Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The field grew, the programme did not: differential growth and substrate portability in artificial intelligence and education research, 2016–2026

Yu Miao[1]
Hengshui University
Mengting Zhao[2]
Hengshui University
Jianwei Yu[3]
Shandong Huayu University of Technology
Chang-Hyun Jin[4]
chjin@kgu.ac.kr

Kyonggi University

Systematic Review

Keywords: Growth-adjusted indicators, Field transition, Generative artificial intelligence, Learning analytics, Topic modelling, Research-front dynamics

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10857127/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Reviews of a field in transition routinely read a falling share of records as evidence of decline. Where annual output rises more than fivefold that inference fails: a research form can lose four-fifths of its share while publishing more papers each year than before. This study reconstructs the artificial-intelligence-in-education literature across a decade — 1,218 records from Scopus, Crossref and Google Scholar, screened under a PRISMA 2020-aligned protocol and split at ChatGPT's release into pre-generative (2016–2022, n = 333) and generative (2023–June 2026, n = 885) periods — converting every prevalence figure into an annualised count evaluated against pre-specified criteria with exact rate-ratio intervals. A growth-adjusted index G expresses each form's growth as a fraction of the field's own 5.32-fold increase in annual output; G is algebraically the prevalence ratio, so the share ratio is already growth-adjusted and the share difference is what misleads. The result is differential growth rather than displacement. Educational data mining contracts (rate ratio 0.31, 95% CI 0.15–0.58), as does learning analytics (0.67, 0.46–0.95), though that contraction is design-dependent and reverses under a retrieval exposing both periods to both vocabularies; intelligent tutoring is statistically flat (0.96, 0.64–1.42); three further forms grow at a fifth to three-fifths of the field's rate, one keeps pace, and four outgrow it. Manual coding of a stratified sample (n = 200, κ = 0.80) supplies the code-level series a term-based instrument cannot recover: constructs presupposing instrumented longitudinal data return G = 0.18 and 0.52, those that do not return 1.12, 1.15 and 2.67. Substrate portability, not functional definition, predicts survival.

Introduction

Artificial intelligence entered education long before ChatGPT. Through the 2010s the field was organised around learning analytics, educational data mining and intelligent tutoring systems — technologies concerned with observing learners, predicting difficulty and adapting instruction (Romero and Ventura 2020; Siemens and Baker 2012; VanLehn 2011). That work had dedicated journals, conferences and a methodological repertoire built on logged behavioural data (Aldowah et al. 2019; Blikstein and Worsley 2016). Since late 2022 a different technology has occupied the field's attention, and the volume of publication has risen steeply. Reviews have documented that rise (Bahroun et al. 2023; Crompton and Burke 2023; Lo 2023).

What they have not established is what happened to the earlier programme, and the reason is methodological rather than empirical. Rising totals are consistent with at least four histories: the earlier programme may have grown at the field's rate, so that its share is preserved; grown more slowly, so that its share falls while its output rises; stopped growing, holding output constant while the field expanded around it; or contracted outright. These cannot be told apart from the statistic that reviews of this literature conventionally report. A within-period share is a ratio whose denominator is itself the object of study. Treating a falling share as evidence of decline is a compositional error, and treating a stable share as evidence of stability is the same error in the opposite direction. The distinction is not a technicality: it separates a literature that has been abandoned from one that has merely been outgrown, and only the first would oblige a late-entering subfield to rebuild the capability for itself.

Scientific literatures grow exponentially as a matter of course (Bornmann and Mutz 2015; Price 1963), so any claim that a line of inquiry has declined must be made against the growth of its own field rather than against zero. That benchmark is easy to state and almost never applied. This study applies it. Prevalence is reported as a share, because share determines the composition of the literature a newcomer encounters; it is also converted to an annualised absolute count, because annualised counts determine whether the underlying activity continues; and the two are combined into a growth-adjusted index expressing how fast a form grew relative to the field. Criteria for contraction, stasis and survival are fixed before the estimates are read, and every comparison carries an interval estimate.

Three questions follow. How did output, knowledge structure and topic mass redistribute across the transition, in relative and in absolute terms? Which convergence forms grew, stalled or contracted once the field's own growth is the benchmark? And what accounts for the ordering among them? A fourth, applied question motivates the case: management education entered this literature only in 2022–2023 (Lim et al. 2023; Ratten and Jones 2023; Sollosy and McInerney 2022), roughly two years after the general education literature, and therefore inherited an already-formed problem framing. What a late entrant can import is determined by composition, and composition is exactly what the growth-adjusted analysis separates from activity.

The study contributes a measurement framework that separates absolute from relative change in a rapidly expanding literature with decision rules stated in advance; a revised convergence typology in which each construct carries an operational rule and a reported measurement status; a demonstration that the growth-adjusted index is algebraically the prevalence ratio, which clarifies which share statistic is safe to use; and a substrate-portability account of which research forms survive a change of technological substrate, corroborated on code-level data the account did not see.

Background

The pre-generative programme and the generative turn

The AI–education literature of the 2010s cohered around three related programmes. Learning analytics and educational data mining sought to make learning visible through behavioural traces, supporting diagnosis, prediction and early warning (Romero and Ventura 2020; Siemens and Baker 2012). Intelligent tutoring systems sought to approximate one-to-one instruction through model-based adaptation (VanLehn 2011). Adaptive and personalised learning sought to adjust pathways to the individual learner. Syntheses of the period describe the same triad and the same infrastructural dependency — instrumented platforms and longitudinal data — whether surveying techniques (Chassignol et al. 2018), the turn to large-scale behavioural data (Luan et al. 2020) the reconfiguration of learner–instructor interaction (Seo et al. 2021), the demands placed on teachers (Celik et al. 2022) or the empirical record in online higher education (Ouyang et al. 2022). That dependency divides the pre-generative repertoire along a line that is not the same as the line between techniques and functions, and the analysis below argues it is the more consequential of the two.

Generative systems removed the dependency. A study of ChatGPT's effect on assessment requires neither instrumented platforms nor longitudinal data, and the technology became simultaneously available to every instructor and student. The immediate literature was dominated by academic integrity and assessment security (Cotton et al. 2024; Perkins 2023; Sullivan et al. 2023), followed by institutional policy analysis (Moorhouse et al. 2023) and systematic assessments of practical and ethical risk (Yan et al. 2024). Appraisals appeared within months of release (Baidoo-Anu and Owusu Ansah 2023; Rudolph et al. 2023) and were followed by structured evaluations (Farrokhnia et al. 2024). Existing bibliometric treatments establish scale and map keyword clusters for the generative period (Bahroun et al. 2023; Chiu et al. 2023; Crompton and Burke 2023; Wang et al. 2024; Zawacki-Richter et al. 2019), and decade-scale reviews exist for the earlier period (Zhai et al. 2021), but none relates the two periods in growth-adjusted terms, because each is confined to one era, aggregates both without comparison, or compares them on shares alone.

A revised typology of convergence forms

Table 1 sets out seven convergence forms. Three features of its construction matter for measurement. First, *codes* and *indicators* are separated: a code is a theoretical construct with a record-level inclusion rule, an indicator is a term-level expression used to detect it, and the mapping is many-to-one and, for three codes, empty. Second, two widely used indicators — intelligent tutoring and conversational or pedagogical agent — name technology families that cut across several codes rather than measuring any one of them, and are reported as descriptive series outside the code system. An agent that answers a question instantiates augmentation (C2), not the negotiated joint problem-solving that defines human–AI teaming (C6). Third, automated scoring is assigned to substitution (C1) rather than to generation (C4), because marking is an evaluative act; where a system also composes the feedback text the record satisfies both codes and is resolved by a primary-form rule.

The measurement-status column reports the outcome of the validation exercise rather than an a priori judgement. Three codes — C1, C2 and C6 — are functional categories with no proprietary vocabulary, and the automated measure does not recover them; they are therefore measured by manual coding alone.

Table 1

Typology of AI–education convergence forms, with operational rules and measurement status

Code	Form	Record-level inclusion rule (principal exclusion in italics)	Term indicators	Measurement status (κ vs. coders)
C1	Substitution	A system performs, unassisted, a task previously performed by a human instructor or assessor. *Excludes systems that return information to a human who then acts.*	Automated scoring / grading (partial)	Not recoverable (− 0.13); manual only
C2	Augmentation	A system extends instructor or learner capability while the human retains the task and its authorship. *Excludes supervisory-only human roles.*	None proprietary	Not recoverable (0.11); manual only
C3	Personalisation	Pathway, sequence, difficulty or material is adjusted to an individual learner model. *Excludes cohort-level tailoring without an individual model.*	Adaptive / personalised	Recoverable (0.81)
C4	Generation	The primary output is a newly composed instructional or assessment artefact. *Excludes evaluative output such as a score or flag.*	Generative AI / LLM	Recoverable with caution (0.43)
C5	Analytics	Behavioural or performance data are analysed to diagnose, predict or warn; output is an inference delivered to a human decision-maker.	Learning analytics; educational data mining; prediction / early warning	Recoverable (0.82)
C6	Human–AI teaming	Joint problem-solving in which task allocation is negotiated within the episode and neither party is subordinate. *Excludes*	None	Not recoverable (0.01); manual only

Note: κ values are code-level agreement between the automated measure and the consensus of two independent coders on a stratified sample of 200 records, reported under Code-level prevalence, overlap and manual validation below. The codes draw on established distinctions in the AI-in-education literature — analytics and prediction (Romero and Ventura 2020), adaptive instruction (VanLehn 2011), literacy and governance (Long and Magerko 2020) — and on the affordance–risk framing of Lim et al. (2023). The typology is a measurement scaffold, not a claim about exhaustiveness.

Code	Form	Record-level inclusion rule (principal exclusion in italics)	Term indicators	Measurement status (κ vs. coders)
		question-answering and single-turn assistance.		
C7	Governance	Policy, ethics, integrity regimes and AI literacy provision at institutional or curricular level.	AI literacy; ethics / policy / governance; academic integrity	Recoverable (0.59)
—	*Technology families reported outside the code system*	Named families that cut across codes and cannot measure any one of them.	Intelligent tutoring; conversational / pedagogical agent	Descriptive series only

Note: κ values are code-level agreement between the automated measure and the consensus of two independent coders on a stratified sample of 200 records, reported under Code-level prevalence, overlap and manual validation below. The codes draw on established distinctions in the AI-in-education literature — analytics and prediction (Romero and Ventura 2020), adaptive instruction (VanLehn 2011), literacy and governance (Long and Magerko 2020) — and on the affordance–risk framing of Lim et al. (2023). The typology is a measurement scaffold, not a claim about exhaustiveness.

Methods

Corpus construction

Records were retrieved through Publish or Perish 8 (Harzing 2007) across three sources. Scopus served as the reference index for inclusion because its coverage is curated and its retrieval deterministic; Crossref extended coverage to indexed but unlisted venues and supplied the only substantial body of abstracts; Google Scholar was used solely to populate citation counts, and no record entered the corpus on a Google Scholar hit alone. That hierarchy confines the known non-determinism of Google Scholar retrieval to the impact statistics rather than allowing it to shape corpus composition.

Retrieval used paired term blocks combining AI terminology with education terminology, executed separately for the two periods. This was necessary: an initial retrieval using *artificial intelligence* as the primary AI term returned a pre-2023 subsample dominated by medical AI and containing almost no education-specific technology terms. Retrieval was re-run with period-appropriate terminology — *intelligent tutoring system*, *learning analytics*, *educational data mining*, *adaptive learning*, *automated essay scoring*, *pedagogical agent*, *recommender system* — which recovered the pre-generative literature including its principal venue, the *Journal of Learning Analytics*. The design invites an obvious objection, addressed below. This re-execution constitutes an amendment to the retrieval strategy and is reported as such, the superseded strategy and its yield being given in Supplementary Table S2.

The review was not registered and no protocol was prepared in advance. Eligibility criteria were nevertheless fixed before screening and were applied programmatically rather than by human raters, so

no screening reliability statistic is reported; the criteria followed PRISMA 2020 reporting conventions (Page et al. 2021) adapted for bibliometric rather than effect-size synthesis (Donthu et al. 2021). A record was retained where the title, abstract or keyword field contained at least one AI term and one education term; an explicit educational actor or setting was identifiable; the item was a journal article or peer-reviewed conference paper published between 1 January 2016 and 30 June 2026; and the record was in English. Records were excluded where the AI term occurred only within a compound belonging to another domain, where the education term denoted clinical patient education, or where the item was a book, chapter, preprint, editorial, erratum or citation-only record. Two screening rules warrant note: treating learning as an education signal admits large volumes of machine-learning research from unrelated domains, so compound forms were neutralised before testing, which removed the 37 records in which the AI term survived only as such a compound; and the phrase management education retrieves substantial diabetes self-management literature, of which 104 records were excluded as clinical or patient education. De-duplication operates on normalised title rather than DOI, because a number of Crossref items share a container DOI while being distinct works; adding DOI matching to the title rule removes only five further records, while a DOI-only rule would remove 2,425 — a figure that is an artefact of the 1,018 records carrying no DOI, which such a rule collapses into a single key. Figure 1 reports the resulting flow: 7,829 records identified, 1,764 duplicates removed, 1,830 excluded on document type or publication year and 3,017 on topical eligibility, leaving 1,218. No full-text eligibility stage is shown because none was conducted; the two corpora are nested rather than parallel, Corpus B (management education, n = 287) being a domain subset of Corpus A.

The corpus is split at the public release of ChatGPT (30 November 2022): pre-generative 2016–2022 (n = 333) and generative 2023–2026 (n = 885). Because the publication window closes on 30 June 2026, the generative period spans 3.5 years against the pre-generative period’s 7.0. Every count-based comparison below is annualised on those exposures, so that the truncated final year neither inflates nor deflates the generative rate. All retrieval runs were executed on 2 August 2026; 30 June 2026 is the closing date of the publication window, not the date of search.

Analytical procedures

Five analyses were conducted. First, descriptive bibliometrics by period. Second, term-frequency and bigram analysis on titles, with the six terms present in the retrieval strings flagged throughout so that query imprint is not reported as finding. Third, co-word network analysis (Callon et al. 1991) on the 40 most frequent title terms within each period, with co-occurrence matrices binarised at the 70th percentile, centrality computed following Freeman (1979), and CONCOR blockmodelling (Breiger et al. 1975) estimated separately for each period so that block structures are internally derived rather than imposed from a pooled solution.

Fourth, latent Dirichlet allocation (Blei et al. 2003) over the full corpus. The number of topics was not assumed: solutions for k = 4 to 12 were estimated on the 1,192 titles carrying three or more content

terms and compared on C_v coherence and held-out perplexity. Neither criterion identifies an optimum — coherence moves non-monotonically from 0.279 at k = 4 to 0.352 at k = 12, dipping to 0.287 at k = 8, while perplexity rises monotonically from 102.7 to 284.8 — which is characteristic of title-length documents; k = 8 (C_v = 0.287) was selected as the point at which topics remain distinguishable on inspection while the two lines of inquiry the study tracks are separately represented. Re-estimation across 20 seeds gives mean modal-assignment agreement of 37.8% against the reference seed, but the analytics topic and the adaptive-tutoring topic were recovered in 100% of seeds; inferences rest on topic mass rather than document assignment, and it is at that level that the solution is stable. Because a topic share is subject to the same compositional ambiguity as an indicator share, each period share is also reported as an implied document mass (share × N) and annualised.

Fifth, convergence-form prevalence. Each record's title and abstract were matched against eleven term-level indicators, and prevalence computed as the proportion of records in each period returning a match. Four quantities follow. Prevalence differences are reported with Newcombe hybrid-score intervals (Newcombe 1998) and Fisher exact tests. Annualised output is the implied record count divided by the period exposure. The rate ratio RR = r_2 / r_1 carries an exact conditional (binomial) interval, the standard estimator for a ratio of Poisson rates over fixed exposures (Rothman et al. 2008). And a growth-adjusted index G = RR ⁄ F expresses each form's growth as a fraction of the field's own growth ratio, F = 252.9 ⁄ 47.6 = 5.32 records per year.

The index has an algebraic property worth stating, because it determines which share statistic is safe to use. Writing p for prevalence, N for period size and y for exposure, $G = [(p_2 N_2 / y_2) / (p_1 N_1 / y_1)] \,/\, [(N_2 / y_2) / (N_1 / y_1)] = p_2 / p_1$. The growth-adjusted index *is* the prevalence ratio. The share ratio is therefore already growth-adjusted and requires no correction; it is the share *difference*, the statistic reviews conventionally report, that confounds composition with activity. Intervals for G are accordingly computed directly on the proportions by the Katz log method rather than by scaling the rate-ratio bounds, which avoids the rounding induced by reconstructing counts.

Two aggregation rules govern how indicator results may be combined. Indicator-level percentage-point changes are never summed to a code-level figure, since records can satisfy several indicators of the same code and a sum double-counts by an unknown amount; code-level prevalence is the union of a code's indicators, and where the union has not been recomputed from record-level data it is reported as a bounded interval whose lower limit is the largest single indicator and whose upper limit is the sum capped at 100%, with inferences drawn only where they hold across the interval. And where a record satisfies more than one code, a primary-form rule assigns it to the code describing the system's principal output. The automated measure is non-exclusive by construction and its figures are read as indicator prevalence, not as a partition.

Pre-specified criteria for contraction, stasis and survival

Claims that a research form has declined or survived require decision rules fixed before the estimates are inspected. Three principles govern those used here. A change in share is not evidence about activity, because the denominator changes. A rate ratio indistinguishable from unity is evidence of stasis only if the interval also excludes the field's growth ratio; otherwise the estimate is uninformative. And a form whose interval straddles a threshold is reported as straddling it rather than assigned to the nearer side (Gelman and Stern 2006). Each indicator is assigned to exactly one of five classes (Table 2).

Table 2
Pre-specified classification of period-to-period change

Class	Decision rule	Substantive reading
Absolute contraction	Upper bound of the 95% interval for RR < 1	Fewer records per year than before; the activity is being wound down
Absolute stasis	RR interval contains 1 and excludes F	Output per year is maintained but does not grow
Sub-proportional growth	Lower bound of RR > 1 and upper bound of G < 1	Output grows but more slowly than the field; a falling share records dilution, not decline
Proportional growth	G interval contains 1	The form grows at the field's rate and retains its position
Super-proportional growth	Lower bound of G > 1	The form grows faster than the field and gains position

A form is described as having *survived the transition* only where it falls in the proportional or super-proportional class. A form whose share is stable but whose G interval lies below unity has not survived in the relevant sense: it has been outgrown while retaining a constant slice of a much larger field. Conversely, a form whose share falls sharply but whose rate ratio exceeds unity has not declined; it has been diluted. No statement of stability or decline below rests on a share difference alone.

Sensitivity to retrieval design

If the pre-generative period is retrieved with analytics vocabulary and the generative period is not, a fall in analytics prevalence is guaranteed by construction rather than observed. Because the growth comparison is the central claim, three checks were run. The first required no new retrieval: two of the exports — the pre-generative and generative block runs logged as R20 and R8 in Supplementary Table S2 — were each executed across the whole 2016–2026 window rather than by period, so pooling them exposes both periods to both vocabularies. The pooled set yields 516 screened records (96 pre-generative, 420 generative), an annual output rising from 13.7 to 120.0 and a field growth ratio of F = 8.75. The second is a reciprocal search of each period's corpus for the other period's term set. The third deletes the retrieval strings from the text and re-estimates on what remains.

The union retrieval changes one conclusion and leaves the other intact. Under that design learning analytics moves from 5.6 to 11.1 records per year, RR = 2.00 (1.25–3.20); intelligent tutoring from 4.3 to

10.9, RR = 2.53 (1.53–4.23); educational data mining from 4.1 to 3.4, RR = 0.83 (0.38–1.67). The absolute-contraction finding for learning analytics is therefore design-dependent. What does not change is the growth-adjusted result: against the union set's own F = 8.75, learning analytics returns G = 0.23 (0.14–0.37), intelligent tutoring 0.29 (0.17–0.48) and educational data mining 0.09 (0.04–0.19) — in every case an order of magnitude below the field and with an interval excluding unity. The reciprocal search identifies 105 post-2023 records carrying pre-generative vocabulary (11.9% of the generative corpus) against a single pre-generative record carrying generative vocabulary, so the earlier vocabulary is in the minority rather than absent because unsearched. The imprint-free subset preserves a redistribution of 20.9 percentage points in share. Magnitudes soften under every check; the growth-adjusted ordering does not, and the interpretation below is stated in those terms. Full panels are given in Supplementary S8.

Results

Output, venues and knowledge structure

Table 3 describes Corpus A, its two periods and the management-education subset Corpus B.

[Table 3 near here]

Table 3
Corpus description by period

Corpus	N	Citations	Mean	Median	h	Uncited
Corpus A, full	1,218	76,252	62.6	3	137	32.7%
Corpus A, 2016–2022	333	16,830	50.5	8	60	14.7%
Corpus A, 2023–2026	885	59,422	67.1	2	116	39.4%
Corpus B (management education)	287	4,613	16.1	1	28	45.3%

The pre-generative period sustained a mean of 47.6 records per year. Output then rose to 135 in 2023 and 279 in 2025, giving a generative-period mean of 252.9 once the half-year 2026 window is taken into account (Fig. 2). The ratio of the two, F = 5.32, is the benchmark for every form-level comparison below; it is a property of the corpus rather than a finding, but it is the quantity whose omission makes share statistics uninterpretable. Declining mean citations after 2023 and the rise in uncited proportion from 14.7% to 39.4% reflect accrual time.

The venue structure changes as completely as the vocabulary. The *Journal of Learning Analytics* is the dominant pre-generative outlet with 53 records and does not appear in the generative-period top ten; *The International Journal of Management Education* takes its place as the largest generative-period outlet with 23. Two readings should be kept apart. As a statement about where attention now sits, this is a finding. As a statement about the analytics programme it is weaker than it looks, because a venue can fall out of

a top-ten list while publishing at an unchanged rate in a corpus that has tripled; the venue data are therefore used for composition only.

Network density rises from 0.244 to 0.288, but the bridging structure changes character. In the pre-generative network betweenness is distributed across several terms — system (0.165), data (0.125), intelligent (0.118), analytics (0.111), tutoring (0.084) — with no single term dominating. In the generative network generative alone carries 0.249, two and a half times the next term, with degree centrality 0.872. A field with several bridging terms has multiple semi-autonomous conversations connecting at more than one point; a field with one dominant bridge has a single conversation. Unlike a share statistic this comparison is not confounded by corpus growth, since both networks are estimated on the forty most frequent terms within their own period. The CONCOR solutions are not comparable in kind either: all four pre-generative blocks partition the field by system type, whereas the generative solution partitions by institutional problem, gathering twenty-two of forty terms into a single block. One generative block is worth noting — educational, automated, feedback, tutoring, adaptive, analytics, system, enhancing, systems — because a vocabulary that had been abandoned would dissolve into surrounding terms rather than reassemble into a coherent structural position. Centrality scores, block memberships and the network diagrams are given in Supplementary S8 and Figs. S1 and S2.

Topic dynamics

Table 4
Eight LDA topics with period shares and implied annual document mass (k = 8)

Topic	Leading terms	2016–22 share	2023–26 share	Mass/yr pre	Mass/yr post
T1 Pedagogical agents	agent, pedagogical, design, conversational, business	15.8%	6.2%	7.5	15.7
T2 LLMs in teaching	language, large, models, generative, teaching, accounting	4.7%	12.3%	2.2	31.1
T3 Academic integrity	academic, integrity, generative, chatgpt, higher, impact	3.7%	24.1%	1.8	61.0
T4 GenAI in business	generative, higher, challenges, business, management	5.5%	19.9%	2.6	50.3
T5 Curriculum and feedback	generative, curriculum, design, feedback, assessment	7.6%	16.8%	3.6	42.5
T6 Automated scoring	automated, scoring, assessment, essay, feedback, framework	12.6%	7.7%	6.0	19.5
T7 Tutoring and adaptation	intelligent, tutoring, system, personalized, adaptive	22.7%	7.5%	10.8	19.0
T8 Analytics and EDM	data, analytics, educational, mining, student, big	27.5%	5.5%	13.1	13.9

Implied annual mass is the period share multiplied by the number of records in the period and divided by the period exposure (7.0 and 3.5 years). Because document–topic proportions sum to one within each period, shares are compositional and their sum across topics is fixed; mass is not. Year-wise shares are given in Supplementary Table S8.

Two redistributions are visible, one nested inside the other (Fig. 3). The primary redistribution occurs at the period boundary: the analytics topic (T8) falls from 35.5% of topic mass in 2016 to 3.6% in 2026, and the tutoring topic (T7) from 22.7% to 9.1%; together the two topics that constituted the field in 2016 fall from 58.2% to 12.7% of its topic mass. The final two columns of Table 4 show why these figures cannot be read as descriptions of activity. Expressed as implied mass per year, T8 moves from 13.1 to 13.9 and T7 from 10.8 to 19.0. The analytics topic holds its absolute mass almost exactly constant while losing four-fifths of its share; every topic in the table gains mass, because the corpus grew faster than any single topic lost share. The topic model therefore corroborates the redistribution of attention and does not corroborate the abandonment of the earlier programme. The secondary redistribution occurs within the generative period: academic integrity peaks at 35.2% in 2023 and falls to 21.1% by 2026 while

curriculum design and feedback rises from 13.8% to 21.9%. The field’s initial response was defensive; within three years it had become constructive.

Growth of convergence indicators

Table 5 reports both panels together: the share of each period’s records matching each indicator, and the annualised output that the share implies. Read as composition the shares are informative — a reader entering the field in 2016 met analytics vocabulary in more than a third of records, one entering in 2024 meets it in about one in twenty. Read as evidence about activity they are uninterpretable, because the denominator grew 5.32-fold.

Table 5
Convergence-indicator prevalence, annualised output, rate ratios and growth-adjusted index, with pre-specified classification

Indicator	2016–22	2023–26	Pre /yr	Post /yr	RR [95% CI]	G [95% CI]	Class
Educational data mining	23.1%	1.4%	11.0	3.4	0.31 [0.15, 0.58]	0.06 [0.03, 0.11]	Absolute contraction
Learning analytics	37.8%	4.7%	18.0	12.0	0.67 [0.46, 0.95]	0.13 [0.09, 0.18]	Absolute contraction
Intelligent tutoring	24.9%	4.5%	11.9	11.4	0.96 [0.64, 1.42]	0.18 [0.12, 0.27]	Absolute stasis
Prediction / early warning	9.6%	3.7%	4.6	9.4	2.06 [1.23, 3.46]	0.39 [0.23, 0.65]	Sub-proportional
Conversational / pedagogical agent	9.9%	4.5%	4.7	11.4	2.42 [1.49, 3.97]	0.46 [0.28, 0.75]	Sub-proportional
Automated feedback / scoring	8.4%	4.9%	4.0	12.3	3.07 [1.86, 5.13]	0.58 [0.35, 0.97]	Sub-proportional
Adaptive / personalised	16.5%	15.6%	7.9	39.4	5.02 [3.65, 6.99]	0.94 [0.69, 1.32]	Proportional
AI literacy	0.9%	3.3%	0.4	8.3	19.3 [5.99, 99.2]	3.64 [1.13, 18.7]	Super-proportional
Ethics / policy / governance	4.2%	20.9%	2.0	52.9	26.4 [15.4, 49.3]	4.97 [2.89, 9.27]	Super-proportional
Academic integrity	0.3%	20.2%	0.1	51.1	358 [63.4, 14218]	67.4 [11.9, 2675]	Super-proportional
Generative AI / LLM	0.3%	54.0%	0.1	136.6	956 [171, 37838]	180 [32.1, 7119]	Super-proportional

n = 333 (2016–2022) and n = 885 (2023–June 2026); exposures 7.0 and 3.5 years. Share differences carry Newcombe hybrid-score 95% intervals and are tested with two-sided Fisher exact tests at α = 0.05; the exact p value for every indicator is given in Supplementary Table S9. All eleven differences return $p < 0.05$ except adaptive / personalised (− 0.9 pp, $p = 0.73$). RR is the ratio of annualised rates with an exact conditional interval; $G = p_2/p_1$ with a Katz log interval; $F = 5.32$. Classes follow Table 2. Intervals for the

two rarest pre-generative indicators are very wide because their pre-generative counts are one record; only the sign of those ratios should be interpreted.

The pattern is differential growth, not uniform displacement (Fig. 4). Of eleven indicators, two contract absolutely, one is statistically flat, three grow more slowly than the field, one keeps pace and four grow faster. The forms that contract are the two that depend most heavily on instrumented longitudinal data: educational data mining falls from 11.0 to 3.4 records per year, and learning analytics from 18.0 to 12.0 with an interval whose upper bound sits just below unity, a modest contraction rather than a collapse.

Intelligent tutoring shows most sharply why the two panels must be read together. Its share falls by 20.4 percentage points, significant at any conventional level, while its annual output moves from 11.9 to 11.4 records, a rate ratio of 0.96 whose interval is entirely consistent with no change. The 20.4-point fall is a fact about the composition of the literature, not about how much intelligent tutoring research is being done. The same logic revises the status of the forms a share-based reading would call survivors: conversational agents fall 5.4 points in share while rising from 4.7 to 11.4 records per year, and automated feedback falls 3.5 points while rising from 4.0 to 12.3. Neither has declined, but both return G well below unity, so both have been diluted — they grew at roughly half the field's rate. Describing them as essentially unchanged would convert a substantial proportional decline into apparent stability; describing them as declining would convert a threefold increase in output into a loss. Neither description is available, which is what the classification of Table 2 exists to prevent. Exactly one pre-generative indicator survives on the stated criterion: adaptive and personalised work rises from 7.9 to 39.4 records per year with G = 0.94 (0.69–1.32), the only indicator for which stability of share and stability of position coincide.

Code-level prevalence, overlap and manual validation

Indicators are non-exclusive, so a code with more than one indicator has a bounded rather than a fixed prevalence: at least the largest single indicator, at most their sum. Those bounds are given in Supplementary Table S13. Two conclusions follow and only two. The direction of change is established for all four measurable codes, because the sign of Δ is invariant across the admissible range of overlap; the magnitude is not established for C5 and C7, and no point estimate of it is reported. In particular, an additive figure of the kind obtained by summing the three C5 indicators' percentage-point changes is not admissible.

Whether the prevalence figures record substance as well as vocabulary was tested rather than assumed. A sample of 200 records was drawn with equal allocation by period — 100 pre-generative and 100 generative, seed 20260630 — and coded independently by two researchers against the definitions in Table 1, working from titles and abstracts only because that is the field on which the automated measure operates, and blind to the automated classification. Reliability was computed before any reconciliation. The equal allocation maximises precision within each period, so within-period proportions are unbiased estimates of period prevalence, at the cost of making the pooled proportion across all 200 records a quantity that is not used as an estimate of corpus-wide prevalence anywhere below.

Table 6
Validation of the convergence-form measure (n = 200 records)

Code	Inter-coder κ	Inter-coder agreement	κ vs. automated	Agreement vs. automated
C1 Substitution	0.81	97.5%	−0.13	74.5%
C2 Augmentation	0.70	86.0%	0.11	59.0%
C3 Personalisation	0.95	98.5%	0.81	94.5%
C4 Generation	0.75	94.0%	0.43	77.5%
C5 Analytics	0.85	94.5%	0.82	93.0%
C6 Human–AI teaming	0.61	97.0%	0.01	87.5%
C7 Governance	0.72	88.5%	0.59	83.5%
Pooled	0.79	93.7%	0.38	81.4%

Note: Inter-coder reliability is computed on the seven binary code decisions before reconciliation. Consensus coding assigns a code where either coder assigned it. Agreement on the single primary-form judgement, a seven-way rather than binary decision, is 78.0% (κ = 0.73). The automated series compared against each code are: intelligent tutoring for C1 and C2, adaptive / personalised for C3, the union of automated feedback / scoring and generative AI / LLM for C4, the union of learning analytics, educational data mining and prediction / early warning for C5, conversational / pedagogical agent for C6, and the union of AI literacy, ethics / policy / governance and academic integrity for C7. Where Table 1 records no proprietary indicator, the nearest available series is used and the resulting κ is reported as evidence that the code is not recoverable by vocabulary.

Inter-coder reliability is substantial and even: pooled κ = 0.79 computed on the stacked binary decisions, with 93.7% agreement, and every code clears 0.60. The typology can therefore be applied consistently by independent coders working from titles and abstracts. How far the automated measure recovers those judgements determines which codes may enter a term-based growth analysis at all. It performs well where the indicator phrase names both a technique and the function it serves — C5 at κ = 0.82, C3 at 0.81 — and fails for C1, C2 and C6, where κ is at or below zero. The reason is structural: substitution, augmentation and teaming are functional categories with no proprietary vocabulary. Coders recognised substitution or augmentation in 97 records while the nearest term-based indicator, intelligent tutoring, fired on 34. A paper in which a system marks student work is coded C1 by a reader and missed by any expression that waits for a phrase. That failure does not leave the three codes unmeasured, because the manual coding is itself a prevalence measure and the stratified design makes it usable as one. Table 7 reports the resulting code-level series.

Table 7
Manually coded code-level prevalence by period (n = 200)

Code	Form	2016–22 (n = 100)	2023–26 (n = 100)	RR	G [95% CI]	Class
C1	Substitution	8.0% [4.1, 15.0]	9.0% [4.8, 16.2]	5.98	1.12 [0.45, 2.80]	Proportional
C2	Augmentation	41.0% [31.9, 50.8]	47.0% [37.5, 56.7]	6.09	1.15 [0.84, 1.57]	Proportional
C3	Personalisation	25.0% [17.5, 34.3]	13.0% [7.8, 21.0]	2.76	0.52 [0.28, 0.96]	Sub-proportional
C4	Generation	6.0% [2.8, 12.5]	28.0% [20.1, 37.5]	24.8	4.67 [2.02, 10.78]	Super-proportional
C5	Analytics	44.0% [34.7, 53.8]	8.0% [4.1, 15.0]	0.97	0.18 [0.09, 0.37]	Absolute stasis
C6	Human–AI teaming	3.0% [1.0, 8.5]	8.0% [4.1, 15.0]	14.2	2.67 [0.73, 9.76]	Proportional
C7	Governance	14.0% [8.5, 22.1]	54.0% [44.3, 63.4]	20.5	3.86 [2.30, 6.48]	Super-proportional

Note: Percentages are within-period proportions with Wilson 95% intervals. G is the prevalence ratio with a Katz log interval; RR = G × F, and the interval for RR is the interval for G scaled by F, so the classification in the final column can be read directly from the G interval. Intervals are wide because estimates rest on 100 records per period; point estimates should not be cited alone.

All three constructs the automated measure failed to recover keep pace with the field: substitution G = 1.12, augmentation 1.15, human–AI teaming 2.67, none of the intervals excluding or falling below unity. Analytics returns G = 0.18 with RR = 0.97 — the same absolute stasis found at indicator level for intelligent tutoring, now observed for the analytics construct as a whole and by human readers rather than a regular expression. Personalisation is the case where the two instruments disagree: as a term-based indicator it returns G = 0.94 and keeps pace, as a construct requiring an individual learner model it returns 0.52 and does not. The intervals overlap between 0.69 and 0.96, so the estimates are not statistically incompatible, and the difference has a substantive reading: generative-period papers use the vocabulary of personalisation far more often than they build the learner model that C3 requires.

One reliability result belongs alongside these figures. The coders diverged on whether a record was codable at all — 17 flags against 54, with 14 in common and 78.5% agreement — a difference in coder severity rather than an artefact of the missing abstracts, which are absent for 42.5% of the sample and

distributed similarly across both coders' flags. Both coders returned assignments for all 200 records, so Table 7 is complete; two sensitivity analyses bound the risk. Excluding the 14 records both flagged moves the estimates to C1 1.23, C2 1.25, C3 0.57, C5 0.20, C6 2.54; excluding all 57 either flagged gives 1.39, 1.15, 0.64, 0.23 and 1.23. The ordering is unchanged under both restrictions and C5 remains an order of magnitude below the field, though C3's interval crosses unity in the restricted sets, which is why it is treated below as a boundary case.

The management education subcorpus

Corpus B (n = 287) contains 7 pre-generative records, so the period comparison conducted for Corpus A cannot be conducted for the subfield. It is equally important that those 7 records are not read in the opposite direction, as evidence that pre-2023 management education work is scarce. Corpus B is nested within Corpus A, whose pre-2023 block was retrieved with general-education technology vocabulary; a pre-2023 management education literature organised around business simulation, decision support systems or assurance of learning would not have been retrieved whether or not it exists. The number is a property of the retrieval design, not a measurement of the subfield.

What the corpus does support is a cross-sectional comparison. Generative-period vocabulary in Corpus B is dominated by *business* (113), *management* (74), *accounting* (60), *generative* (55) and *adoption* (22) — the disciplinary-integration cluster rather than the integrity terms that dominate the general corpus, and containing essentially none of the analytics vocabulary that dominates Corpus A's pre-generative period. The inheritance argument therefore takes one form and no stronger: management education began publishing at a moment when the general literature it drew on was composed as the right-hand column of Table 5 describes, and what it imported was that composition. Whether the subfield also had an earlier tradition of its own is a separate question this corpus cannot answer.

Discussion

Differential growth

The generative turn redistributed the field's attention comprehensively and reduced its absolute research activity hardly at all. On the composition side the change is as complete as any bibliometric record is likely to show: the leading venue changed, thirteen of the twenty-two most frequent title terms were replaced, the block structure moved from a partition by system type to a partition by institutional problem, and a single term acquired a bridging centrality no term held before. On the activity side, eight of eleven indicators publish more records per year than in the earlier period; of the three that do not, the two most dependent on instrumented longitudinal data fall significantly, while intelligent tutoring holds a rate statistically indistinguishable from its earlier one.

The correct description is differential growth. The field grew 5.32-fold and its earlier programme did not grow with it. Framing this as displacement, in the sense of one programme evicting another, overstates it: under a union retrieval that exposes both periods to both vocabularies even the contraction of learning

analytics becomes twofold growth. What survives every retrieval design and both measurement instruments is the growth-adjusted comparison, and it is a statement about relative rather than absolute abandonment. A prosaic institutional reading fits better than a Kuhnian one (Kuhn 1962). Research capacity is finite in the short run but expands in response to opportunity, and a technology that becomes available to every instructor and student at once generates a large stock of immediately publishable questions that draws in new capacity rather than only reallocating existing capacity. Almost nothing was cannibalised outright; almost nothing kept up either, because the new questions could be answered without the instrumented platforms and longitudinal datasets the old ones required.

Substrate portability

The ordering in Tables 5 and 7 is not random with respect to how the forms are defined. Table 8 classifies them on two attributes assigned from their definitions in Table 1: whether the form is defined by an educational function or by a technique, and whether instantiating it requires instrumented longitudinal behavioural data. Panel A covers the pre-generative indicators of the term-based measure; Panel B the pre-generative constructs of the manual coding. Generative and governance forms are excluded from both, being the arrivals against which the others are compared rather than candidates for survival.

Table 8
Attributes of pre-generative convergence forms and growth outcome

Form	Definition	Requires instrumented longitudinal data	G [95% CI]	Class
Panel A. Indicator level — term-based measure				
Educational data mining	Technique	Yes	0.06 [0.03, 0.11]	Contraction
Learning analytics	Technique	Yes	0.13 [0.09, 0.18]	Contraction
Intelligent tutoring	Technique	Partly	0.18 [0.12, 0.27]	Stasis
Prediction / early warning	Function	Yes	0.39 [0.23, 0.65]	Sub-proportional
Conversational / pedagogical agent	Function	No	0.46 [0.28, 0.75]	Sub-proportional
Automated feedback / scoring	Function	No	0.58 [0.35, 0.97]	Sub-proportional
Adaptive / personalised	Function	Not as detected by vocabulary	0.94 [0.69, 1.32]	Proportional
Panel B. Code level — manual coding				
C5 Analytics	Function	Yes	0.18 [0.09, 0.37]	Stasis
C3 Personalisation	Function	Yes — individual learner model	0.52 [0.28, 0.96]	Sub-proportional
C1 Substitution	Function	No	1.12 [0.45, 2.80]	Proportional
C2 Augmentation	Function	No	1.15 [0.84, 1.57]	Proportional
C6 Human–AI teaming	Function	No	2.67 [0.73, 9.76]	Proportional

The two panels separate what the indicator-level analysis alone could not. In Panel A both attributes vary and the outcomes order monotonically: technique-defined and data-dependent forms at the bottom, function-defined and data-independent forms at the top, and the single case that is function-defined but data-dependent — prediction and early warning — between them, closer to the bottom than its functional definition alone would predict. In Panel B the first attribute does not vary at all, because every code in the typology is defined by an educational function; yet the outcomes still order, and they order on the second attribute alone. The two codes that presuppose instrumented longitudinal data sit at 0.18 and 0.52, the three that do not at 1.12, 1.15 and 2.67. Functional definition is a property of how a form is named and does useful work at the level of vocabulary; infrastructural dependence is a property of what a form requires in order to exist, and it is what determines whether the form survives a change of substrate.

Personalisation is where the panels part company, and the divergence is the account's strongest evidence rather than an inconsistency in it. Measured by vocabulary, adaptive personalisation looks infrastructure-free and keeps pace at G = 0.94; measured as the construct C3, which requires an individual learner model and therefore an instrumented platform, the same form falls to 0.52 and joins the data-dependent group. This is one research form, measured twice, crossing the predicted boundary exactly when the definition carrying the infrastructural requirement is applied. The account generates a testable prediction: where a new substrate supplies the missing infrastructure — for instance where generative systems are coupled to institutional data platforms — the data-dependent forms should recover, and C3 should recover before C5. Because the classification was assigned from the operational rules before the code-level tabulation but was not pre-registered, the result of Table 7 is corroboration rather than confirmation.

Two further implications follow. The outgrown work was not completed: prediction of student difficulty, early warning and data-driven diagnosis were open problems in 2022 and remain open, the difference being that they are now pursued at a fraction of the field's rate rather than at its centre. The distinction between *unfinished and outgrown* and *finished* has not been drawn in the review literature, which treats the generative period as an undifferentiated expansion. And the generative-period literature is structurally less differentiated than its predecessor despite being three times larger, which is itself a scientometric finding about what rapid growth around a single technology does to a field's internal organisation.

Implications for a late-entering subfield

The practical question for a business school is not whether analytics knowledge exists but what arrives when it imports the current literature, and that is answered by composition: analytics vocabulary appears in about 5% of the records a 2024 entrant encounters against 38% for a 2016 entrant. A school adopting the contemporary literature as its evidence base receives a repertoire in which diagnostic capability is a residual category, even though the underlying research continues at close to its earlier volume — roughly 12 records a year inside a field publishing 253. The material exists and is being added to, but it is not indexed by the venues, terms or reviews a newcomer would consult first. The recommendation is a search strategy rather than a research programme, and a school that assumes otherwise will rebuild

what it could have adopted. More generally, the case illustrates a hazard that applies wherever a subfield enters an established literature late: what it inherits is the composition of that literature at the moment of entry, and composition and activity diverge sharply in periods of rapid growth.

Conclusion, limitations and future work

Across a decade, AI–education research grew 5.32-fold and its founding programme did not grow with it. Of seven pre-generative research forms, two contracted in absolute annual output, one held level, three grew at between a fifth and three-fifths of the field's rate, and one kept pace; at code level the analytics construct held its output constant at 18% of the field's growth while substitution, augmentation and human–AI teaming kept pace. The forms that travelled best did not depend on instrumented longitudinal data. Methodologically, the study shows that share differences cannot support claims about the trajectory of a research programme in a rapidly expanding literature, that the share ratio is the growth-adjusted quantity and is safe to use, and that estimating network and block structures separately by period rather than pooling is necessary to see the redistribution at all.

Several limitations bound the claims. The corpus analysed here is a reconstruction. The original retrieval yielded 1,268 records, 305 pre-generative and 963 generative; re-executing the screening rules on the raw exports, which is what makes the corpus auditable record by record, yields 1,218, 333 and 885. Every figure reported in this paper is computed on the reconstruction. The two sets differ by 3.9% in total but by more in composition, and the field growth ratio falls from 6.31 to 5.32 as a result; that direction is conservative for the central claim, because a smaller benchmark raises every growth-adjusted index rather than lowering it. The counts underlying Table 5 are implied by prevalence figures and period totals rather than tabulated independently, and should be verified against record-level data in any replication; the growth-adjusted index is unaffected, since it is computed on proportions. The classification of Table 8 was assigned from the operational rules rather than pre-registered, so the portability account is abductive. The code-level series rests on 100 records per period, so its intervals are wide, and the coders disagreed materially on codability, which the sensitivity analyses bound but do not eliminate. The coding instrument also exists in two versions whose treatment of C6 differs, one admitting conversational and pedagogical agents to the code and one restricting it to symmetric collaboration as Table 1 requires; both are deposited, and the C6 estimate should be read as sensitive to that difference. The corpus draws on three sources rather than one curated index; Google Scholar returns only the 1,000 most-cited items per query, Crossref includes non-refereed material, and Scopus retrieval was narrower. Abstracts were available for 746 of 1,218 records, so topic modelling was estimated over titles, which are semantically sparse, and the eight-topic solution is a coarse partition. Query imprint remains the most serious threat to the central comparison; the checks reported above preserve the growth-adjusted ordering but not the absolute-contraction finding for learning analytics, which is design-dependent and reported as such. Author, institution and country analysis was not possible because the retrieval tool does not supply affiliation fields.

Four extensions follow. Record-level recomputation of code-level unions would replace the overlap bounds reported above with point estimates. Extending the manual coding beyond 200 records would

narrow the intervals of Table 7 and settle whether C3 is a boundary case or a genuine member of the data-dependent group. A dedicated pre-2023 retrieval for the management education subfield — combining its domain vocabulary with both the pre-generative technology terms and its own analytic tradition, or sweeping its core journals exhaustively by ISSN where index access is constrained — would convert the subfield argument from supply-side inference into measurement. And the mechanism proposed here invites an author-level test: whether researchers active in learning analytics before 2023 migrated to generative topics, continued in place, or exited, and whether those who continued did so where institutional data infrastructure was available to them. More broadly, the growth-adjusted index and the pre-specified classification are not specific to this literature. Any field undergoing rapid expansion around a new technology presents the same inferential hazard, and the same two-panel reading — share for composition, annualised count for activity — is available wherever period boundaries can be defined.

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

Ethical approval was not required as the study analyses published bibliographic metadata only and involved no human participants or their data.

Informed consent

Informed consent was not required as the study involved no human participants.

Author Contribution

Conceptualization: Yu Miao, Chang-Hyun Jin; Methodology: Yu Miao; Software: Jianwei Yu; Formal analysis: Jianwei Yu; Investigation: Yu Miao, Mengting Zhao; Data curation: Mengting Zhao; Validation: Jianwei Yu; Visualization: Mengting Zhao; Writing – original draft preparation: Yu Miao, Mengting Zhao; Writing – review and editing: Jianwei Yu, Chang-Hyun Jin; Supervision: Chang-Hyun Jin; Project administration: Chang-Hyun Jin. All authors read and approved the final manuscript.

Acknowledgements

This research received no specific grant from funding agencies in the public, commercial or not-for-profit sectors.

Data Availability

The manual coding data generated in this study are provided as Online Resource 2 and the analysis scripts as Online Resource 3; the retrieval and screening specification is given in Online Resource 1. The screened corpus metadata are derived from Scopus, Crossref and Google Scholar and cannot be redistributed in full under those providers' terms; they are available from the corresponding author on reasonable request.

Code availability

The scripts used for retrieval processing, de-duplication, screening, indicator matching, latent Dirichlet allocation, co-word and CONCOR estimation, and interval computation are provided as Supplementary S9 and archived with the data deposit at https://doi.org/10.7910/DVN/FULQPK. Analyses were carried out in Python 3 using pandas, numpy, scipy, scikit-learn, gensim, networkx and matplotlib. The exact versions of the environment in which the reported results were produced are not recoverable and are therefore not stated. Supplementary S9 also records the run order, which script produced which deposited file, and three known defects in the code.

References

1. Aldowah H, Al-Samarraie H, Fauzy WM (2019) Educational data mining and learning analytics for 21st century higher education: a review and synthesis. Telematics Inform 37:13–49. https://doi.org/10.1016/j.tele.2019.01.007
2. Bahroun Z, Anane C, Ahmed V, Zacca A (2023) Transforming education: a comprehensive review of generative artificial intelligence in educational settings through bibliometric and content analysis. Sustainability 15(17):12983. https://doi.org/10.3390/su151712983
3. Baidoo-Anu D, Owusu Ansah L (2023) Education in the era of generative artificial intelligence (AI): understanding the potential benefits of ChatGPT in promoting teaching and learning. J AI 7(1):52–62. https://doi.org/10.61969/jai.1337500
4. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022
5. Blikstein P, Worsley M (2016) Multimodal learning analytics and education data mining: using computational technologies to measure complex learning tasks. J Learn Analytics 3(2):220–238. https://doi.org/10.18608/jla.2016.32.11
6. Bornmann L, Mutz R (2015) Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. J Association Inform Sci Technol 66(11):2215–2222. https://doi.org/10.1002/asi.23329

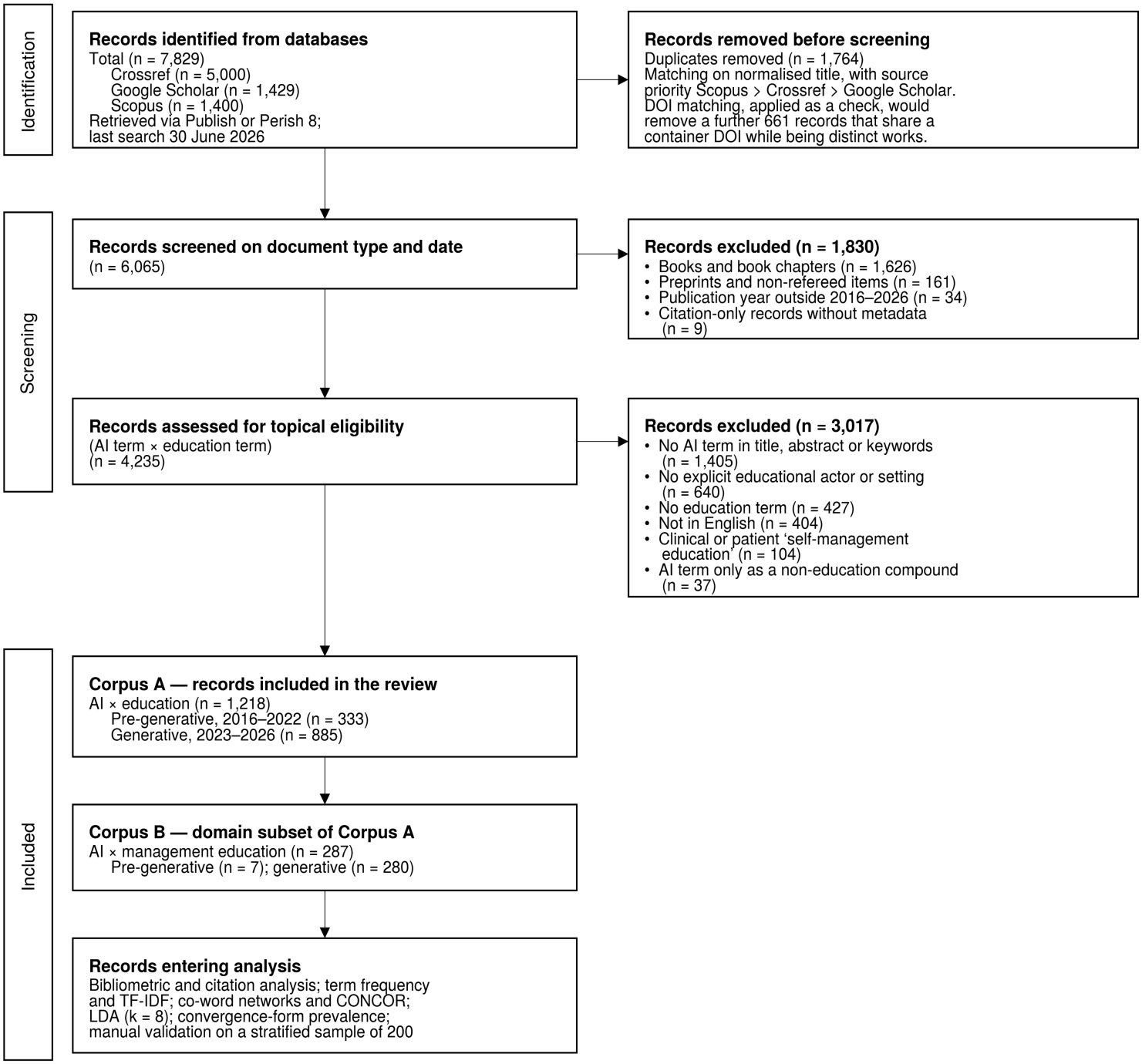
7. Breiger RL, Boorman SA, Arabie P (1975) An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. J Math Psychol 12(3):328–383. https://doi.org/10.1016/0022-2496(75)90028-0
8. Callon M, Courtial J-P, Laville F (1991) Co-word analysis as a tool for describing the network of interactions between basic and technological research. Scientometrics 22(1):155–205. https://doi.org/10.1007/BF02019280
9. Celik I, Dindar M, Muukkonen H, Järvelä S (2022) The promises and challenges of artificial intelligence for teachers: a systematic review of research. TechTrends 66:616–630. https://doi.org/10.1007/s11528-022-00715-y
10. Chassignol M, Khoroshavin A, Klimova A, Bilyatdinova A (2018) Artificial intelligence trends in education: a narrative overview. Procedia Comput Sci 136:16–24. https://doi.org/10.1016/j.procs.2018.08.233
11. Chiu TKF, Xia Q, Zhou X, Chai CS, Cheng M (2023) Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. Computers Education: Artif Intell 4:100118. https://doi.org/10.1016/j.caeai.2022.100118
12. Cotton DRE, Cotton PA, Shipway JR (2024) Chatting and cheating: ensuring academic integrity in the era of ChatGPT. Innovations Educ Teach Int 61(2):228–239. https://doi.org/10.1080/14703297.2023.2190148
13. Crompton H, Burke D (2023) Artificial intelligence in higher education: the state of the field. Int J Educational Technol High Educ 20:22. https://doi.org/10.1186/s41239-023-00392-8
14. Donthu N, Kumar S, Mukherjee D, Pandey N, Lim WM (2021) How to conduct a bibliometric analysis: an overview and guidelines. J Bus Res 133:285–296. https://doi.org/10.1016/j.jbusres.2021.04.070
15. Farrokhnia M, Banihashem SK, Noroozi O, Wals A (2024) A SWOT analysis of ChatGPT: implications for educational practice and research. Innovations Educ Teach Int 61(3):460–474. https://doi.org/10.1080/14703297.2023.2195846
16. Freeman LC (1979) Centrality in social networks: conceptual clarification. Social Networks 1(3):215–239. https://doi.org/10.1016/0378-8733(78)90021-7
17. Gelman A, Stern H (2006) The difference between significant and not significant is not itself statistically significant. Am Stat 60(4):328–331. https://doi.org/10.1198/000313006X152649
18. Harzing A-W (2007) Publish or Perish [Computer software]. https://harzing.com/resources/publish-or-perish
19. Kuhn TS (1962) The structure of scientific revolutions. University of Chicago Press, Chicago
20. Lim WM, Gunasekara A, Pallant JL, Pallant JI, Pechenkina E (2023) Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. Int J Manage Educ 21(2):100790. https://doi.org/10.1016/j.ijme.2023.100790
21. Lo CK (2023) What is the impact of ChatGPT on education? A rapid review of the literature. Educ Sci 13(4):410. https://doi.org/10.3390/educsci13040410

22. Long D, Magerko B (2020) What is AI literacy? Competencies and design considerations. In: Proceedings of the 2020 CHI conference on human factors in computing systems. Association for Computing Machinery, New York, p 1–16. https://doi.org/10.1145/3313831.3376727
23. Luan H, Geczy P, Lai H, Gobert J, Yang SJH, Ogata H, Baltes J, Guerra R, Li P, Tsai C-C (2020) Challenges and future directions of big data and artificial intelligence in education. Front Psychol 11:580820. https://doi.org/10.3389/fpsyg.2020.580820
24. Moorhouse BL, Yeo MA, Wan Y (2023) Generative AI tools and assessment: guidelines of the world's top-ranking universities. Computers Educ Open 5:100151. https://doi.org/10.1016/j.caeo.2023.100151
25. Newcombe RG (1998) Interval estimation for the difference between independent proportions: comparison of eleven methods. Stat Med 17(8):873–890. https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8%3C873::AID-SIM779%3E3.0.CO;2-I
26. Ouyang F, Zheng L, Jiao P (2022) Artificial intelligence in online higher education: a systematic review of empirical research from 2011 to 2020. Educ Inform Technol 27:7893–7925. https://doi.org/10.1007/s10639-022-10925-9
27. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, Moher D (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372:n71. https://doi.org/10.1136/bmj.n71
28. Perkins M (2023) Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. J Univ Teach Learn Pract 20(2):07. https://doi.org/10.53761/1.20.02.07
29. de Price DJ S (1963) Little science, big science. Columbia University, New York
30. Ratten V, Jones P (2023) Generative artificial intelligence (ChatGPT): implications for management educators. Int J Manage Educ 21(3):100857. https://doi.org/10.1016/j.ijme.2023.100857
31. Romero C, Ventura S (2020) Educational data mining and learning analytics: an updated survey. WIREs Data Min Knowl Discov 10(3):e1355. https://doi.org/10.1002/widm.1355
32. Rothman KJ, Greenland S, Lash TL (2008) Modern epidemiology, 3rd edn. Lippincott Williams & Wilkins, Philadelphia
33. Rudolph J, Tan S, Tan S (2023) ChatGPT: bullshit spewer or the end of traditional assessments in higher education? J Appl Learn Teach 6(1):342–363. https://doi.org/10.37074/jalt.2023.6.1.9
34. Seo K, Tang J, Roll I, Fels S, Yoon D (2021) The impact of artificial intelligence on learner-instructor interaction in online learning. Int J Educational Technol High Educ 18:54. https://doi.org/10.1186/s41239-021-00292-9
35. Siemens G, Baker RSJ (2012) Learning analytics and educational data mining: towards communication and collaboration. In: Proceedings of the 2nd international conference on learning

analytics and knowledge. Association for Computing Machinery, New York, p 252–254. https://doi.org/10.1145/2330601.2330661

36. Sollosy M, McInerney M (2022) Artificial intelligence and business education: what should be taught. Int J Manage Educ 20(3):100720. https://doi.org/10.1016/j.ijme.2022.100720
37. Sullivan M, Kelly A, McLaughlan P (2023) ChatGPT in higher education: considerations for academic integrity and student learning. J Appl Learn Teach 6(1):31–40. https://doi.org/10.37074/jalt.2023.6.1.17
38. VanLehn K (2011) The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychol 46(4):197–221. https://doi.org/10.1080/00461520.2011.611369
39. Wang S, Wang F, Zhu Z, Wang J, Tran T, Du Z (2024) Artificial intelligence in education: a systematic literature review. Expert Syst Appl 252:124167. https://doi.org/10.1016/j.eswa.2024.124167
40. Yan L, Sha L, Zhao L, Li Y, Martinez-Maldonado R, Chen G, Li X, Jin Y, Gašević D (2024) Practical and ethical challenges of large language models in education: a systematic scoping review. Br J Edu Technol 55(1):90–112. https://doi.org/10.1111/bjet.13370
41. Zawacki-Richter O, Marín VI, Bond M, Gouverneur F (2019) Systematic review of research on artificial intelligence applications in higher education — where are the educators? Int J Educational Technol High Educ 16:39. https://doi.org/10.1186/s41239-019-0171-0
42. Zhai X, Chu X, Chai CS, Jong MSY, Istenic A, Spector M, Liu J-B, Yuan J, Li Y (2021) A review of artificial intelligence (AI) in education from 2010 to 2020. Complexity 2021:8812542. https://doi.org/10.1155/2021/8812542

Figures



Note: adapted from the PRISMA 2020 statement (Page et al., 2021) for bibliometric synthesis. No full-text eligibility stage is shown because none was conducted: screening operated on title, abstract and keyword fields. Corpus B is nested within Corpus A rather than retrieved separately.

Figure 1

PRISMA 2020 flow diagram of corpus construction, adapted for bibliometric synthesis. Screening operated on title, abstract and keyword fields; retrieval exports, the screening script and the record-level log of excluded records with reasons are provided as Supplementary S3, S4 and S9.

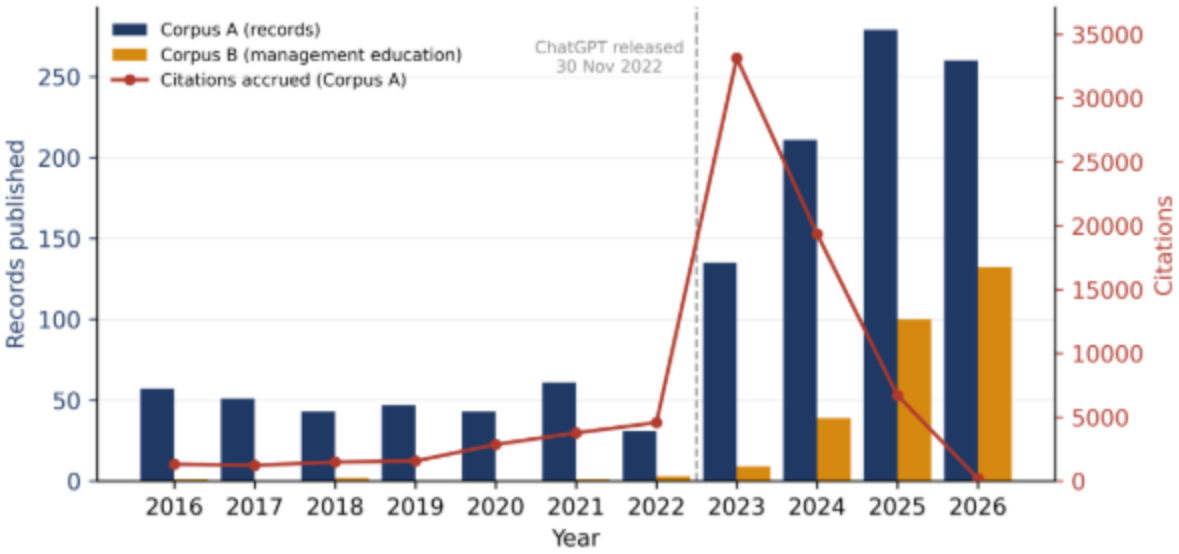


Figure 2

Annual output and citations, 2016–2026. The pre-generative period sustains 30–60 records per year; output reaches 279 in 2025.

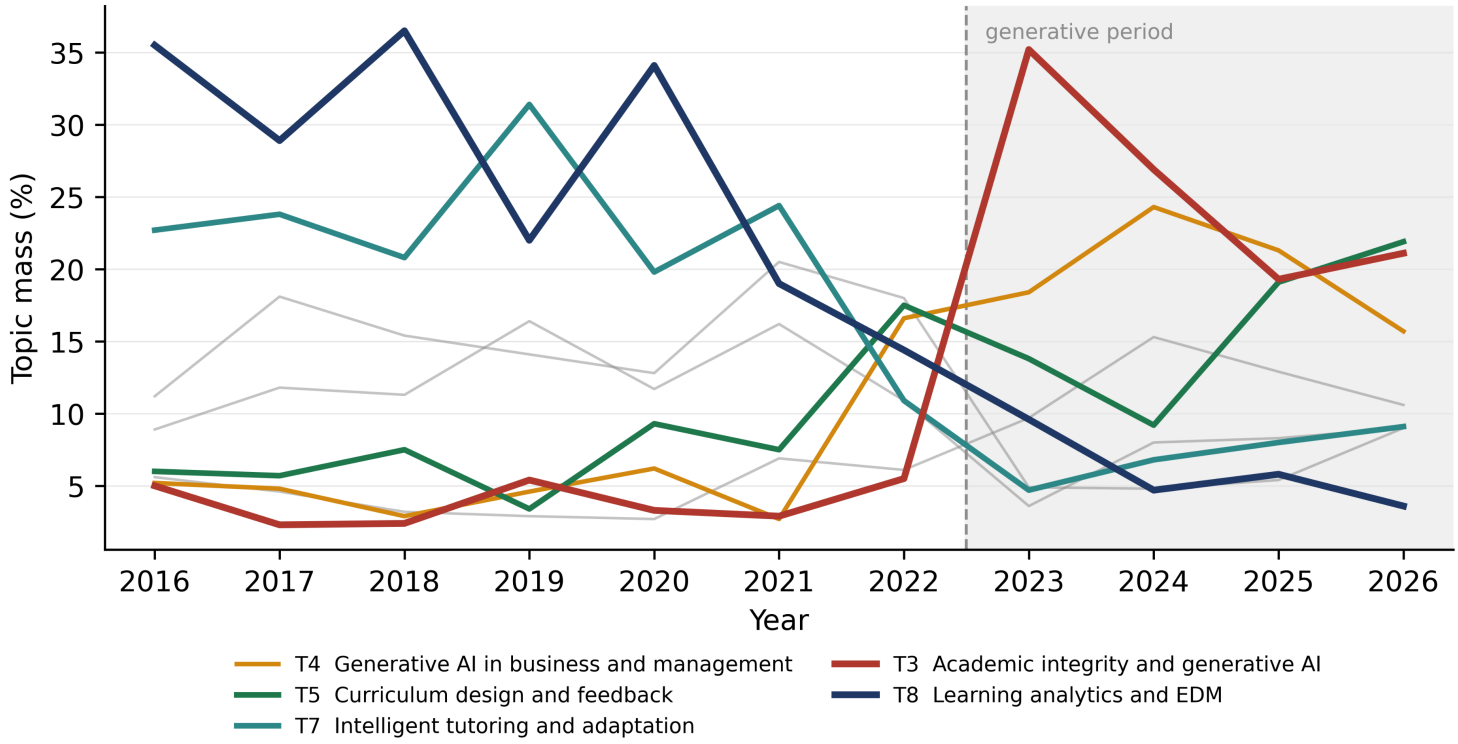


Figure 3

Topic shares over time. Two redistributions are visible: one at the period boundary, one within the generative period.

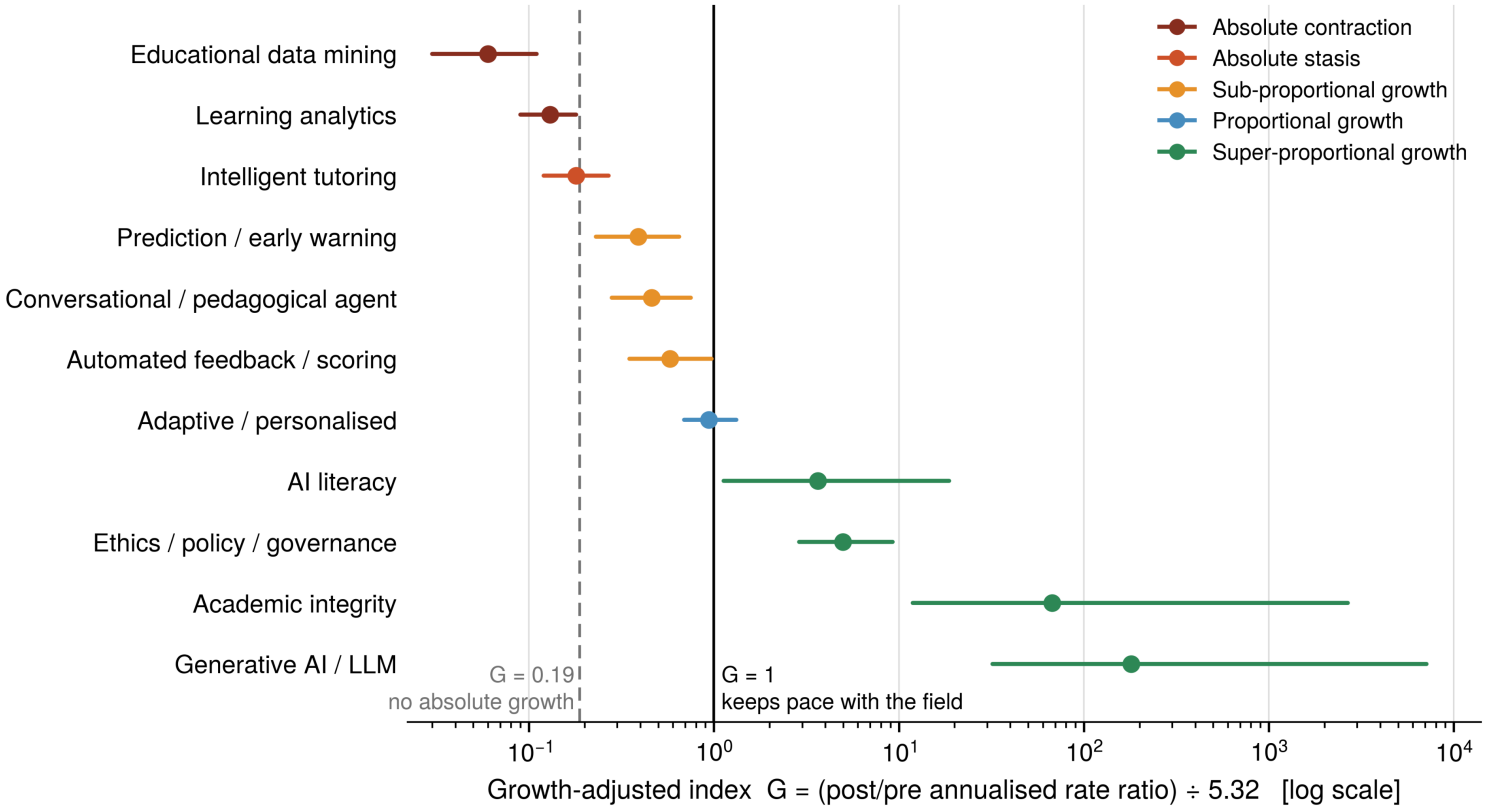


Figure 4

Growth-adjusted index with 95% confidence intervals. Values left of the dashed line indicate no absolute growth; values overlapping the solid line indicate growth at the field's own rate.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- S1PRISMA2020checklist.docx
- SupplementaryDatamanualcoding.xlsx
- S2PRISMAflowdiagram.pdf
- HSSCmanuscriptR20910marked.docx
- SupplementaryInformationREVISEDmarked.docx