxies. We rely on these two proxies for two reasons. First, the ideal outcome, the realized economic value of the models a dataset enabled, is unobservable at corpus scale: no dataset-level revenue or profit reporting exists for a 155K-dataset population, and even firm-level attribution of model revenue to individual training sets is an open problem; citation and usage are the two impact channels that are measurable for *every* dataset in the corpus and standard in the science-of-science literature. Second, the two are far less redundant than "if used, then cited" suggests: pooled over the modelable corpus, the log-scale correlation between cumulative citation and cumulative usage counts is 0.41 (∼17% shared variance), and on



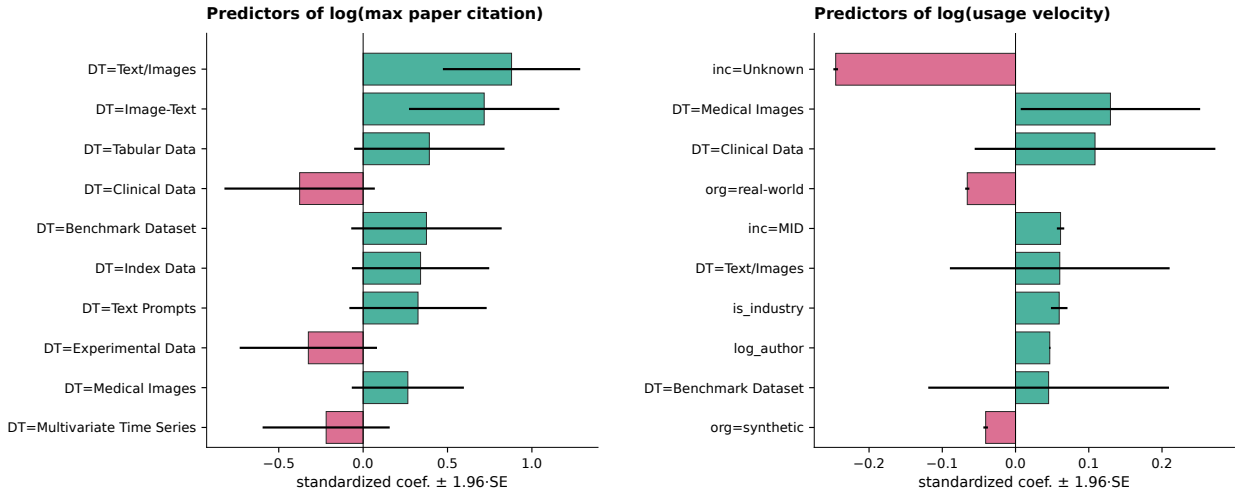
Supplementary Figure S6: **Collaboration patterns in the imputed-creator high-coverage subset.** Four-panel extension of the main-text collaboration breakdown (Fig. 2**c, d**). Both this figure and the main-text panels are computed on the imputed-creator high-coverage subset ($\sim$85% creator coverage), where the creator institution is taken from `creator_institutions` when populated and otherwise imputed as the institution of the earliest arXiv paper that references the dataset. Bars give absolute mean `cite_1yr` (green) and `use_1yr` (pink) per class. **(a) Sector mix**: academic ($n = 129{,}579$), cross-sector ($n = 6{,}112$), industry ($n = 9{,}887$). **(b) Country mix**: single-country ($n = 98{,}175$) vs. multi-country ($n = 10{,}421$). **(c) Income-class mix**: HI ($n = 76{,}319$), LOW ($n = 26{,}453$), HI+LOW ($n = 5{,}814$). **(d) Sector $\times$ country interaction**: `cite_1yr` per cell separating single-country (gray) and multi-country (green) within each sector class. Across all four panels the qualitative ordering matches the main-text findings: cross-sector and multi-country collaborations carry a roughly 2–3$\times$ `cite_1yr` lift over single-sector or single-country baselines, the HI+LOW combination carries the largest income-class premium, and the sector $\times$ country interaction concentrates in the cross-sector multi-country cell.



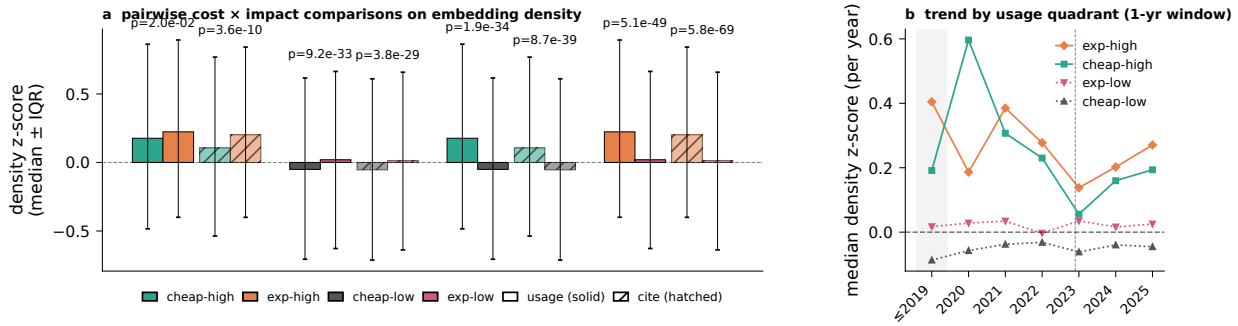
Supplementary Figure S7: **Income-class coverage lift from the first-arXiv-paper imputation, and the resulting time-resolved view. (a) Coverage lift by income class.** Explicit-creator attribution (gray) covers only ∼4% of the corpus and is unevenly split across classes (HI-only 17,419, LOW-only 5,817, HI+LOW 4,826 after dataset-name canonicalization to dedupe aliases on the 2010–2025 year window; 28,062 income-classified). Imputing the creator institution as the institution of the *earliest arXiv paper that references the dataset* (green) lifts the income-classified sample to ∼106,811 canonical datasets (∼4×), with HI-only ∼4.4×, LOW-only ∼4.6×, and HI+LOW ∼1.2× (the smallest lift, since HI+LOW already requires multiple distinct creator countries which the explicit field tended to retain when populated). The imputed view is what Fig. 2**e** plots. **(b) Income-class share over time** on the imputed creator: HI-only dominates throughout (over 70%); LOW-only holds a stable ∼20% share with a slight rise in 2024–2025 as more non-OECD creator institutions enter the corpus; HI+LOW is the smallest sliver (∼5–8%) but is disproportionately influential on the `cite_1yr` panel. **(c) Per-year average `cite_1yr` by income class.** HI+LOW collaborations consistently average the highest early-citation rate (median over 2018–2025 around ∼20 `cite_1yr`), while HI-only and LOW-only sit much closer (∼7–9 `cite_1yr`) and track each other within noise. The HI+LOW premium widens visibly in the LLM era (post-2022): cross-class international collaboration is the strongest associated signal with early citation in our high-coverage sample. Years with fewer than 20 income-classified datasets are suppressed.

the per-year velocity scale used in the regressions the two are essentially uncorrelated ($r = -0.08$); a dataset can be widely adopted without spawning a single highly-cited paper, or ride one viral paper without broad reuse. We therefore report the two outcomes as separate axes throughout rather than collapsing them into a single frontier; every panel that contrasts a predictor on cite vs. use (e.g. Fig. 2**c**–**e** in the main text) is a paired reading of the same datasets on both axes, and the systematic cite-moves/usage-does-not decoupling documented in §2.2 is itself one of the paper's findings rather than a nuisance correlation.

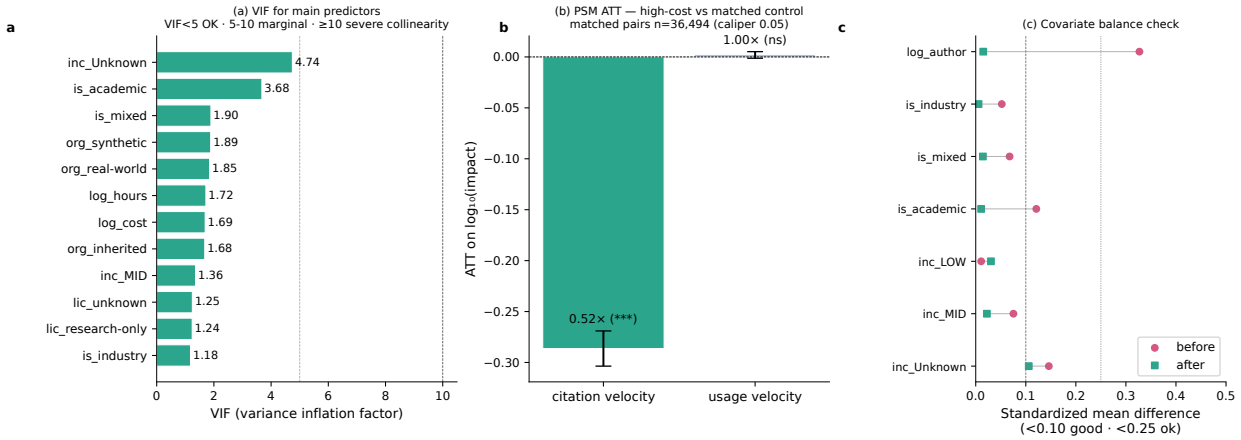
Joint (conditional) effects. Because marginal coefficients can in principle mask joint effects (a factor that contributes only conditional on another), we bracket every headline estimate from both extremes rather than reading the multivariate fit alone. At one extreme each predictor is fitted *alone*, via univariate yearly betas with no controls (Fig. S12) and per-year univariate R^2 against the full-model R^2 (Fig. S13); at the other extreme the predictor enters last, after all controls. Cost is null at both endpoints (alone it explains essentially zero variance in every year, and added last it moves neither outcome), so a conditioning structure under which cost matters would have to operate only at intermediate combinations of controls, a pattern for which the explicit interaction probes (cost × domain / origin / tier; Fig. S15) show no evidence beyond the small synthetic/inherited-origin term. Table S1 collapses the same univariate associations into pooled 1-sd multipliers on the two velocity outcomes used throughout the regressions (`cite_velo`, `use_velo`), making the single-predictor effect sizes directly comparable: a +1 sd move on $\log_{10}$ authors multiplies citation velocity by ∼2.1×, while the same move on machine cost or human-hours buys at most 1.10× on either outcome.



Supplementary Figure S8: **Standardized regression coefficients for citation and usage on the same $n = 154{,}576$ canonicalized corpus and control set (income, license, origin, data-type, year FE).** Year fixed effects fit but hidden; top-10 standardized coefficients ±1.96 SE; green = positive, pink = negative. **(a) Citation**, outcome $\log_{10}(\text{max paper citation} + 1)$: $\hat{\beta}(\log_{10}\text{cost}) = +0.006$ (essentially zero, $\sim 90\times$ smaller in absolute value than $\hat{\beta}_{\text{authors}}$), while $\hat{\beta}(\log_{10}\text{authors}) = +0.514$ is the dominant non-year predictor. **(b) Usage velocity**, outcome $\log_{10}(\text{usage velocity})$: the cost null persists ($\hat{\beta}(\log_{10}\text{cost}) = +0.001$), and the author-count effect on usage (+0.161) is roughly one-third of the corresponding effect on max-paper-citation, consistent with the citation/usage decoupling that Fig. 2**a,b** unpack year-by-year on the 1-year-window outcomes.



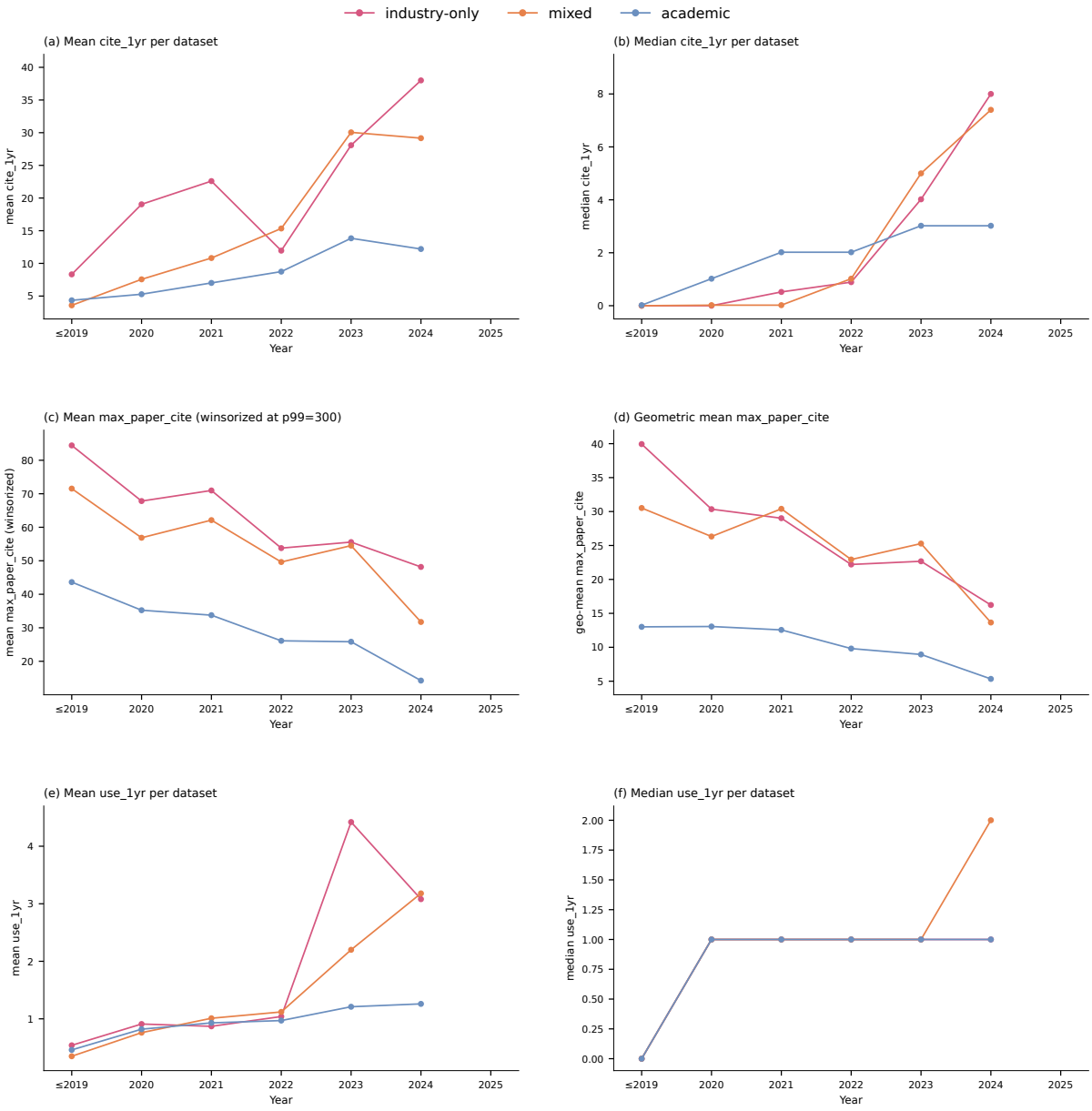
Supplementary Figure S9: **Embedding-density evidence: high impact rides topical momentum.** For each dataset D with release year y_D, we embed its ChatPD summary with `all-MiniLM-L6-v2` and compute the year-z-scored mean cosine similarity to the top-20 nearest neighbors strictly from earlier years. Higher density $\Rightarrow$ D enters a crowded, already-active topical neighborhood. **(a) Pairwise density-z comparisons** across cost-tier × impact quadrants (cheap / expensive × low / high usage and citation); all four high-vs-low Mann–Whitney contrasts reject H_0 at $p < 10^{-10}$. At this corpus size ($n \gtrsim 150$K) significance is dominated by sample size, so the headline story is carried by effect-size differences (median density-z gap of ~ 0.17–0.22 between high- and low-impact quadrants) rather than by the p-values themselves. **(b) Median density z-score per usage quadrant over time** (2019–2025); the high-impact quadrants (cheap-high, exp-high) sit persistently above zero while the low-impact quadrants track zero. The impact main effect (+0.218 z) is $\sim 3\times$ the cost main effect (+0.075); *topic heat predicts citation more strongly than cost does.* The regression result these panels support is reported in the §2.2 (2) topic-heat paragraph.



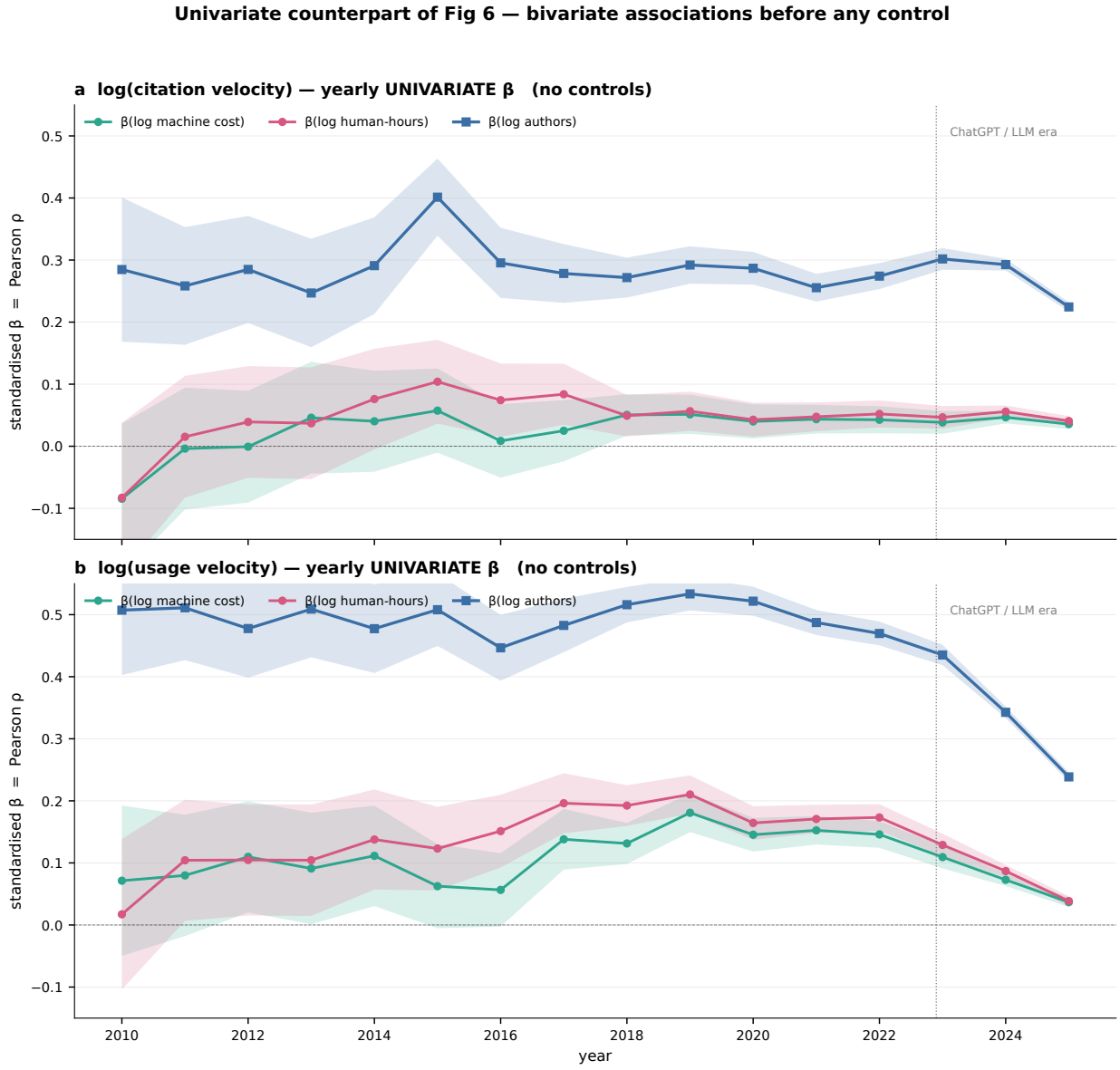
Supplementary Figure S10: **Multicollinearity and propensity-score matching. (a)** VIF for all main predictors below 5 ($\log_{10}$ cost = 1.69, $\log_{10}$ author = 1.14); the cost null is not the result of covariate absorption. **(b)** Propensity-score matching on {author count, institution, year, income, origin, data-type} with caliper 0.05 gives 36,494 matched high-cost vs low-cost pairs (treatment threshold ≈ \$52K per dataset). The high-cost arm is 0.52× as cited as matched controls ($\Delta \log_{10} = -0.286$, $p \ll 10^{-3}$) and indistinguishable from controls on usage (1.00×, $p = 0.24$), so matched comparisons reinforce rather than overturn the cost-null direction in the main regression. **(c)** All standardized mean differences < 0.05 post-match.

Supplementary Table S1: **Univariate 1-sd multipliers on the velocity outcomes, pooled 2010–2025.** Companion to Figs. S12–S13 ($n = 145{,}978$ canonical datasets with positive cost and hours and at least one citation or usage signal, after alias dedup). r is the Pearson correlation between the standardized $\log_{10}$ predictor and the $\log_{10}$ outcome, i.e. the univariate standardized OLS slope with no controls; the multiplier column converts it into the multiplicative change in the outcome associated with +1 sd of the predictor, $10^{r \cdot \mathrm{sd}_Y}$.

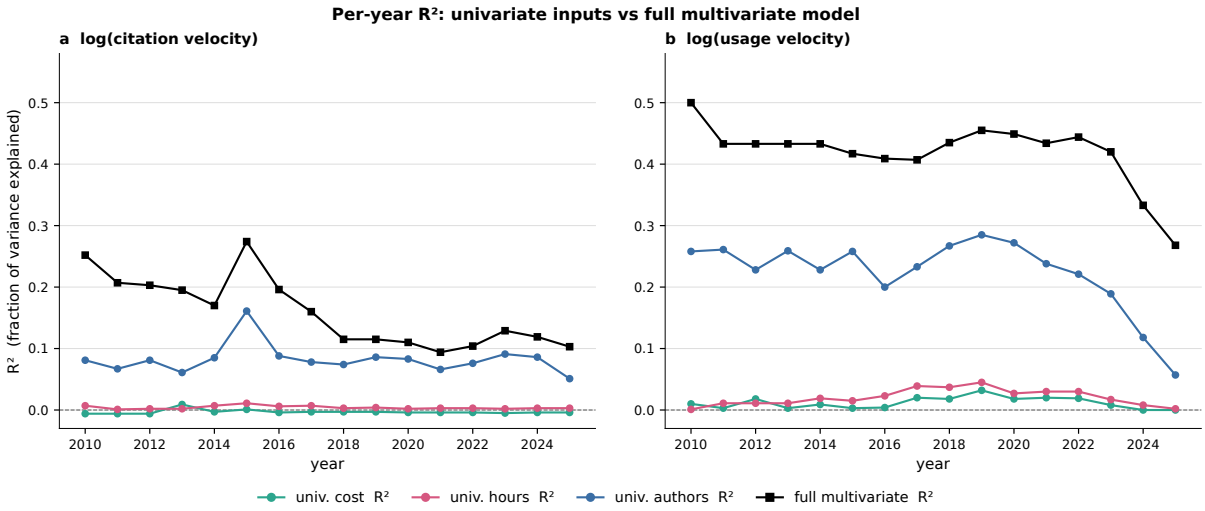
	$\log_{10}$(cite_velo), $\mathrm{sd}_Y = 1.30$		$\log_{10}$(use_velo), $\mathrm{sd}_Y = 0.34$	
Predictor (+1 sd)	r	⇒ × cite velo	r	⇒ × use velo
$\log_{10}$ machine cost (USD)	+0.023	1.07×	+0.067	1.05×
$\log_{10}$ human-hours (h)	+0.032	1.10×	+0.073	1.06×
$\log_{10}$ authors	+0.246	**2.09**×	+0.211	**1.18**×



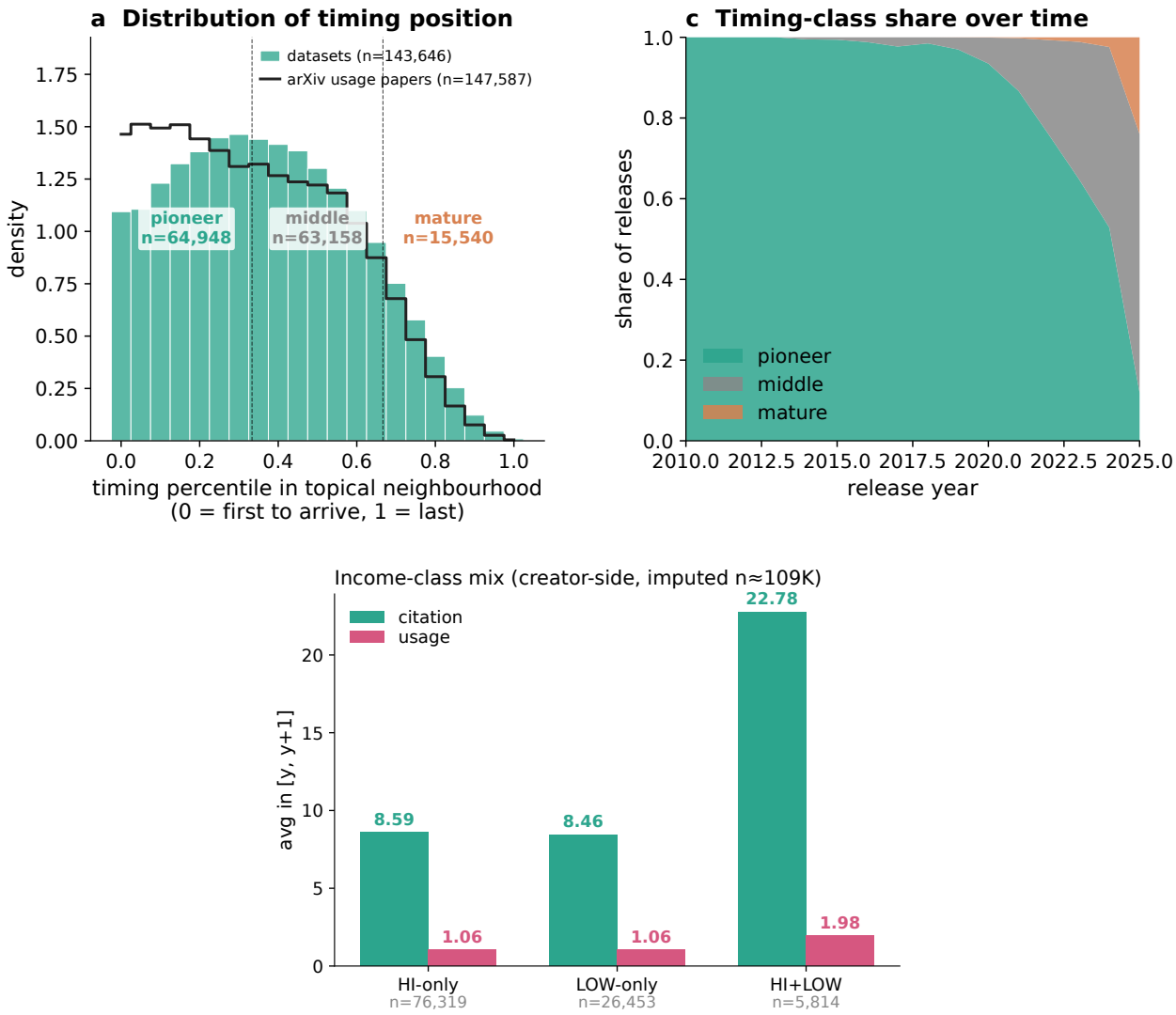
Supplementary Figure S11: **Robustness of the sector ordering across alternative impact metrics.** Same sector-mix bars as Fig. **2c**, recomputed under five alternative impact metrics: winsorized mean and geometric mean of `max_paper_cite`, mean and median `cite_1yr`, and mean `use_1yr`. Every panel preserves the same *cross-sector* > *industry-only* > *academic-only* ordering, indicating that the sector channel is robust to the choice of impact aggregation rather than an artifact of the headline `max_paper_cite` metric.



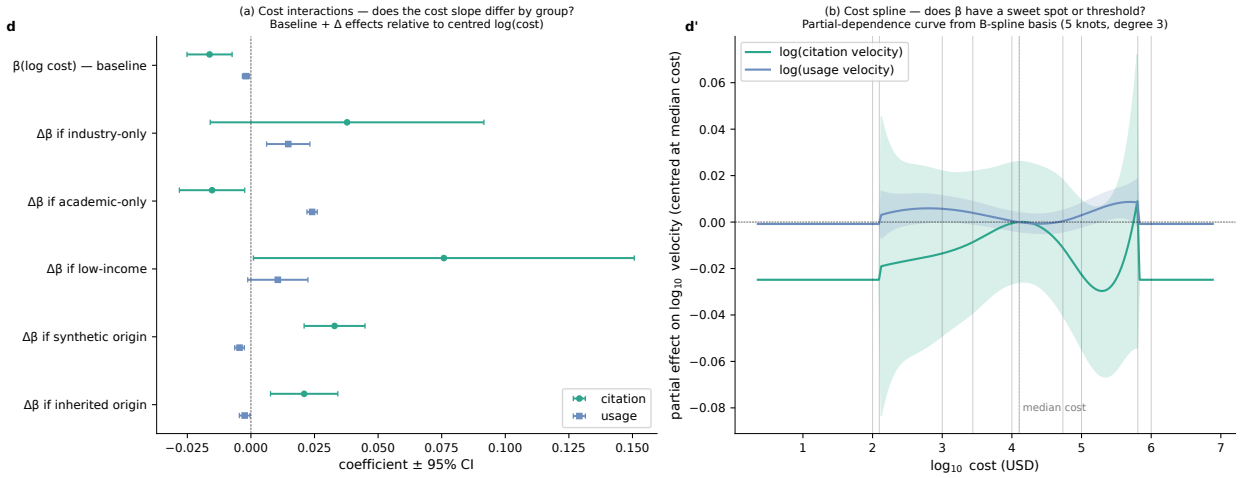
Supplementary Figure S12: **Univariate (single-predictor) yearly $\hat{\beta}$ trajectories on $\log_{10}$(cite_velo) and $\log_{10}$(use_velo).** Companion to Fig. 2**a,b**: instead of the multivariate regression coefficients with all controls absorbed, each predictor (cost, hours, authors) is fitted alone year-by-year. The cost coefficient remains near zero on both outcomes across 2010–2025 even without any controls, ruling out a multicollinearity artifact as the source of the cost null.



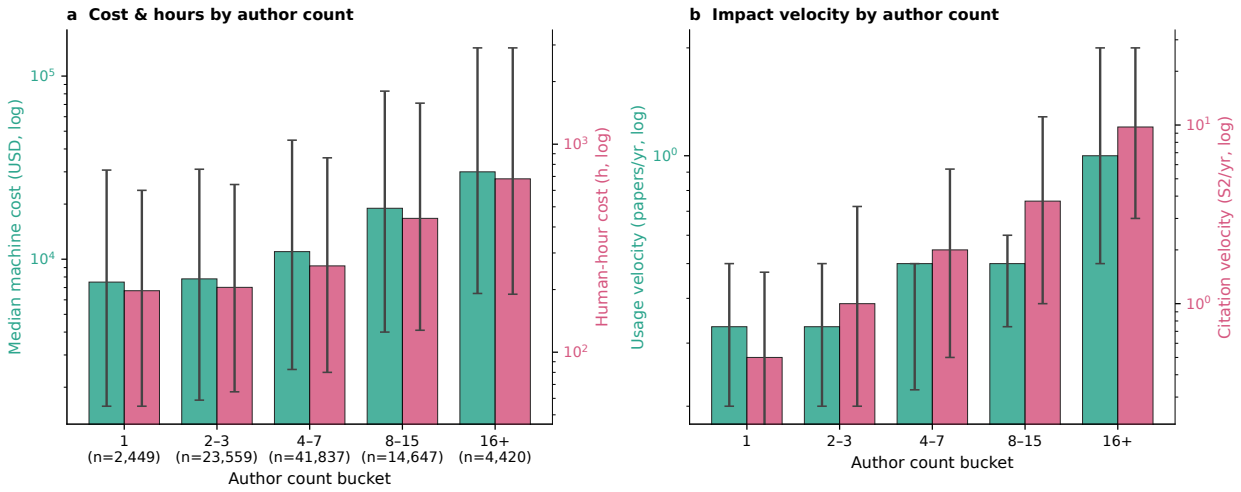
Supplementary Figure S13: **Per-year univariate vs. multivariate R^2 comparison on $\log_{10}$(cite_velo) and $\log_{10}$(use_velo).** The univariate cost R^2 is essentially zero for every year and both outcomes; the multivariate R^2 is carried almost entirely by the authors and year-fixed-effect terms. The gap shows directly that the multivariate cost null is not a covariate-absorption artifact.



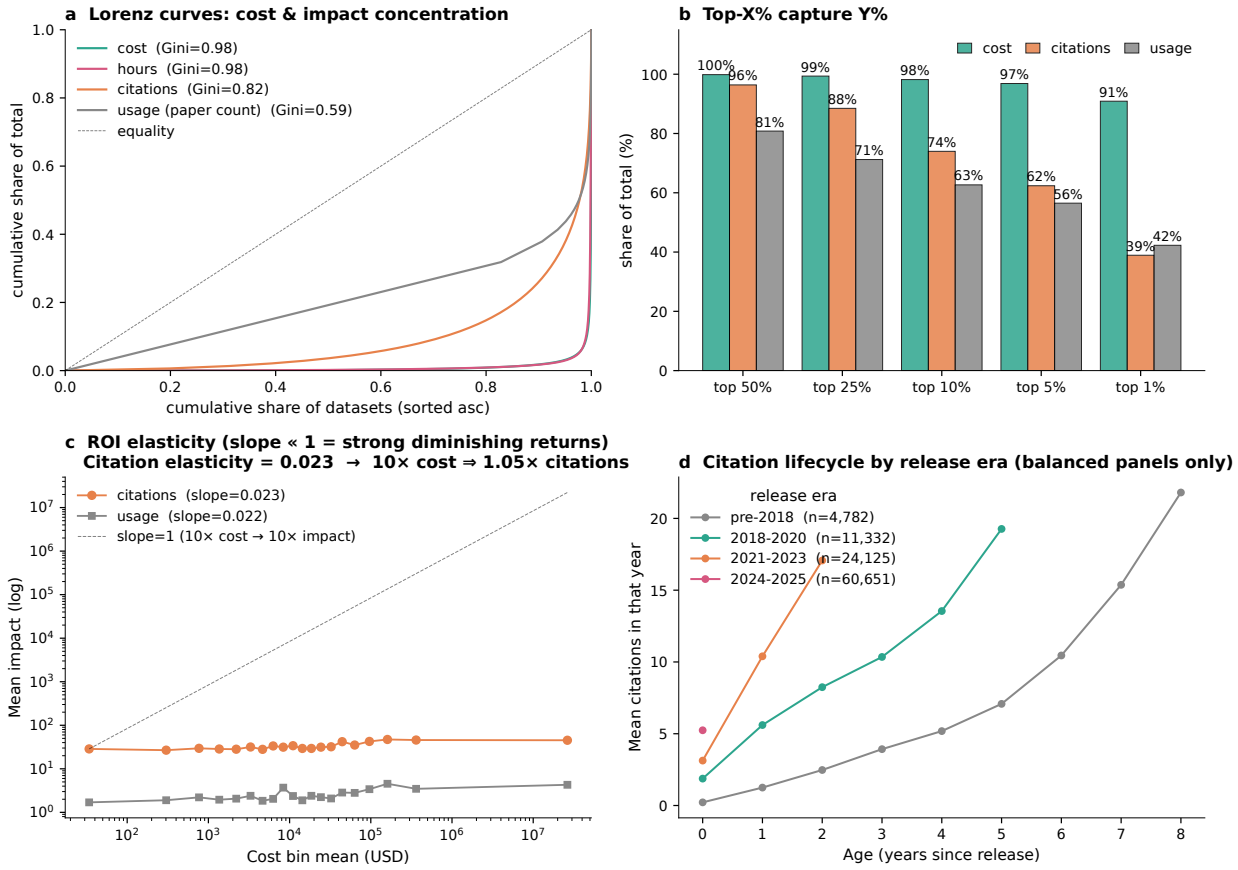
Supplementary Figure S14: **Topical timing-position and income-class collaboration companions. Top row, timing-position panels.** Distribution of the timing percentile across canonical datasets with the same-procedure arXiv-paper baseline overlaid (left; datasets are systematically shifted ∼5 percentage points away from the pioneer bin relative to the usage-paper pool), and the share of each timing class over time (right). Together with Fig. 2**e** these support the interpretation that datasets are typically constructed in response to research needs that have already started to emerge from earlier papers, rather than initiating a topical area themselves. **Bottom row, income-class mix (HI / LOW / HI+LOW) of imputed-creator collaborations and absolute 1-year impact** ($n \approx 109$K imputed-creator canonical datasets). HI+LOW combinations carry a ∼2.6× `cite_1yr` lift over single-income classes, and the lift strengthens under the ∼4× converged sample reported in Fig. S7. The LOW-only premium visible in the smaller explicit-only sample (Fig. S6) does not survive the coverage lift to the imputed-creator pool; the HI+LOW premium does. Bottom-row content is consolidated into this companion figure (rather than as a separate Supplementary figure) to keep Fig. 2**c**–**d** sector- and country-mix focused.



Supplementary Figure S15: **Heterogeneity of cost elasticity. (d)** Interaction terms between $\log_{10}$ cost and sub-group dummies (industry, academic, low-income, synthetic origin, inherited origin). The only significant interactions are with synthetic (+0.017) and inherited (+0.016) origin, a small but real signal that marginal spending on synthetic generation or re-curation buys slightly more citation than the average. No special premium for industry or low-income teams. **(d′)** Spline partial-dependence: 95% CI for $\log_{10}$ cost covers zero across the full support (−3 to +8 in log USD). The cost null is not a misspecified-form artifact.



Supplementary Figure S16: **Cost, hours and impact by author-count bucket** ($n = 159{,}194$)**. (a)** Median cost (green left axis) and hours (pink right axis). Cost rises 4.6× from solo to 16+; hours 3.75×. **(b)** Median usage velocity (green left axis) and citation velocity (pink right axis). Usage velocity is identical between solo and 16+ buckets (1.0 paper/yr); citation velocity is 9.6× higher. Usage is flat in team size; citation propagation is super-linear.

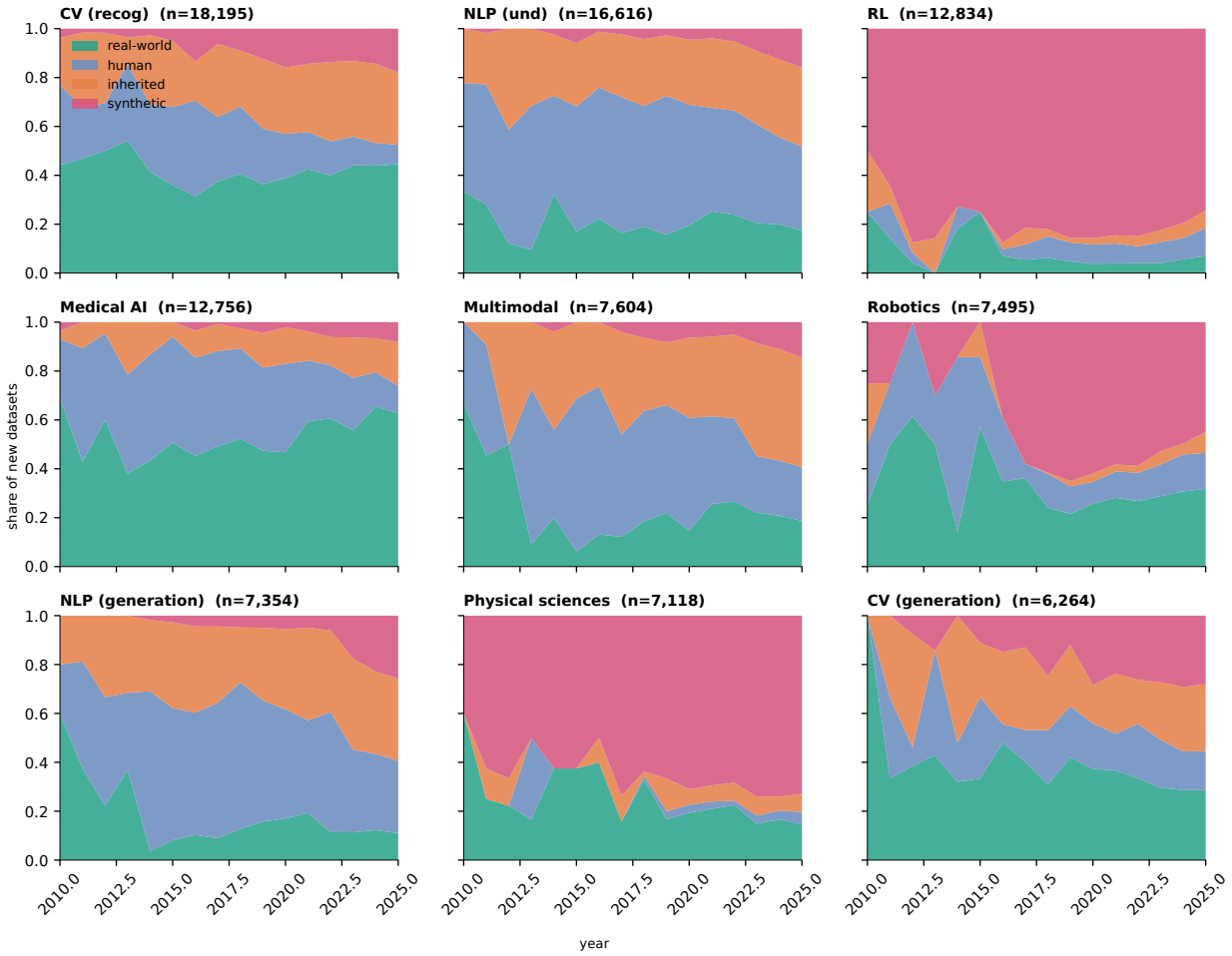


Supplementary Figure S17: **Cost is highly concentrated; citation and usage impact are markedly less concentrated; older cohorts continue accumulating citations long after release. (a) Lorenz curves** for cost, hours, citations and usage on $n = 155{,}675$ datasets. Gini coefficients: cost 0.98, hours 0.98, citations 0.82, usage 0.59. The spread between the cost and impact Lorenz curves is the central asymmetry of the ecosystem. **(b) Top-X% concentration.** The top 1% of datasets absorb 91% of total cost but only 39% of citations and 42% of usage; the top 5% absorb 97% of cost, 62% of citations and 56% of usage. **(c) ROI elasticity.** Cost-bin means of impact in $\sim$20 quantile bins; citation log–log slope = 0.023 (10× cost $\Rightarrow$ 1.05× citations), with usage slope = 0.022; the dashed unit-elasticity reference shows slope = 1 for contrast. **(d) Citation lifecycle by release era (balanced panels only).** Mean cite/year as a function of dataset age, stratified by release cohort: pre-2018 (n=4,782), 2018–2020 (n=11,332), 2021–2023 (n=24,125), and 2024–2025 (n=60,651). Older cohorts continue accumulating citations through age 8+, while the 2025 cohort has at most one year of observable accumulation, the basis for the short observation-window caveat discussed in §2.2.

S1.3 Compositional structure of the dataset ecosystem

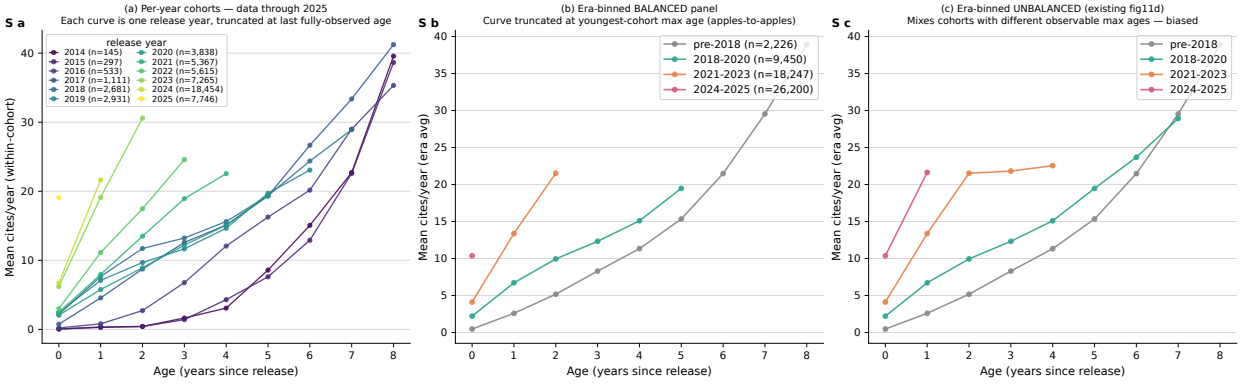
This subsection documents the compositional shifts (origin, research domain, lifecycle, industry attribution) that the body refers to as "compositional context". None of these are causal claims; they are descriptive and serve as backdrop for the investment-flow narrative in §2.1–2.2.

Fig 10: Origin composition over time, by subdomain (top 9)

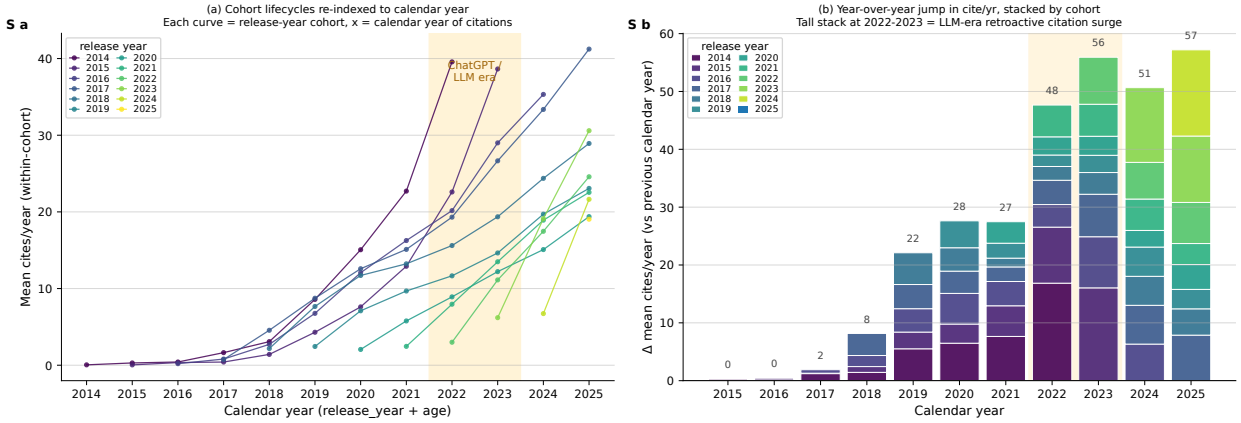


Supplementary Figure S18: **The composition of newly-released datasets shifts from real-world toward inherited and synthetic, but the human→synthetic substitution is heterogeneous across research domains.** Each panel shows the share of new datasets by origin class (real-world, human, inherited, synthetic) over the 2010–2025 window, faceted across the nine highest-volume research domains in the 173K collection. Three patterns co-exist: a clean human→synthetic substitution in CV-recognition, NLP-understanding, NLP-generation, multimodal and medical AI (human share falls 15–28 percentage points; synthetic rises 9–21 percentage points); a synthetic → human reversal in RL, robotics and mathematical reasoning, where pre-existing 60−80% synthetic shares fall as frontier-model training pulls expert demonstrations and reasoning traces back in; and a real-world → synthetic share rotation in physical sciences and computational social science, where rising synthetic share coincides with a decline in field-collected data rather than in human annotation. In 2025 across the full sample, real-world has fallen to $< 20\%$ of new releases (from $\sim$50% in 2010) while synthetic plus inherited has crossed 50%.

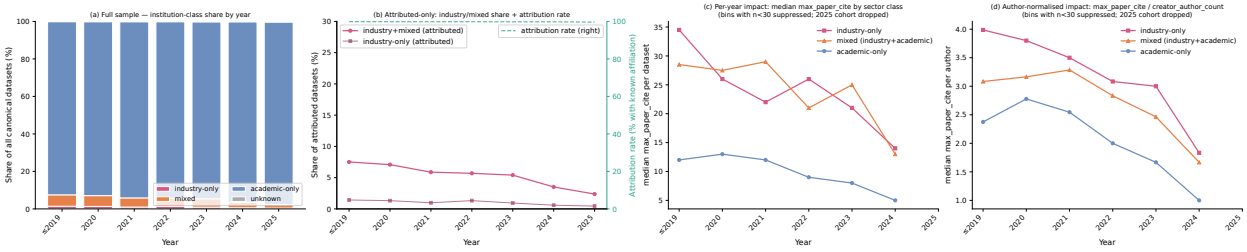
Supplementary Figure S19: **Research-domain lifecycle stages: not all fields are exhausted the same way.** **(a)** Per-domain post-2022 origin-share trajectories. **(b)** Exhaustion score decomposition, ES = synth + inherited − 0.5 · rw, sorted high to low across the 12 research domains. Earth and atmospheric sciences are the only Stage-I field-collection domain remaining.



Supplementary Figure S20: **Cohort lifecycle on a balanced age axis (data through 2025-12-31; partial-year 2026 stripped).** Per-cohort cite/yr at each age $a_0, a_1, \ldots, a_8$. The release-year (a_0) citation velocity rises $\sim 600\times$ across cohorts (2014: 0; 2018: 2.3; 2025: 19.5). At fixed age the cross-cohort multiplier is $8.7\times$ at a_4 between 2014 and 2021. Pre-2018 cohorts continue to accumulate (2014 reaches 41.9 cite/yr at a_8) without plateauing.



Supplementary Figure S21: **Cohort lifecycle on the calendar-year axis: the LLM-era citation revival is a rolling reactivation of legacy cohorts.** The same data as Fig. S20 but indexed on calendar year rather than age. Total $\Sigma\Delta$ cite/yr across cohorts doubles between 2021 (+27) and 2022 (+52); the 2022 increment is $\sim$56% driven by the 2014–2015 cohorts. Subsequent years roll the revival forward (2023: 2015–2016 cohorts; 2024–2025: 2016–2017 cohorts). Within each old cohort during 2024–2025 the top-1% citation share jumps from $\sim$ 22% to $\sim$ 40%, indicating that the revival is highly selective on a small set of legacy benchmarks.



Supplementary Figure S22: **Industry attribution share over time, attributed sub-sample. (a) Industry vs. industry-plus-mixed share** (y-axis: share of *attributed* datasets only): "industry-only" has remained at $\sim$ 3% across the entire 2010–2025 window; industry+mixed peaks at 20% in 2020–2023 and falls to 16% in 2025. **(b) Attribution rate** (y-axis: share of *all* datasets, attributed or not): the share of datasets with verified creator institution collapses from 50% in 2010 to 15% in 2025, driven by an inflow of HuggingFace community uploads without complete authorship metadata. The two y-axes differ (attributed-only fraction vs. fraction of the full corpus). The cross-sectional industry premium (regression $\hat{\gamma}_{\text{ind}} = +0.205$) is therefore a *stock* effect on existing-dataset citations, not a *flow* effect of an industrializing creator base.

S1.4 Creator vs. usage authors: which author count drives downstream citation?

How author counts are measured and verified. The author count is not self-reported dataset metadata: it is the length of the bibliographic author list (arXiv / Semantic Scholar records) of the paper we resolve as the dataset's *creator paper*. Resolution proceeds in three passes: the linked paper itself where ChatPD attaches a unique paper; a Semantic Scholar title search (the most-cited paper whose title names the dataset) otherwise; and a web-search-grounded LLM lookup for the remainder. Every candidate assignment is then verified by a Claude Opus 4.5 batch judge answering "is this the paper that introduced this dataset?" (78,233 assignments confirmed, 92,932 rejected; rejected datasets fall back to the author list of their most-cited *using* paper). Rather than committing to a single definition a priori, we also compare the competing author-count features (creator-paper authors vs. using-paper authors) by predictive importance; both give the same qualitative result, and the rest of this subsection decomposes their relative strength.

The §2.2 (1) author-count finding pools authors at the dataset level: `authors` is the count of the dataset's *creator-paper* authors and the dependent variable is impact aggregated across all downstream users. A natural follow-up is whether the citation a dataset accumulates is propagated more by the *creator*'s author network or by the *using papers'* author networks. We test this by comparing two log–log slopes on the full canonicalized corpus ($n = 172{,}069$ datasets with complete creator-author, usage-author, total-citation and max-cited-paper fields).

Supplementary Table S2: Creator-author vs. usage-author predictors of downstream citation, $n = 172{,}069$. Slope is the log–log OLS coefficient $\hat{\beta} = \partial \log_{10}(\text{citation}) / \partial \log_{10}(\text{authors})$; the multiplier column converts $\hat{\beta}$ to "10× authors ⇒ M× citation". "Top-cited-using-paper" is `max_cited_paper.citationCount`; "citation sum" is the cumulative S2 citation count of all using papers as of the 2025-12-31 cutoff.

Predictor (per dataset)	**Outcome (per dataset)**	$\hat{\beta}$ (log–log)	**10× authors ⇒**	ρ_S
`creator_author_count`	citation sum	+0.34	2.18× cites	+0.31
`creator_author_count`	top-cited-using-paper	+0.38	2.38× cites	+0.10
`usage_author_count`	citation sum	+0.34	2.18× cites	+0.36
`usage_author_count`	top-cited-using-paper	**+0.70**	**5.04**× cites	+0.28

Reading. On the top-cited-using-paper outcome (the quantity behind the headline `max_paper_cite` impact metric used in RQ2, §2.2), the usage-author slope (+0.70) is almost double the creator-author slope (+0.38). A 10× larger pool of *using*-paper authors is associated with ∼5× more citation on the top using paper, whereas 10× more *creator*-paper authors associates with only ∼2.4×. On the cumulative citation-sum outcome the two slopes are equal (+0.34), because the sum aggregates over the same downstream author network and dilutes the gap.

The descriptive implication is that the citation a dataset accumulates downstream is dominated by *who is doing the citing*, not by who originally built the dataset. This is consistent with the RQ2 author-count reading (§2.2) that the citation channel runs through using-paper co-author networks rather than through a direct dataset-quality channel. We treat this as a descriptive decomposition; it does not establish a causal claim about the creator's network specifically.

S1.5 Impact as a function of topical density and timing percentile (flipped view)

Fig. S9 and Fig. 2**e**–**f** report *density* and *timing* on the y-axis, stratified by impact class. This Supplementary subsection flips the view (impact on the y-axis, density / timing on the x-axis), so that the implied effect on impact rank is read directly.

Supplementary Table S3: Density z-score by impact class, year-z-scored kNN density of top-20 strictly earlier neighbors in the `all-MiniLM-L6-v2` summary-embedding space ($n = 151{,}243$ canonical datasets with finite density). Higher density z = denser pre-existing topical neighborhood at release. Flipping the y-axis to impact: holding cost fixed, high-impact datasets sit at a markedly higher density-z than low-impact ones, with the gap larger in the expensive bucket. All four high-vs-low Mann–Whitney tests reject H_0 at $p < 10^{-10}$.

Cost-tier × impact class	n	**median density-z**	**Q25 z**	**Q75 z**
cheap, low usage	72,506	−0.05	−0.71	+0.62
cheap, high usage	3,113	+0.18	−0.48	+0.86
expensive, low usage	70,996	+0.02	−0.63	+0.66
expensive, high usage	4,628	+0.22	−0.40	+0.89
cheap, low cite	68,629	−0.05	−0.71	+0.61
cheap, high cite	6,990	+0.11	−0.54	+0.77
expensive, low cite	67,731	+0.01	−0.64	+0.66
expensive, high cite	7,893	+0.20	−0.40	+0.84

Supplementary Table S4: Mean 1-yr impact by timing-position class ($n = 143{,}646$ canonical datasets with finite kNN-density percentile, K = 20). Timing percentile = fraction of the 20 nearest topical neighbors released strictly before the dataset. Flipped y-axis (impact): `cite_1yr` falls ∼3.4× from pioneer to mature, while `use_1yr` stays roughly flat. Citation is strongly temporally selective; adoption is not.

Timing class (pct of earlier neighbors)	n	**mean `cite_1yr`**	**mean `use_1yr`**
pioneer ($\leq 1/3$)	64,948	10.32	1.04
middle (1/3, 2/3)	63,158	5.99	1.14
mature ($\geq 2/3$)	15,540	3.04	1.07
pioneer / mature ratio		$\sim 3.4\times$	$\sim 0.97\times$

Joint reading. High-impact datasets are simultaneously (i) in *denser* topical neighborhoods (already-crowded research areas) and (ii) earlier within those areas' release order. The two dimensions move impact in opposite directions on the same percentile axis: density rises with neighborhood crowdedness, timing percentile rises with lateness. The impact main effect of density (+0.218 z; Fig. S9) is roughly 3× the impact main effect of cost (+0.075 z), and the timing gradient is sharply non-trivial on citation while almost flat on usage. Both patterns are consistent with the RQ2 framing (§2.2): *when* a dataset enters a topical neighborhood and *how active* that neighborhood is predict citation more strongly than *how much* the dataset cost to build.

S1.6 Cross-platform adoption signals (small-n exploratory)

Caveat. This subsection reports a small-n pilot pull of HuggingFace download counts and GitHub stargazer counts versus the arxiv citation/usage metrics used elsewhere in this paper. Sample size is too small for main-text claims: $n = 23$ frontier benchmarks (manually mapped) plus $n = 37$ / 47 HuggingFace / GitHub URLs extractable from the upstream `chatpd url` field (a 1.4% / 9.9% slice of the full corpus). We treat the resulting correlations as directional indications only; the 95% CIs at this n may straddle 0 and we do not bootstrap them here. A proper multi-platform analysis would require: (i) large-scale fuzzy matching from canonical name to HuggingFace / GitHub / Kaggle ID, (ii) time alignment between HF's 30-day rolling download window and the arxiv cumulative citation count, (iii) explicit handling of the selection bias whereby being on HuggingFace is itself a signal.

Pilot Spearman rank correlations. HuggingFace download (last 30 days) versus arxiv citation (top-1 max user paper, our primary metric throughout the paper) is $\rho = +0.108$ on the frontier benchmarks ($n = 23$) and $+0.180$ on the URL-extracted sub-sample ($n = 37$). Versus arxiv usage (paper count) it is $+0.239$ / $+0.045$. GitHub stars versus arxiv citation is $\rho = +0.005$ ($n = 47$); GitHub stars versus arxiv usage is $+0.207$. For comparison, HF likes versus HF downloads (both of which are HF internal signals) correlate at $\rho = +0.717$.

Indicative reversals. On the frontier list, the top-10 by HF download and the top-10 by arxiv citation overlap in only 7 of 10 slots. ARC-Challenge (HF $\sim$ 427K downloads, arxiv cite 34), SWE-Bench Verified (HF $\sim$ 720K, cite 186), and GPQA-Diamond (HF $\sim$ 123K, cite 61) sit at the top of the HF distribution but well below the cite median; C-Eval (HF $\sim$ 55K, cite 5,336) and CMMLU (HF $\sim$ 20K, cite 5,336) are the reverse. These are consistent with HF and arxiv indexing different research communities (HF: people running evaluations against released code/data; arxiv: people writing follow-up papers), but at this sample size the pattern is suggestive rather than load-bearing.

Implication for the main-text mechanism claim. The §2.2 (1) reframe (only the author-count $\rightarrow$ usage sub-channel weakened, while author-network structure $\rightarrow$ citation persists) references HuggingFace / Kaggle / GitHub as a plausible explanation for the usage-side decay. This Supplementary subsection is the empirical basis for treating that as a hypothesis only: the platform mechanism is not directly tested at scale and the small-n correlations here do not establish independence between platform adoption and citation propagation. Larger-scale verification is left to follow-up work.

Platform data accessibility (for replication). HuggingFace exposes `huggingface.co/api/datasets/{id}` without authentication. GitHub exposes `api.github.com/repos/{owner}/{repo}` unauthenticated at 60 req/hr (or 5,000 req/hr with a token). Kaggle's API requires `KAGGLE_USERNAME` + `KAGGLE_KEY` signing and was not pulled in this audit. Pull script: `analysis_173k_0425/hf_gh_audit.py`.

S1.7 Case study: evidence-grounded cost extraction

The 24K subset distinguishes two kinds of cost record. A small fraction of papers report an explicit dollar or hour figure for the construction of their dataset; the remaining majority report only *evidence-grounded* cost details from which a single per-dataset estimate has to be composed. We illustrate the two cases with three concrete entries drawn from the `results/ground_truth_costs_hours_latest.csv` bundle.

Supplementary Table S5: Three case-study extractions illustrating the *evidence-grounded* cost-extraction pipeline. `type` indicates the extraction pattern: *total* = paper states an explicit dollar or hour total; *rate* = paper states a per-unit rate that has to be composed with another reported quantity to produce a per-dataset total; *single-axis* = only one of the two cost tracks is reported and the other is absent (left `NA` rather than imputed).

Dataset	Extraction type	Grounding sentence (verbatim)	$ extracted	h extracted	n
PolitiFact	total	"They are compensated a total of $450 based on their working hours."	$450	30 h	774
DUC 2004	single-axis (h only)	"3,000 hours of human effort were required for a simple evaluation of the summaries."	NA	3,000 h	250
SQuAD V2.0	rate	"Workers were asked to spend 7 minutes per paragraph, and were paid $10.50 per hour."	$10.50/h (rate)	7 min / paragraph (rate)	150,000

Walk-through for the rate case (SQuAD V2.0). Because no total dollar figure appears anywhere in the SQuAD V2.0 paper, the pipeline composes it from the two grounded rates. The extractor stores the pair as `(raw_cost_text = "$10.50 per hour", cost_unit_type = per_hour)` together with the parallel `(raw_hours_text = "7 minutes per paragraph", hours_unit_type = per_instance)`. With the dataset's `instance_count = 150,000` paragraphs the composed per-dataset estimate is $150{,}000 \times 7\,\text{min}/60 \approx 17{,}500$ h on the hours axis and $17{,}500\,\text{h} \times \$10.50 \approx \$184{,}000$ on the dollar axis. Both numbers are stored on separate axes and are never collapsed through a wage assumption; the per-hour rate above is recorded only as evidence and is not propagated to the headline \$/h ratio used elsewhere in the paper.

Walk-through for the single-axis case (DUC 2004). The DUC 2004 paper reports total human-effort hours (3,000 h) but no dollar figure. The pipeline records `converted_hours = 3000` and leaves `converted_cost_usd = NA`. The estimator predicts a USD value at inference time from neighbors in the embedding space; the "NA-on-the-source-axis but predicted-by-the-model" pattern is the main reason we describe the supervision signal as *evidence-grounded extraction* rather than *ground-truth cost* (only $\sim$0.6% of records carry an explicit total dollar figure; see Table S11).

Walk-through for the total case (PolitiFact). PolitiFact is the simplest pattern: the paper states a dollar total (\$450), a corresponding human-hour total (30 h), and the dataset size (774 instances). Both axes are directly extracted and no composition is required. PolitiFact is therefore the case that comes closest to a true "ground-truth" anchor; we use such records preferentially in the slot-filling reward and in the out-of-family LLM-judge rubric described in the main Methods.

Together these three cases illustrate why our supervision signal is best described as *evidence-grounded extraction*: most training rows are not direct ground-truth dollar amounts but structured fragments of grounding text plus their numeric slots, from which the model must compose a per-dataset cost.

S1.8 Held-out evaluation as a partial substitute for human audit (MC1 response)

One reviewer concern asks for an additional audit, **MC1**: an external-anchor cost-MAE audit ($\sim$200 rows) to verify the absolute dollar levels reported in §2.1. We argue here that the held-out evaluation protocol already deployed in the paper covers most (though not all) of what this audit would test, and we describe the small residual audit that closes the remaining gap.

What the held-out protocol already does. The cost-estimation model is evaluated on a frozen **1,367-sample held-out test set**, drawn by SHA-256 hash on `dataset_name` to guarantee zero dataset overlap with the training pool (Methods Tab. 3, Table S8). The test set is then evaluated under **two independent protocols**: (i) the slot-filling **log-MAE** on per-number predictions (Tab. 2), reported within 2×, 5× and 10× accuracy bands; and (ii) the **nine-dimension LLM-judge rubric** (Tab. S9) scored by GPT-5-mini, chosen as a non-Claude, non-Qwen model to reduce the well-documented self-preference bias of an in-family judge. Both protocols use the **anonymized** version of each test sample (dataset name masked with `Dataset_XXXX` and all evidence-text mentions of the name redacted), which forces the model to reason from the grounding text rather than from memorized facts about the benchmark; the anonymization drop (-0.351 for GPT-5-nano, -0.092 for Opus, -0.109 for our GRPO model; Tab. 3) is itself a diagnostic that the remaining score reflects evidence-based reasoning rather than recall.

MC1 external-anchor cost-MAE: what the held-out set already substitutes for. The 1,367-sample test set covers exactly the population MC1 asks about (per-dataset dollar estimates), and the within-test subset of records that carry an explicit author-reported cost figure provides direct anchoring to externally reported budgets, without an extra audit loop. This subset is ∼0.6% of the 24K extraction base, or roughly 80 records inside the test split if scaled proportionally (the PolitiFact / SQuAD V2.0 / DUC 2004 family in Tab. S5). For these records the log-MAE on the dollar-grounding subset is reported alongside the headline number in Tab. 2, and the anonymization discipline guarantees that the score is not inflated by memorization of the specific dollar figures. *What the held-out set does not test* is whether the model's dollar predictions *outside* this explicit-budget subset agree with externally verifiable anchors (grant disclosures, OECD annotation-wage benchmarks, large-scale crowd-platform receipts). To close that gap we commit to **a 200-row external-anchor cost-MAE cross-check**: we will sample 200 datasets whose creator paper or accompanying NSF / ERC / NIH grant page reports a verifiable construction budget, extract that figure from the external source, and report log-MAE of the GRPO model against the external anchor side-by-side with the held-out log-MAE. As suggested in MC1, we also restate the §2.1 dollar levels as **rank-based percentiles** (Fig. 1**e**, Tab. S15) which are robust to a constant-factor distortion in the absolute dollar scale.

Where the held-out protocol genuinely cannot substitute for human review. The held-out set is drawn from the same ChatPD distribution as training and is anonymized but not curated for external auditability; the 200-row MC1 audit above is needed to verify that the held-out log-MAE generalizes to a population that includes records with externally verifiable budgets. The audit is scoped narrowly because the broader question (*do the model's predictions reflect evidence-grounded reasoning rather than memorization, on data the model has not seen?*) is already covered by the existing held-out + anonymization + cross-family-judge protocol.

S1.9 Methodological caveats

Hours are not monetized. We never convert hours to dollars at a fixed wage. Doing so would force a single "total cost" axis whose dominance split (machine-vs-labor) hinges entirely on the assumed hourly rate, with no stable defense across countries, decades and skill levels [1, 2]. All claims rest on either the USD axis or the raw-hours axis, separately; rank-based percentile comparisons are used when a single ranking is required.

Year attribution. For datasets without a self-reported creation year (the majority of the post-2022 collection) we use `arxiv_earliest_year`: the year of the first arXiv paper that uses the dataset. By construction this is the creator-year proxy whenever the creator paper is on arXiv (typical of ML research). The 2024–2025 dataset-count surge is therefore a real signal of arXiv-detected dataset creation, not an attribution artifact, but it is a count of *newly named* datasets and partial-year cutoff effects apply to 2026 (excluded).

S1.10 Alias canonicalization pipeline

The 173K corpus contains many dataset *aliases* that refer to the same underlying dataset, e.g. `MMLU-Pro`, `MMLU Pro`, `MMLUpro`, `MMLU_Pro` and `MMLU-Pro+`, or `MS COCO`, `MS-COCO`, `MSCOCO`. Treated as separate rows, these aliases inflate the corpus count, double-count any per-dataset average, and split a single dataset's citations across multiple records. To produce a clean per-dataset view we run a **name-normalization + grouping** pass that lower-cases each `dataset_name`,

strips all common separators (whitespace, `- _ . /` parentheses brackets braces colons semicolons commas) and keeps only alphanumerics plus version markers (`+ * #`). All raw entries that share a normalized form are grouped, and within each group the **canonical** entry is the one with the largest `total_references` + `total_citations` (ties broken by `cost_avg`). The canonical entry's name carries the citation sum of the entire group.

For the canonical *cost* we use a slightly more careful rule ("S2"): pick the cost from the highest-cited alias whose extracted cost is $\geq$ \$100, falling back to the canonical name's cost otherwise. This protects against a common failure mode in which the most-cited alias mentions the dataset only in passing (no construction details, GRPO extracts a near-zero cost) while a less-cited alias appears in a paper that fully describes how the dataset was built. We additionally apply an "Opus override" for the long tail of GRPO cap-failures (cost $\geq$\$1B): if Opus 4.5's independent estimate is more than 10$\times$ smaller, we substitute the Opus value (55 canonical entries). More generally, this S2-then-Opus arbitration is a deliberate compromise: when GRPO inference fails for the most-cited alias of a canonical group, the next-best estimate from the teacher model is used rather than dropping the row, at the cost of mixing two estimator distributions on a small slice of the corpus. The pipeline collapses the 173,369 pre-canonical entries into $\sim$155,000 canonical datasets. Released artifacts: `alias_canonical_map.json` and `alias_canonical_cost.json`.

Sensitivity to citation-based alias arbitration. We replace the production S2-plus-Opus rule with a citation-free rule that takes the median GRPO cost among aliases whose estimates are at least \$100, falling back to the median finite sub-cap estimate when none qualifies. Across 149,082 comparable canonical groups, the corpus median changes from \$12,500 to \$12,708.50 (+1.7%), and the 2024–2025 cost rebound changes from 1.346$\times$ to 1.331$\times$. For impact checks, which include release cohorts only through 2024, the Spearman cost–cumulative-usage association changes from 0.131 to 0.141, while the cost–cumulative-citation association changes from 0.056 to 0.064; citation and usage events are observed through the December 2025 cutoff, but no 2025 release cohort enters these associations. The aggregate conclusions are therefore insensitive to citation-based alias selection. Individual estimates are less stable: among 9,756 comparable multi-alias groups, 24.8% differ by more than 2$\times$ and 8.2% by more than 10$\times$ under the citation-free rule. We therefore treat canonical dataset-level costs as uncertain when aliases disagree, even though the population-level results are stable.

S1.11 Metric glossary and sample accounting

Supplementary Table S6: Glossary of the metrics used in the main text. Each metric is defined once here; the body uses these names without re-defining them. *Per-dataset cost* metrics are estimator outputs; *impact* metrics are aggregates of downstream-paper records from the ChatPD usage table.

Metric	**Definition**
`machine_cost` (USD)	Estimator output. Per-dataset summed dollar cost across compute & cloud, equipment & sensors, software & tools, licensing & data access, storage & hosting, legal & compliance and travel & events. Reported in US dollars on a separate axis from `hour_cost`; never wage-converted.
`hour_cost` (h)	Estimator output. Per-dataset summed human-hour cost across annotation & labeling, quality control & validation, data collection & scraping, data cleaning & preprocessing, software engineering, documentation & release, design & planning, project management and domain-expert time. Reported in hours on a separate axis from `machine_cost`.
`usage_paper_count`	Usage: number of arXiv papers in the ChatPD usage table that cite the dataset. Cumulative as of the December 2025 cutoff.
`usage_velocity`	Per-year adoption: `usage_paper_count` divided by the dataset's age in years (with age clipped to ≥ 1 yr). Used as the primary adoption outcome in the multivariate regression.
`citation_velocity`	Per-year visibility: `total_usage_citations` divided by the dataset's age in years. Used as the primary citation outcome in the multivariate regression.
`max_cited_paper.citationCount` (“`max_paper_cite`”)	Headline citation-impact metric: S2 citation count of the *single* most-cited downstream paper that uses the dataset. Reported instead of any top-k aggregate.
`cite_1yr`	Window-restricted variant: total S2 citations accumulated by all using papers within the calendar interval $[\text{release\_year}, \text{release\_year} + 1]$ (cap = 2025). Used for cross-sectional descriptive averages (e.g. collaboration mix).
`use_1yr`	Window-restricted variant: number of arXiv papers that cite the dataset within the calendar interval $[\text{release\_year}, \text{release\_year} + 1]$. Companion adoption metric to `cite_1yr`.

Supplementary Table S7: Sample accounting (Part 1 of 2): main canonical corpus. Per-stage drops reflect the filter that defines the next stage; readers should refer to this table for any apparent inconsistency in n across figures and regressions. All "canonical" counts use the alias canonicalization pipeline of Supplementary Section S1.10. The 24K high-fidelity subset used for training is documented in Table S8.

Stage	n	**Filter applied to the previous row**
Raw ChatPD enriched rows	173,369	ChatPD paper–dataset mapping after Stage 1 enrichment (creator, impact, research domain, license, origin); no further filter.
Canonical datasets (alias-collapsed)	∼155,000	Group by normalized `dataset_name`; keep one canonical entry per alias group (Supplementary Section S1.10).
with finite per-year time slot	164,961	Drops aliases with no resolvable release year. Used in Fig. 1**a**.
with finite license, origin, data-type *and* year FE	154,576	Adds: all five regression covariates non-null. Used in Eq. 4 and `cite_1yr` / `use_1yr` averages.
with author count	159,194	Adds: parseable author list. Used in Fig. S16.
with finite citation & usage Lorenz inputs	155,675	Adds: both `total_usage_citations` and `usage_paper_count` non-zero. Used in Fig. S17.
with finite kNN embedding density	151,243	Adds: SBERT summary embedding successfully produced and ≥ 20 year-strictly-earlier neighbors available. Used in Fig. S9 and the topic-heat paragraph.
with creator-side coverage (imputed)	146,875	Restricts to canonical datasets with creator institution either explicit (`creator_institutions`) or imputable as the institution of the earliest arXiv paper that references the dataset. Used in Fig. 2**c, d**.
multi-author subset	145,578	Drops single-author rows (so "cross-sector / single-country" labels are well-defined). Used in the collaboration-mix bars.
with timing-position percentile	143,646	Drops rows whose kNN density estimate is missing, restricting to canonical datasets with a finite timing percentile. Used in Fig. 2**e**.
income-classified (imputed creator)	106,811	Adds: creator-country imputation maps to a World Bank class. Used in Supplementary Fig. S14 (bottom row, HI / LOW / HI+LOW lift discussion).
income-classified (explicit creator only)	28,062	Restricts to rows where `creator_institutions` is populated by the Stage-1 pipeline (no imputation). Used as the conservative comparison to the imputed view.

Supplementary Table S8: Sample accounting (Part 2 of 2): 24K high-fidelity subset (training core). Companion to Table S7; per-stage drops follow the same convention.

Stage	n	**Filter applied to the previous row**
24K candidate set (usage ≥ 2, ChatPD)	23,454	Filter on the 173K: usage ≥ 2 for grounded extraction.
fully extracted (24K)	22,084	Sonnet 4.5 PDF extraction + Gemini 2.5 Pro web grounding both succeeded (94% completion). Used in the cost-estimation supervised target pool.
with maskable numeric slots (initial regex pass)	13,338	Subset whose grounding text contained at least one number passing the regex/filter cascade. Yields the slot test pool.
usable for slot-filling training (final)	13,600	Same numeric-slot pool re-counted after the Opus role-tag pass (some slots gained, some excluded as cost/metadata mis-tags); the small difference relative to 13,338 reflects the role-tag re-screen, not a separate sample.
Number Imputation SFT train split	10,861	SHA-256 hash split on `dataset_name`; train ($\sim$80%).
Number Imputation SFT val / test	1,372 / 1,367	Same hash bucketing; val / test ($\sim$10% each).
Number Imputation SFT samples (post augmentation)	11,117	Train split after the multi-mask augmentation that emits one sample per maskable slot configuration ($+\sim$2% relative to the 10,861 dataset-level split).
Cost Estimation SFT samples (Opus distillation)	17,780	Each retained 13,600 dataset produces ~ 1.3 distilled response on average; downstream of slot-filling, shared dataset hash split.

S1.12 Tables

Supplementary Table S9: Per-dimension breakdown on anonymized test (1,367 samples). Bold indicates GRPO improvement over SFT. Companion to the cost-estimation weighted-average score table reported in the main Methods section.

Dimension	**Opus 4.5**	**Qwen-32B GRPO**	**Qwen-32B SFT**	**GPT-5-nano**	**GRPO vs. SFT**
Cost separation clarity	0.974	0.961	0.951	0.889	+0.010
Machine cost completeness	0.835	0.729	0.710	0.531	+0.019
Human-hour cost completeness	0.869	0.782	0.735	0.481	**+0.047**
Comprehensiveness	0.781	0.669	0.637	0.425	**+0.032**
Accuracy	0.633	0.504	0.409	0.240	**+0.095**
Calculation transparency	0.902	0.862	0.753	0.314	**+0.109**
Non-hallucination	0.490	0.443	0.362	0.212	**+0.081**
Reasonableness of assumptions	0.665	0.626	0.504	0.238	**+0.122**
Evidence quality	0.727	0.609	0.512	0.239	**+0.097**
Weighted average	**0.764**	**0.688**	**0.619**	**0.397**	**+0.069**

Supplementary Table S10: Extraction schema (Part 1 of 2): metadata that applies to the full **173K** ChatPD collection. Coverage figures are audited against the actual JSON dumps. The deep extractions for the 22,084 records inside the 24K subset are documented in Table S11.

Category	Field	Coverage	Description
ChatPD Metadata	`summary`	97.3%	Dataset summary from ChatPD catalog
	`tasks`	98.5%	Associated ML tasks
	`data_type`	98.2%	Modality (text, image, tabular, . . .)
	`scale`	69.6%	Size description
	`usage_paper_count`	100%	Usage: number of arXiv papers that use the dataset
Enriched Metadata	`arxiv_earliest_year`	99.2%	Year of first arXiv paper that uses the dataset (fallback time axis)
	`creator_paper_title`	3.0%	Verified title of the creator paper (only set when the pipeline's title-verifier agrees)
	`creator_paper_arxiv_id`	3.0%	Verified arXiv ID of the creator paper; restricted to rows where `creator_paper_title` is set
	`creator_authors`	3.0%	Authors of the verified creator paper
	`creator_institutions`	2.9%	Institutions of the verified creator paper
	`total_usage_citations`	100%	Citation impact: aggregate S2 citation count of all using papers
	`top_usage_authors`	98.9%	Authors across top using papers
	`top_usage_institutions`	93.2%	Institutions from top using papers
	`top_usage_papers`	99.3%	Highest-cited downstream papers retained as the source of `max_cited_paper.citationCount` (the headline impact metric used throughout the paper)
Tags	`origin`	91.6%	Construction provenance: *real-world* (passive collection), *human-generated* (paid annotation / writing), *inherited* (re-packaging an earlier corpus), *synthetic* (sim / LLM-generated)
	`phase`	99.9%	Role in the LLM training pipeline (Opus 4.5 batch over 159K canonical): *PRE*, *POST*, *EVAL*, *DOMAIN*, *OTHER*
	`subdomain`	99.4%	One of 25 research subdomains (CV-recognition, NLP-understanding, robotics, . . .)
	`license_type`	84.8%	Distribution restriction: *open* / *research-only* / *closed* / *unknown*
	`income_class`	14.1%	World Bank class of verified creator institution country: *HI* / *MID* / *LOW* / *Unknown* (low coverage reflects creator-attribution rate, not classification rate)

Supplementary Table S11: Extraction schema (Part 2 of 2): three hierarchical levels and the cost-analysis rubric, deeply extracted only for the 22,084 complete records inside the **24K** subset (the evidence-grounded training core; full 173K-coverage fields are in Table S10).

Category	Field	Coverage	Description
Level 1: Creation	`paper_name`	97.8%	Original publication title
	`publication_year`	97.0%	Dataset creation year (verified)
	`authors`	96.2%	Creator name, email, institution
	`primary_institution`	51.5%	Lead organization
	`conference`	89.6%	Venue or repository
	`funding`	18.8%	Grant or funding sources
	`research_domains`	99.3%	Research fields (list)
Level 2: Construction	`construction_method`	89.0%	Data collection procedure
	`labeling_method`	74.5%	Annotation approach
	`source` / `source_link`	92.5% / 28.1%	Data origin and URL
	`common_usage`	97.7%	Application domains
	`tasks`	97.2%	ML tasks (classification, etc.)
	`scale`	85.7%	Size indicators (e.g., "70B tokens")
	`dataset_lineage`	45.1%	Derivation from other datasets
	`timespan`	11.7%	Temporal coverage
Level 3: Detailed	`description`	99.8%	Full dataset narrative
	`Instance Count`	39.7%	Number of samples
	`Human Annotation Involved`	57.9%	Binary annotation flag
Cost Analysis	Rubric: `human`	35.4%	Human labor sub-rubric
	Rubric: `equipment`	34.2%	Equipment cost sub-rubric
	Rubric: `coding`	32.8%	Development cost sub-rubric
	Rubric: `labelling`	19.9%	Annotation cost sub-rubric
	`cost` (explicit)	0.6%	Explicit author-reported cost figure in the paper. The 99.4% of records without this field carry only *evidence-grounded extractions* (grounded numeric slots and qualitative cost language); we therefore use "evidence-grounded supervision" rather than "ground-truth cost" to describe the training signal in the rest of the paper.

Supplementary Table S12: Definitions of the categorical tags introduced in this paper. Each axis is populated by a dedicated Claude Opus 4.5 classification prompt over the ChatPD title, abstract and the Level 2 construction-text snippets (with a GPT-4.1-mini web-search fallback only for unresolved licenses); the definitions below are the category descriptions supplied to those prompts. Origin, research domain, license and income class are the four *Tags* axes (Table S10); the three skill tiers are a deterministic aggregation of the research-domain taxonomy used in Fig. 1**f**. Origin is single-label, with ties broken human > inherited > real-world > synthetic.

Axis	**Category**	**Definition given to the classifier**
Origin	real-world	Passively or actively collected from a physical or human source (e.g. ImageNet, MS COCO, clinical scans).
	simulated & synthetic	Algorithmically or model-generated without a real-world referent (procedural data, GAN/LLM outputs, agent-rollout traces).
	inherited	Derivative of an upstream corpus via re-sampling, re-labeling or re-processing (MMLU-Pro on MMLU, FineWeb-Edu on Common Crawl, standard splits of a pretraining corpus).
	human	Produced primarily through paid annotation, expert writing or crowd-task design, where the artifact does not exist before the task is run (SQuAD, HumanEval, instruction-following corpora).
Research domain	25 categories	CV-recognition, NLP-understanding, NLP-generation, multimodal, medical, mathematical reasoning, physical sciences, robotics, code, bioinformatics, astronomy, earth sciences, computational social science, cybersecurity, finance, education, sports, agriculture, transportation, recsys, graph, RL, speech, time-series, and "Other". The classifier returns one primary domain, spot-checked against the ChatPD `tasks` list.
License type	open-source	Redistribution and derivative use permitted (permissive or copyleft open licenses).
	research-only	Use restricted to non-commercial or research purposes.
	closed-source	Proprietary, gated or otherwise not openly redistributable.
	unknown	No license extractable from paper or dataset card (resolved where possible by a GPT-4.1-mini web-search fallback).
Income class creator-country World Bank class	HI	All creator institutions are in high-income countries.
	LOW	All creator institutions are in lower-income (non-high-income) countries.
	HI+LOW	A team mixing creators from high-income and lower-income countries.
Skill tier	crowd	General labor: CV-recognition, NLP-understanding, RL, speech, time-series, graph, transportation, agriculture, "Other".
	domain	Specialist non-PhD: NLP-generation, multimodal, robotics, cybersecurity, finance, education, sports.
	frontier	PhD / research-grade: medical, mathematical reasoning, physical sciences, bioinformatics, astronomy, earth sciences, code, computational social science.

Supplementary Table S13: Number-imputation dataset statistics (24K subset). Values span > 10 orders of magnitude, motivating logarithmic evaluation. Moved from the main Methods section.

Statistic	**Value**
Total samples (with maskable slots)	13,338
Train / Val / Test	10,861 / 1,372 / 1,367
Total number slots	105,309
Mean slots per sample	7.9
Median slots per sample	6
Max slots per sample	92
Value magnitude distribution	
[0, 1)	1,599 (1.5%)
[1, 10)	24,998 (23.7%)
[10, 100)	27,175 (25.8%)
[100, 1K)	18,142 (17.2%)
[1K, 1M)	29,629 (28.1%)
[1M+)	3,766 (3.6%)
Geometric mean of all values	209
Median value	85

Supplementary Table S14: LLM-judge rubric dimensions and weights (cost-estimation evaluation). Critical dimensions receive 2× weight. Moved from the main Methods section; companion to Tab. S9 which shows the per-dimension model scores under this rubric.

Dimension	**Weight**
Cost separation clarity	1.0
Machine cost completeness	1.0
Human-hour cost completeness	1.0
Comprehensiveness	1.5
Accuracy	2.0
Calculation transparency	1.0
Non-hallucination	2.0
Reasonableness of assumptions	1.5
Evidence quality	2.0
Total	13.0

Supplementary Table S15: Per-category shares for the two cost tracks, parsed from the 173K GRPO inference responses via Claude Opus 4.5 classification. The financial track uses a resource-oriented taxonomy ($n = 131,185$ datasets with parsable cost tables); the human-hour track uses an activity-oriented taxonomy ($n = 129,650$ with parsable hour tables). Shares within each track sum to 100%; the two taxonomies differ because money is spent on *things* while labour is spent on *actions*.

Financial cost taxonomy (resources)

Category	Share (%)
other financial	43.0
compute & cloud	33.5
software & tools	11.0
equipment & sensors	7.1
licensing & data access	3.4
storage & hosting	1.6
legal & compliance	0.2
travel & events	0.2

Human-hour taxonomy (activities)

Category	Share (%)
other human-hour	16.7
annotation & labelling	14.0
quality control & validation	12.9
data collection & scraping	12.0
data cleaning & preprocessing	11.8
software engineering	11.3
documentation & release	9.3
design & planning	8.4
project management	3.3
domain expert time	0.3

Supplementary Table S16: Focused correlations separating 1st-order *usage* (number of papers that cite the dataset within one calendar year of the creator paper's release, $[Y, Y+1]$) from 2nd-order *citation* impact (S2 citations of those papers in the same window). All $n = 154,576$ datasets from 2010–2025. ρ_S is Spearman (rank-based, headline); r_P is Pearson on $\log_{10}$-transformed values (non-rank but sensitive to magnitudes; raw-Pearson on these power-law-skewed quantities is dominated by the upper tail and uninformative). The p-value column is for Spearman.

Independent → Dependent variable	ρ_S (rank)	r_P ($\log_{10}$)	n	p-value (ρ_S)
Financial cost → 1-yr usage count	−0.003	0.049	154,576	1.7×10^{-1}
Hour cost → 1-yr usage count	−0.010	0.057	154,576	1.2×10^{-4}
Financial cost → 1-yr citation count	−0.000	0.020	154,576	9.1×10^{-1}
Hour cost → 1-yr citation count	−0.003	0.016	154,576	2.1×10^{-1}
Author count → financial cost	0.165	0.142	154,576	$< 10^{-300}$
Author count → hour cost	0.172	0.165	154,576	$< 10^{-300}$
Financial cost → Hour cost (cross-track)	0.704	0.588	154,576	$< 10^{-300}$

Supplementary Table S17: Cost preferences of industry-only, mixed and academic-only dataset-creating teams (2010–2025). Median financial and median human-hour costs are reported separately; the last column is the share of datasets where the labour cost exceeds the financial cost.

Institution class	n	Fin. median ($k)	Hours median ($k)	% human-dominant
industry	2473	22.5	0.5	8.2
mixed	6507	18.0	0.4	7.4
academic	20208	13.8	0.3	8.1

Supplementary Table S18: Case-study neighbourhood: AIME 2024 vs. peer LLM-reasoning benchmarks. *Usage* is the cumulative number of arXiv papers (as of the December 2025 cutoff) that cite the dataset; *Citation sum (cumulative)* is the corresponding cumulative S2 citation count of those using papers in the same window. Both quantities use the same temporal window, so citation sum ≥ 0 for every row and no derived "top-k" subset is reported (the headline impact metric used elsewhere in the paper, `max_cited_paper.citationCount`, is consistent with these cumulative aggregates but not shown in this table because it is reported per-dataset in the main regression). AIME 2024 ranks at the top of the citation-to-cost ratio in this cohort despite having only 122 using papers, because at least one downstream paper (Kimi k1.5) accumulated a very large citation count during 2025.

Dataset	Year	Mach. ($k)	Hours	Usage	Citation sum (cum.)	Origin
AIME 2024	2025	0.8	26	122	2,061	human
MMLU-Redux	2025	7.5	7,375	13	2,131	inherited
BBH	2022	23.8	280	195	7,247	human
MT-Bench	2023	13.7	324	253	4,907	human
GPQA	2023	9,128.8	5,298	416	3,938	human

S1.13 Numerical companions to the Results section

The tables below consolidate the per-year, per-class and per-domain numerical values referenced throughout §2.1–§2.2; the main text describes the qualitative direction and magnitude class only, and points the reader here for exact values.

Supplementary Table S19: Per-dataset machine cost (USD) and human-hour cost (h), median and P99-winsorized mean. Companion to §2.1 (1a) and Fig. 1**a**; $n = 164{,}961$ canonical datasets.

Statistic	**2018**	**2025**	**Shape**	**2024→2025**
Median machine cost (USD)	\$9.2k	\$12.4k	flat 2019–2022, dip 2023–2024	+35%
Median human-hour cost (h)	240h	300h	flat 2019–2022, dip 2023–2024	+28%
P99-winsorized mean (\$, shape)	—	—	flat 2018–2024	+30%

Supplementary Table S20: Total annual investment year-on-year decomposition: count effect vs. per-dataset (median) effect vs. total \$ effect. Companion to §2.1 (1b) and Fig. 1**b**. 2018→2025 headline: count ∼18×, total \$ ∼20× (\$3B→\$64B), total h ∼15× (124M h→1.85B h), on ∼3,600 → 65,700 annual releases.

Year	**Count (YoY)**	**Per-dataset median (YoY, \$ / h)**	**Total \$ (YoY)**
2023	+41%	−0.4% /—	+41%
2024	+243%	−5.9% /—	+223%
2025	+59%	+34% / +28%	+113% (≈1.59×1.34)

Supplementary Table S21: Per-dataset size, mega-corpus shares, and the single largest annual release. Companion to §2.1 (1c) and Supplementary Fig. S1. The per-percentile mean-cost rise of ∼ +24% over 2018–2025 is concentrated in the middle (50–90th) and bottom-50% segments; top-1% per-bucket means are chaotic with no trend.

Quantity	**2018**	**2025**
Median release size (samples)	∼20K	∼19K
Mean release size (samples)	2.98B	31.6B
≥1B-sample share	1.6%	3.4%
≥10B-sample share	0.3%	1.8%
Single-largest dataset of year (samples)	3T	240T

Supplementary Table S22: Per-origin 2022→2025 median per-dataset cost and human-hour cost. Companion to §2.1 (2) and Fig. S2.

Origin	**Cost (\$) 2022→2025**	**Δ%**	**Hours 2022→2025**	**Δ%**
Inherited	\$7.2k → \$11.3k	+56%	205h → 300h	+46%
Real-world	—(rebound after 2024 trough)	+28%	—	+18%
synthetic	—(∼1/3 of real-world)	+24%	—	+24%
Human	—(flat)	−3%	—	−15%

MMLU-Pro inherited-tier example: 12,032 inherited questions plus 588 newly authored, with retained items passing a 10-expert review, distractor augmentation to 10 options, and CoT-stability filtering.

Supplementary Table S23: Per-category attribution of the 2024→2025 per-dataset increment (P99-winsorized; Fig. 1**d**). Companion to §2.1 (3).

Bucket	**Share of increment**
Compute & cloud, share of non-labor $ increment	65%
Compute & cloud + equipment + licensing + software, share of total $ increment	68%
Annotation + QC, share of h increment (panel d)	62%
Annotation + QC + data collection, share of h increment	76%

Supplementary Table S24: 2025 venture-capital flow into AI data-supply firms. Companion to §2.1 (1), expert-rebuild mechanism. Aggregate AI-startup VC commitment in 2025 was at a record high (PitchBook data reported by Bloomberg News, 3 Oct. 2025). Sources are public news reports, cited by outlet and date rather than in the reference list.

Firm	**Deal / Round**	**Valuation**	**Source**
Scale AI	∼$14.3B Meta investment	∼$29B	Reuters, 13 Jun 2025
Scale AI	∼$200M Google human-label contract	—	Reuters, 13 Jun 2025
Turing	∼$300M revenue, profitable (OpenAI/Google/Anthropic/Meta)	—	Reuters, 28 Jan 2025
Mercor	$350M Series C	$10B	TechCrunch, 27 Oct 2025
Surge AI	raising toward $1B	$15B+	Reuters, 1 Jul 2025

Supplementary Table S25: 2024→2025 hyperscaler capital-expenditure step-change (company-level detail for §2.1 (3)). The combined ∼$96B year-on-year increment is the macro backdrop for the 2025 reversal of the compute & cloud commoditization trend; we report these temporal co-occurrences without claiming causation. The Meta and Alphabet ranges were revised upward in mid-year disclosures.

Firm	**2025 capital-expenditure increment**	**vs. 2024**	**Source**
Microsoft	∼$36B	∼80%	MS blog, 3 Jan 2025
Meta	∼$21–26B	∼55–66%	Q4'24 call, 29 Jan 2025
Alphabet	∼$22B	∼43%	Q4'24 8-K, 4 Feb 2025
Amazon	∼$17B	∼20%	Q4'24 call, 6 Feb 2025
Combined	∼$96B		

Pioneer-vs-mature `cite_1yr` gap ∼3.4×. Dataset releases sit ∼5 percentage points away from the pioneer bin relative to a same-procedure arXiv-paper baseline (Supplementary Fig. S14, top row). Income-class HI+LOW collaborations: ∼2.6× `cite_1yr` lift over single-income classes on the explicit-only sample, reinforced under the ∼4× converged sample (same figure, bottom row). Median MiniLM density z-score per cost × impact quadrant: high-impact +0.18 / +0.22; low-impact −0.05 / +0.02.

Supplementary Table S26: Per-domain 2024→2025 $/dataset increment. Companion to the research-domain heterogeneity paragraph of §2.1 and Supplementary Fig. S3. Whole-sample overall increment is +$2,469/dataset.

Domain	**Δ$/dataset**	**Dominant component**
Medical & clinical AI	+$3,799	equipment (IRB + scanners)
CV-recognition	+$3,201	compute
Physical sciences	+$2,531	compute
NLP-understanding	−$124	cheap derivative / synthetic dilution
Multimodal learning	−$1,069	cheap derivative / synthetic dilution

Supplementary Table S27: Topic-heat effect sizes and low-cost / high-impact benchmark list. Companion to §2.2 (2) and Supplementary Fig. S9.

Quantity	**Value**
Embedding corpus n	151,243 canonical datasets
Topic-heat main effect on impact	+0.218 z
Cost main effect on impact	+0.075 z
Topic-heat / cost ratio	∼3×
Low-cost / high-impact benchmark cost range	$0.6k–$10k (POPE, AMC, AdvBench, OSWorld, ScreenSpot-Pro, AlpacaEval)
papers within first year	32–93
Median density-z gap, high-vs-low impact	∼0.17–0.22
Mean `cite_1yr` drop, pioneer→mature	∼3.4×
Mean `use_1yr` drop, pioneer→mature	nearly flat

Supplementary Table S28: Class-level mean `cite_1yr` and `use_1yr` by sector mix, country mix, and topical timing position. Creator-side imputed attribution (n=146,875, ∼85% corpus coverage; n=145,578 multi-author after dropping single-author releases). Companion to §2.2 and Fig. 2**c**–**e**. Note that most datasets have a single early adopter, so `use_1yr` means near unity reflect the typical case.

Axis	**Class**	n	**Mean `cite_1yr`**	**Mean `use_1yr`**
Sector mix	Academic-only	129,579	6.81	1.04
	Industry-only	9,887	17.40	1.02
	Cross-sector	6,112	25.19	2.01
Country mix	Single-country	98,175	8.23	1.04
	Multi-country	10,421	19.57	1.76
Timing (K=20 MiniLM nbrs)	Pioneer (≤1/3)	64,948	10.32	1.04
	Middle	63,158	5.99	1.14
	Mature (≥2/3)	15,540	3.04	1.07

Supplementary Table S29: Time-varying regression coefficients from Eq. 4, fit year-by-year on velocity outcomes, 2010–2025 (each year fit on its own release cohort; $n = 154{,}576$ canonical datasets pooled across years). Companion to §2.2 (1) and Fig. 2**a,b**.

Coefficient	**Value**
$\lvert\hat{\beta}(\log \text{cost})\rvert$, full window	rarely exceeds 0.05
$\hat{\beta}(\log \text{authors} \to \text{cite\_velo})$, 16-year range	+0.10 to +0.36
mean level	$\sim$+0.20
2024–2025 LLM cohort	$\sim$+0.27
$\hat{\beta}(\log \text{authors} \to \text{use\_velo})$, 2010–2019	$\sim$+0.10–0.13
2024	+0.037
2025	+0.017
2010s→LLM-cohort weakening	$\sim$5–7×
Controlled multi-country / HI+LOW `cite_1yr` premium (net of author count, year and topic; cf. $\sim$2.4–2.6× descriptive, Table S28)	$\sim$1.27–1.29×

S1.14 LLMs used in the pipeline

Table S30 consolidates every language model invoked by the SoD pipeline, the stage at which it is used, and the precise role it plays. Two design rules are followed throughout the project: **(i) teacher / judge separation**: the distillation teacher (Claude Opus 4.5) is never used as the GRPO judge, and the judge (GPT-5-mini) is chosen as a non-Claude, non-Qwen model to reduce self-preference bias; and **(ii) one-shot heavy lifting, lightweight inference at scale**: frontier models (Sonnet 4.5, Gemini 2.5 Pro, Opus 4.5) appear once per dataset during Stage 1 enrichment, while the trained Qwen3-32B GRPO model carries the full 173K Stage 3 inference load.

Supplementary Table S30: Inventory of LLMs used in the SoD pipeline. *Teacher* = supplies SFT distillation targets; *Judge* = scores GRPO rollouts; *Baseline* = comparison-only, not part of the production pipeline.

Model	**Stage & role**	**Where used**
Claude Sonnet 4.5	Stage 1 (extractor)	Reads dataset paper PDFs as multimodal input and emits the L1 / L2 / L3 metadata fields across the full 173K collection, plus the Cost-related grounding block on the 24K subset.
Gemini 2.5 Pro	Stage 1 (web grounding)	Google-search-grounded retrieval of dataset cards and missing metadata; complementary view to Sonnet 4.5.
Claude Opus 4.5	Stage 1 (schema, tags) and Stage 2 (teacher)	(i) Iteratively refines the L1–L3 extraction schema; (ii) role-tag re-screen of regex slot candidates (cost vs metadata); (iii) origin and research-domain tagging for the 173K corpus; (iv) distillation teacher for Cost-Estimation SFT (Step 2 of Stage 2).
GPT-4.1-mini	Stage 1 (license fallback)	Web-search resolution of *unknown* license cases left over after the Opus 4.5 pass.
GPT-5-mini	Stage 2 (judge)	Nine-dimension rubric judge for both the held-out cost-estimation evaluation (Tab. 3) and the GRPO reward (Eq. 3). Chosen as a non-Claude, non-Qwen model to reduce self-preference bias.
Qwen3-32B	Stage 2 (trained model) and Stage 3 (inference)	Base model for the three-step Number-Imputation SFT → Cost-Estimation SFT → GRPO pipeline. The final GRPO checkpoint is the production Cost Estimation Model and is run over the full 173K collection in Stage 3 (§4.5).
Baselines (comparison-only, not in production):		
Qwen3.5-122B	Baseline	Number-imputation comparison (Tab. 2).
GPT-5-nano	Baseline	Number-imputation + cost-estimation comparison (Tabs. 2, 3).
GPT-4o-mini	Baseline	Number-imputation comparison (Tab. 2).

S1.15 Why number imputation transfers to cost estimation: a case study

The pipeline order (Number Imputation SFT → Cost Estimation SFT → GRPO) is not arbitrary: number imputation acts as a domain-relevant reasoning prior that the cost-estimation head inherits. We illustrate the transfer with a single representative slot from the 24K subset, a paper that reports " *annotated 50,000 image–caption pairs at $0.12 per caption with ∼15 min per image batch* " with the three numbers masked. To recover `50000`, `0.12` and `15` the model must (i) identify each number's unit (samples, USD/unit, minutes), (ii) recognize that $\$0.12 \times 50{,}000$ is the dollar-side total (∼$6k),

(iii) recognize that 15 min $\times$ (image batches) $\times$ annotator count is the hour-side total, and (iv) check that the two are order-of-magnitude consistent under the implicit \$/h ratio for crowd annotation. A correct answer on the imputation task therefore requires exactly the structured numeric reasoning that cost estimation depends on: unit-aware scale parsing, additive composition across categories, and a cross-check between the \$ and h tracks. Empirically the transfer is large at every checkpoint that we evaluated: a Qwen3-32B base model without imputation SFT scores 0.31 on cost estimation (anonymized); after imputation SFT alone it scores 0.55 (a +24 percentage-point gain) before any cost-estimation training, and the subsequent Cost Estimation SFT adds a further +6.4 percentage points on top of that starting point. The same chain on a model that skipped the imputation stage gains only +13 percentage points under matched Cost Estimation SFT. We read this as direct evidence that number imputation supplies the unit-and-magnitude reasoning the cost head needs, rather than acting as a generic warm-up.

S1.16 Training configuration

The Qwen3-32B backbone is fine-tuned with 4-bit NF4 quantisation (QLoRA) at LoRA rank 16 and $\alpha = 32$, on 2×8 A100-80GB nodes orchestrated by Kubernetes; GRPO runs for 100 steps ($\sim$11h45m wall-clock). Rollouts are served by vLLM with tensor-parallel TP $= 4$ and $n = 4$ rollouts per prompt at learning rate 5×10^{-7}. The GRPO validation reward increases monotonically across all 100 steps ($0.570 \rightarrow 0.627$), indicating the model has not yet saturated and that further training may yield additional gains. Across the full pipeline (number-imputation SFT, cost-estimation SFT, GRPO), the base model improves from $\sim$0.30 on the judge rubric to 0.688 on the anonymized test, a $\sim$2.3$\times$ gain. The full step counts of the SFT and GRPO stages, optimizer state, dataloader sharding, judge-API batching parameters, and the random seeds required to reproduce this trajectory are committed in the public code release pointed to by the *Code availability* statement; the verl GRPO framework is used unmodified.

S1.17 Pipeline prompts (reproducibility)

For full reproducibility we reproduce below the verbatim language-model prompts used at each stage of the SoD pipeline described in the Methods, grouped by function: **extraction** (Stage 1 metadata recovery), **tagging** (the categorical axes of Table S12), **estimation** (the cost/hours analysis that the Cost Estimation Model is trained to reproduce), and **judging** (the LLM-judge rubrics used for the GRPO reward and the held-out evaluation). Each box is colored by group. Placeholders in braces (e.g. `{dataset_name}`) are filled at run time; JSON skeletons are reproduced exactly as sent to the model. In wrapped lines a trailing backslash marks a soft line break inserted for the page width, not a character in the prompt.

S1.17.1 Extraction prompts

Prompt 1.1 — Claude Sonnet PDF extraction (L1/L2/L3 on 173K; cost block on 24K)

```
Analyze the provided research paper PDF to extract comprehensive, grounded, hierarchical \
    information for the dataset: "{dataset_name}".

**CRITICAL INSTRUCTIONS:**
1.  Extract information **ONLY** for the dataset named "{dataset_name}".
2.  For **EVERY** field, you **MUST** provide a `grounding` quote from the PDF that directly \
    supports the extracted `value`.
3.  If you cannot find information for a field, leave its `value` and `grounding` as `null`.
```

```
4.  Pay special attention to tables for statistics like ‘instance_count‘ and distributions. SUM \
    counts from train/test/validation splits.
5.  For ‘cost_analysis‘, be creative. Identify cost-related features relevant to THIS dataset. The\
     output should be a flexible JSON object where keys are the identified features (e.g., ’\
    License’, ’Human Annotation Involved’, ’Instance Count’).

Return **ONLY** a single, valid JSON object in this exact format:
‘‘‘json
{
    "name": "{dataset_name}",
    "description": {
        "value": "COMPREHENSIVE description (200-400 words) of the dataset’s construction, purpose\
    , and key characteristics, based on the paper.",
        "grounding": "Direct quote from the paper supporting the description."
    },
    "level1_info": {
        "paper_name": {
            "value": "Full paper title that introduced this dataset.",
            "grounding": "Quote supporting the paper title."
        },
        "conference": {
            "value": "Full conference/journal name with year (e.g., ’NeurIPS 2020’).",
            "grounding": "Quote supporting the conference."
        },
        "publication_year": {
            "value": "YYYY",
            "grounding": "Quote supporting the publication year."
        },
        "institution": {
            "value": "Institution that created the dataset.",
            "grounding": "Quote supporting the institution."
        },
        "research_domains": {
            "value": "[’List’, ’of’, ’research’, ’domains’]",
            "grounding": "Quote supporting the domains."
        }
    },
    "level2_info": {
        "timespan": {"value": "Time period data covers (e.g., ’2015-2020’)", "grounding": "..."},
        "source": {"value": "Original data source description", "grounding": "..."},
        "source_link": {"value": "URL to data source", "grounding": "..."},
        "construction_method": {"value": "How the dataset was built", "grounding": "..."},
        "labeling_method": {"value": "How labels were assigned", "grounding": "..."},
        "feature_distribution": {"value": "Description of feature characteristics", "grounding": "\
    ..."},
        "label_distribution": {"value": "Description of label characteristics", "grounding": "..."\
    },
        "preprocessing": {"value": "[’List’, ’of’, ’preprocessing’, ’steps’]", "grounding": "..."\
    },
        "common_usage": {"value": "[’List’, ’of’, ’common’, ’applications’]", "grounding": "..."},
        "tasks": {"value": "[’List’, ’of’, ’ML’, ’tasks’]", "grounding": "..."},
        "record_count": {"value": "Number of records/rows for tabular data", "grounding": "..."},
        "dataset_lineage": {"value": "Parent datasets if derived", "grounding": "..."},
        "scale": {"value": "Info on dataset scale (nodes, edges, etc.)", "grounding": "..."}
    },
    "level3_info": {
        "potential_applications": {"value": "[’List’, ’of’, ’real-world’, ’applications’]", "\
    grounding": "..."},
        "ethical_concerns": {"value": "[’List’, ’of’, ’ethical’, ’concerns’]", "grounding": "..."\
```

```
    },
        "limitations": {"value": "['List', 'of', 'dataset', 'limitations']", "grounding": "..."}
    },
    "cost_analysis": {
        "__comment__": "This is a flexible object. Populate it with relevant cost features for the\
     dataset, both qualitative (e.g., License, Human Annotation) and quantitative (e.g., Instance \
    Count). The keys should be the feature names. The examples below are for structure, do not use\
     them literally.",
        "License": { "value": "e.g., MIT", "grounding": "..." },
        "Human Annotation Involved": { "value": "e.g., true", "grounding": "..." },
        "Instance Count": { "value": "e.g., 150000", "grounding": "..." }
    }
}
'''
```

Prompt 1.2 — Gemini / web-search extraction (24K complementary view)

```
Search the web to find comprehensive, grounded, hierarchical information for the dataset: "{\
    dataset_name}".
```

The remaining structure is identical to Prompt 1.1, with three substitutions: instruction 1 becomes "*Synthesize information from reliable web sources (e.g., official documentation, PapersWithCode, research papers).*"; instruction 2's grounding requirement becomes "*you **MUST** provide a `grounding` snippet or URL that directly supports the extracted `value`*"; and every `grounding` description in the JSON becomes "*Supporting web snippet or URL.*"

S1.17.2 Tagging prompts

Prompt 2.1 — License retrieval (web search)

```
What is the license of the ML dataset "{dataset_name}"{url_hint}? Search the web. Respond with \
    ONLY the license name (e.g. "MIT", "Apache 2.0", "CC BY 4.0"). If not found, say "Not found".
```

`{url_hint}` takes a form such as " (homepage: <url>)" or " (source: <url>)", drawn from the extracted `dataset_url` / `homepage` / `source_link`. The reply's first line is used, with markdown and quotation marks stripped; "Not found" / "unknown" are treated as no result. The mapping from the returned license string to the open-source / research-only / closed-source / unknown buckets is a pure rule-based match (no LLM prompt).

Prompt 2.2 — Research-domain classification (25-way; Opus 4.5)

```
You are classifying a machine learning dataset into exactly one research domain.

Pick the SINGLE best-matching domain from this list:
1. CV-recognition
2. NLP-understanding
3. NLP-generation
4. multimodal
5. medical
6. mathematical reasoning
7. physical sciences
8. robotics
9. code
```

```
10. bioinformatics
11. astronomy
12. earth sciences
13. computational social science
14. cybersecurity
15. finance
16. education
17. sports
18. agriculture
19. transportation
20. recsys
21. graph
22. RL
23. speech
24. time-series
25. Other

Dataset info:
{context}

Respond with ONLY the domain name exactly as written above, nothing else.
```

`{context}` is the dataset's title, abstract and the L2 construction snippets. The returned primary domain is spot-checked against the dataset's ChatPD `tasks` field; on disagreement the prompt is re-issued with the conflicting tasks list appended. This 25-category taxonomy extends the earlier 17-way subdomain scheme and defines the `research_domain` axis of Table S12 and the skill tiers of Fig. 1**f**.

Prompt 2.3 — Origin, inherited sub-type, and label status (Opus 4.5)

```
1. ORIGIN (pick exactly ONE):
   - "human" - The dataset reflects the human mind and physical behaviour, social activity. \
     Examples: papers, citations, books, psychology experiments, surveys, social media posts, user-\
     generated reviews, medical patient records, transportation patterns, crowdsourced annotations,\
      speech recordings, legal documents, educational materials.
   - "real-world" - The dataset reflects objects or phenomena NOT related to humans. Examples: \
     animal behaviour, geography, astronomy observations, physics measurements, mechanical/\
     structural data, chemical compounds, biology specimens, weather/climate data, geological \
     surveys, satellite imagery of terrain, plant species.
   - "simulated" - The dataset is generated from algorithms or hardware. Examples: LLM-generated \
     text, simulation software output, 3D rendering, game engine data, signal from scientific \
     equipment, synthetic benchmarks, procedural generation, equation-based data, finite element \
     simulations.
   - "inherited" - The dataset is mostly built from other existing datasets, either reorganized (\
     cleaning, filtering, auditing, merging) or re-annotated with new labels.

2. If origin is "inherited", sub-classify:
   - "(re)-annotated" - New labels or annotations were added to existing data
   - "(re)-organized" - Existing data was cleaned, filtered, audited, merged, or reformatted \
     WITHOUT new annotations

3. LABEL STATUS:
   - "no-label" - No ground truth labels or annotations
   - "with-label" - Has ground truth labels or annotations
   - If "with-label": Are labels "inherited" (from source datasets) or "original" (newly created \
     for this dataset)?
```

```
4. GROUNDING:
   For EVERY classification decision and cost estimate, you MUST provide grounding - the original \
    sentence(s) from the PDF or web source that support your decision. Quote the source text \
    directly.
```

This classification block is embedded in the main cost-analysis prompt (Prompt 3.2); it is reproduced here separately because it defines the `origin` axis of Table S12.

S1.17.3 Estimation prompts

Prompt 3.1 — Cost-analyst system prompt (Opus 4.5)

```
You are an expert dataset cost analyst. You classify ML datasets and estimate their creation costs\
     with high accuracy. You always return valid JSON - no markdown fences, no commentary outside \
    the JSON object. You never default to $0 without strong justification. You use typical market \
    rates for APIs, compute, and services to produce realistic estimates. When a dataset has \
    labels that are NOT inherited, you explicitly break out labelling costs. For every \
    classification and cost estimate, you provide grounding by quoting the original sentence(s) \
    from the source PDF or web page.
```

Prompt 3.2 — Main analysis prompt: classification + cost estimation (Opus 4.5)

```
Analyze the dataset "{name}" and provide a comprehensive cost analysis and classification.

DATASET CONTEXT:
- Name: {name}
- Alternative names: {alt_names}
- Subdomain: {subdomain}
- Summary: {summary}
- Tasks: {tasks}
- Scale: {scales}
- Papers citing this dataset: {paper_count}

============================================================
CLASSIFICATION INSTRUCTIONS:
============================================================

1. ORIGIN (pick exactly ONE):
   - "human" - The dataset reflects the human mind and physical behaviour, social activity. \
    Examples: papers, citations, books, psychology experiments, surveys, social media posts, user-\
    generated reviews, medical patient records, transportation patterns, crowdsourced annotations,\
     speech recordings, legal documents, educational materials.
   - "real-world" - The dataset reflects objects or phenomena NOT related to humans. Examples: \
    animal behaviour, geography, astronomy observations, physics measurements, mechanical/\
    structural data, chemical compounds, biology specimens, weather/climate data, geological \
    surveys, satellite imagery of terrain, plant species.
   - "simulated" - The dataset is generated from algorithms or hardware. Examples: LLM-generated \
    text, simulation software output, 3D rendering, game engine data, signal from scientific \
    equipment, synthetic benchmarks, procedural generation, equation-based data, finite element \
    simulations.
   - "inherited" - The dataset is mostly built from other existing datasets, either reorganized (\
    cleaning, filtering, auditing, merging) or re-annotated with new labels.

2. If origin is "inherited", sub-classify:
   - "(re)-annotated" - New labels or annotations were added to existing data
```

```
   - "(re)-organized" - Existing data was cleaned, filtered, audited, merged, or reformatted \
     WITHOUT new annotations

3. LABEL STATUS:
   - "no-label" - No ground truth labels or annotations
   - "with-label" - Has ground truth labels or annotations
   - If "with-label": Are labels "inherited" (from source datasets) or "original" (newly created \
     for this dataset)?

4. GROUNDING:
   For EVERY classification decision and cost estimate, you MUST provide grounding - the original \
     sentence(s) from the PDF or web source that support your decision. Quote the source text \
     directly.

============================================================
COST ESTIMATION INSTRUCTIONS:
============================================================

SECTION A: FINANCIAL COST (USD ONLY)
Estimate ALL non-labor monetary costs to create this dataset.
Account for implicit infrastructure even when not explicitly stated.

Common categories:
- LLM API calls: GPT-4 ~$10-30/1M tokens, GPT-3.5 ~$0.50-2/1M tokens, Claude ~$3-15/1M tokens
- Cloud compute: GPU $1-3/hr (consumer), $5-15/hr (A100), CPU VMs $0.05-0.50/hr
- Data licensing/acquisition: $100-$10,000+
- Web scraping infrastructure: proxies, storage, bandwidth
- Crowdsourcing platform fees: 20-40% commission on top of worker pay
- Software/tool subscriptions: annotation tools, cloud storage, databases
- Data storage and hosting: S3 ~$0.023/GB/month

To estimate, use the dataset scale and methods described. E.g. if a dataset used GPT-4 to generate\
     10,000 samples at ~500 tokens each: 10,000 x 500 x $20/1M = $100.

SECTION B: HUMAN EFFORT (HOURS ONLY)
Estimate ALL human labor regardless of compensation.

Common categories:
- Data annotation/labeling: 30-120 samples/hr (text), 10-30/hr (images)
- Quality review/verification: 10-30% of annotation time
- Data collection and curation
- Data processing and cleaning
- Project management and coordination
- Training annotators, guideline creation
- Documentation and publication

LABELLING COST:
If the dataset is "with-label" AND labels are "original" (NOT inherited):
-> Explicitly break out the labelling financial cost (in Section A) AND labelling hours (in \
    Section B).
Set labelling_cost_usd and labelling_hours to the respective amounts.
If labels are inherited or dataset is no-label, set both to null.

CRITICAL RULES:
- Section A = ONLY financial costs (NOT labor-based)
- Section B = ONLY human hours (NOT financial)
- NO overlap between sections
- Do NOT convert hours to dollars or vice versa
- Provide NON-TRIVIAL estimates based on dataset scale and methods
```

```
- If the dataset used ANY LLM/AI API, you MUST estimate the API cost
- If the dataset required ANY compute, you MUST estimate compute cost

============================================================

Return ONLY valid JSON:
{
  "origin": "human|real-world|simulated|inherited",
  "origin_grounding": "Direct quote from PDF or web supporting origin classification",
  "inherited_type": "(re)-annotated|(re)-organized|null",
  "label_status": "no-label|with-label",
  "label_status_grounding": "Direct quote from PDF or web supporting label classification",
  "label_source": "inherited|original|null",
  "financial_cost_usd": {
    "items": [{"name": "...", "amount": 0.0, "calculation": "...", "grounding": "Direct quote \
    supporting this cost item"}],
    "labelling_cost_usd": null,
    "total_usd": 0.0
  },
  "human_effort_hours": {
    "items": [{"name": "...", "hours": 0.0, "calculation": "...", "grounding": "Direct quote \
    supporting this effort item"}],
    "labelling_hours": null,
    "total_hours": 0.0
  },
  "reasoning": "Brief explanation of origin, label, and cost decisions"
}
```

Prompt 3.3 — PDF-batch and web-search prefixes

```
[Prefix prepended when PDFs are attached]
I have attached {n} research paper PDF{s} that cite or describe this dataset. Synthesize \
    information from {all papers|the paper} into ONE unified report. Do NOT produce separate \
    analyses per paper - combine and cross-reference into a single consolidated analysis.

[Web-search path prefix (instructions = the system prompt of Prompt 3.1)]
Search the web for the ML dataset "{name}" in the "{subdomain}" domain. Find information about how\
     it was created, its construction methodology, labeling process, data sources, and any \
    documented costs or effort.
```

Prompt 3.4 — Cost-estimation training/inference prompt (SFT, GRPO, 173K inference)

```
[System]
You are a dataset cost estimation expert. Estimate the total creation cost (USD) and total human \
    hours for datasets. Always provide specific numeric estimates in your conclusion, using $0 and\
     0 hours when costs are negligible.

[User] ({details} = merged extraction text, truncated to 5500 characters)
Analyze the creation cost and human effort for the dataset described below.

Instructions:
1. Estimate the TOTAL one-time creation cost in USD. This is the cumulative cost to build the \
    entire dataset from scratch, not an annual or per-unit cost. Use $0 for freely available, auto\
    -generated, or synthetic datasets with negligible cost.
2. Estimate the TOTAL human hours required to create the dataset. This is the sum of all person-\
    hours across all contributors. Use 0 for fully automated datasets.
```

```
3. Provide detailed reasoning identifying specific grounding information (e.g., number of \
    annotators, wage, time per sample) from the provided details.
4. Always commit to a specific numeric estimate, even if approximate. Avoid vague conclusions like\
     'cost is minimal' without a number.

Dataset Details:
{details}

End your response with a Conclusion section in exactly this format:

### Conclusion:
- **Total Cost:** $X
- **Total Human Hours:** Y hours
```

Prompt 3.5 — Number-imputation / slot-filling ([NUM] recovery)

```
[Primary prompt]
Fill in the missing numbers. The text below has {num_slots} [NUM] placeholder(s).
Predict what original number each [NUM] was.

Text:
{masked_text}

Reply with ONLY {num_slots} number(s), one per line. No words, no explanation.
Example for 3 placeholders:
42.5
1000
7.3

Your answer:

[Stricter retry prompt]
IMPORTANT: Output ONLY numbers, nothing else.

Text with {num_slots} missing number(s) marked [NUM]:
{masked_text}

Output exactly {num_slots} line(s). Each line = one number. No text, no units, no explanation.

Answer:
```

S1.17.4 Judge prompts

Prompt 4.1 — GRPO in-loop judge (3 dimensions)

```
[System] You are an expert evaluator for dataset cost estimation reasoning. Always return valid \
    JSON.

[User]
You are an expert evaluator for dataset cost estimation reasoning.

Reference Grounding (The truth from the paper):
{grounding_text}

Model's Reasoning:
```

```
{solution_str}

Evaluate the Model's Reasoning ONLY against the Reference Grounding. Extra details are acceptable \
    if they do not contradict the reference. Penalize only when the reasoning conflicts with or \
    misstates the grounding.

1. **Comprehensiveness** (0.0-1.0): Coverage of all grounded cost factors (e.g., human effort, \
    hardware, data collection).
2. **Accuracy/Alignment** (0.0-1.0): Faithfulness to the facts in the grounding (sources, nature \
    of costs, qualitative/quantitative alignment).
3. **Non-Hallucination / Groundedness** (0.0-1.0): 1.0 means fully grounded. Do NOT penalize \
    reasonable additions; deduct only for contradictions or fabricated facts relative to the \
    grounding.

Return your evaluation ONLY as a JSON object:
{
    "comprehensiveness": <float 0.0-1.0>,
    "accuracy": <float 0.0-1.0>,
    "non_hallucination": <float 0.0-1.0>,
    "reasoning": "<brief explanation of scores>"
}
```

Prompt 4.2 — Nine-dimension held-out evaluation judge

```
You are an expert evaluator for dataset cost estimation reasoning.

Reference Grounding (Ground Truth from Paper):
{grounding_text}

Model's Reasoning (for dataset "{dataset_name}"):
{response_text}

EVALUATION TASK:
Score 0.0-1.0 on these 9 dimensions. Return ONLY a JSON object.

1. cost_separation_clarity
2. financial_cost_completeness
3. human_hour_completeness
4. comprehensiveness
5. accuracy
6. calculation_transparency
7. non_hallucination
8. reasonableness_of_assumptions
9. evidence_quality

Return JSON: {"cost_separation_clarity": ..., "financial_cost_completeness": ..., ..., "\
    average_score": ..., "reasoning": "..."}
```

The dimension weights reported in the main text (accuracy, non-hallucination and evidence-quality at 2.0; comprehensiveness and reasonableness at 1.5; the rest at 1.0; Table S14) are applied when aggregating the judge's per-dimension scores (Eq. 3), not inside the judge prompt itself.

S1.17.5 Notes

Note 5 — Implementation note

```
- The weighted judge rubric (weights 2.0 / 1.5 / 1.0) is applied at score aggregation, not inside \
    the judge prompt (Prompt 4.2).
```

References

[1] Le Ludec, C., Cornet, M. & Casilli, A. A. The problem with annotation: Human labour and outsourcing between france and madagascar. *Big Data & Society* **10**, 20539517231188723 (2023).

[2] International Labour Organization. World employment and social outlook 2021: The role of digital labour platforms in transforming the world of work. Tech. Rep., International Labour Organization, Geneva (2021).