

Did LLMs colonize scientific titles? Fifty years of titular colonicity

Magdalena Igras-Cybulska

magda.igras@gmail.com

Interactive Crafts sp. z o.o.

Artur Cybulski

Interactive Crafts sp. z o.o.

Research Article

Keywords: scientific titles, punctuation, language model evaluation, natural language processing, OpenAlex, information retrieval

Posted Date: October 12th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11137993/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Did LLMs colonize scientific titles?

Fifty years of titular colonicity

Magdalena Igras-Cybulska^{1*} and Artur Cybulski¹

^{1*}Interactive Crafts Research, Interactive Crafts sp. z o.o., ul. Marii Curie-Skłodowskiej 3, Lublin, 20-029, Poland.

*Corresponding author(s). E-mail(s): magda.igras@gmail.com;

Abstract

Colon-separated scientific titles were documented decades before generative AI. We examined their 50-year trajectory, prevalence in language-model outputs and relation to retrieval. We analysed 94,527,936 filtered English-language OpenAlex journal-article records from 1976–2025 and generated titles from 1,664 abstracts balanced by year and field, using neutral and field-convention prompts. Qwen 2.5 7B was the primary model; Mistral 7B and OLMo 2 7B provided secondary evaluations. Exact checkpoint revisions were fixed before offline generation. A separate E5 experiment compared colon-bearing titles with colon-to-space variants for 20,000 targets against one million candidate titles. Historical prevalence rose from 24.97% in 2022 to 28.95% in 2024; the provisional 2025 estimate was 34.00%. Qwen neutral-prompt titles contained colons in 41.2% of cases, versus 26.9% of published titles (paired difference 14.3 percentage points, 95% confidence interval 11.7–16.7). The field-convention prompt increased Qwen prevalence by 19.2 points (17.5–21.0); prompt contrasts were also positive in Mistral and OLMo. Colon-to-space replacement increased mean reciprocal rank at 100 from 0.3802 to 0.4007. Together, the studies characterise colon use as a longstanding, changing title convention that is frequent and prompt-sensitive in the evaluated model outputs and relevant to ranking in controlled retrieval.

Keywords: scientific titles, punctuation, language model evaluation, natural language processing, OpenAlex, information retrieval

1 Introduction

Scientific titles guide readers and information-retrieval systems in identifying relevant research, yet their wording and structure vary across disciplines and publication settings. The use of colons attracted scholarly attention as early as the 1980s [1] and was explicitly described as *titular colonicity* in a 1985 *BioScience* article [2]. Subsequent historical analyses have traced changes in title length and compound structures [3, 4]. However, reported associations between title form, downloads and citations are context-dependent and do not isolate punctuation effects from differences in article content or journal composition [5, 6].

Recent practices in using Large language models (LLMs) to support both preparation and retrieval of scientific articles provide a reason to revisit title measurement. Large-scale analyses have documented changes in selected vocabulary associated with LLM writing [7, 8]; related work reports increases after 2022 in selected verbs in article titles [9]. A temporal change in a familiar word or punctuation mark can also reflect editorial practices, indexing or publication composition. Such markers do not reveal how an individual title was written.

Recent studies have expanded research on both the long-term evolution of article titles and their reproduction by generative AI. A 100-year analysis of 4,400 titles from four leading medical journals documented changes in title length, structure, informational content and lexical diversity [10]. Controlled comparisons likewise show that titles generated from abstracts can differ systematically from their published counterparts. In medical samples, ChatGPT produced more frequent colons—100% versus 30% in one 50-pair comparison—or more two-part structures, although findings for title length differed between studies [11, 12]. Reader evaluations have instead examined perceived appeal, preference or willingness to click [13, 14], whereas automatic title-generation benchmarks compare outputs with reference titles, abstracts or human preferences [15, 16]. These strands measure different properties and do not establish a general advantage of any punctuation convention. Moreover, they do not combine broad cross-disciplinary historical evidence, multiple open-weight language models and prompting conditions, and a controlled test of whether replacing the colon itself changes dense-retrieval rankings. We therefore treat historical prevalence, model-generation behaviour and retrieval effects as related but analytically distinct questions.

Two measurement distinctions are central here. Literal ASCII-colon presence is not a validated subtitle boundary: it also captures technical notation, labels and metadata assembly. In generation tasks, a colon in an introductory label such as **Title:** belongs to the response format rather than necessarily to the proposed title. Similarly, success at retrieving a known source from an abstract-derived query is not equivalent to reader discovery in an operational search service. Retrieval performance varies with the task, encoder and comparison system [17].

We examine these distinctions in three components. First, we describe colon prevalence in a fixed OpenAlex snapshot covering 1976–2025, reporting the 2022–2024 change separately from the provisional 2025 endpoint [18]. Second, we compare published titles with titles generated from abstracts by three open-weight 7B models under neutral and field-convention prompts, using an explicit rule to separate the first proposed title from response labels and commentary. Third, we test whether replacing

the sole colon with a space changes known-item retrieval in a fixed E5 system while holding the words, their order and all competing titles constant.

Together, these components establish a 50-year empirical baseline across 94.5 million titles, quantify colon use in 29,952 model responses and isolate a punctuation-sensitive ranking effect in 20,000 retrieval targets. They address distinct descriptive and experimental questions; the historical analysis neither identifies AI-assisted publications nor estimates an effect of model adoption.

2 Materials and methods

2.1 Study design

The study comprised three components: a historical metadata analysis, a controlled title-generation experiment and a retrieval experiment. OpenAlex was selected for the historical analysis because its fixed, multidisciplinary snapshot covered the full 1976–2025 period and provided the titles, publication years, sources, field assignments and identifiers required for consistent cohort construction [18, 19]. Coverage and metadata quality may nevertheless vary across periods, fields and publishers.

The historical component estimated annual colon prevalence and examined sensitivity to language selection, publication dates, colon definitions, title length, field composition and journal composition. Four annual reference functions provided descriptive benchmarks for the recent trajectory. The principal **agreement** cohort comprised records classified as English by OpenAlex and both title-language detectors.

For the generation experiment, the sample, model checkpoints, prompts and generation settings were fixed before inference and before published-title outcomes were joined to the generated outputs. Title extraction and uncertainty estimation are described below. For the retrieval experiment, the model, index, query construction, target sample, contrast, metric and uncertainty procedure were fixed before held-out rankings were examined. Neither experiment was publicly preregistered.

The historical estimand was annual literal-colon prevalence among eligible records in the fixed snapshot. Generation estimands compared colon presence in generated title text with published titles and compared the field-convention prompt (P1) with the neutral prompt (P0), giving equal weight to the 416 publication-year-by-field strata. The retrieval estimand was the mean paired RR@100 difference in the eligible held-out target frame under the fixed E5 checkpoint, query rule and index. Model outputs were generated in September 2026 from abstracts of articles published in 2010–2025; publication year therefore identifies the source article rather than the date of generation.

2.2 Dataset and unit of analysis

The source was the OpenAlex Works Parquet release dated 26 June 2026 [18, 19]. The download contained 2,446 files and 510,372,821 records. Its manifest SHA-256 was 4b5142cb47f22fb5f6e91fc330ed800e355e8bb9421b10d812394780b9a1443d. File manifests, download receipts, schemas and derived-table checksums preserve provenance.

The 1976–2025 window contained 416,128,344 records. The working-population rules required a nonempty source title after whitespace checking, `language=en`, `type=article`, primary-source type `journal`, and `is_xpac=false` [20]. Neither a DOI, an abstract, a field assignment nor open-access status was required. Retracted articles were not excluded. These filters yielded 100,331,652 working records. Of these, 99,729,050 had a resolved title measurement. Unresolved records remained in quality accounting but not in prevalence denominators.

The observational unit is an OpenAlex Work record. The historical corpus has *not* been deduplicated into verified unique publications. DOI-sharing flags were preserved, and sensitivity analyses either removed records whose DOI was shared within the working population or required a recognised, singleton DOI. The former retains records without a usable DOI; the latter does not. A DOI-based exclusion is not a complete resolution of publication versions or near-duplicates. The duplicate groups used in the retrieval frame do not retrospectively change these historical denominators. Supplementary Fig. S9 shows the nested selection from the full snapshot to the main cohort.

2.3 Title preprocessing and language cohorts

The analyses used the stored conservatively cleaned title, `title_clean`, and the saved binary colon measurement from working-corpus construction. Cleaning applied HTML-entity decoding once, removed recognised formatting tags outside preserved MathML blocks, and collapsed whitespace. Empty cleaned titles, preserved MathML, unresolved markup or residual encoded entities left the primary measurement unresolved. Later stages did not translate titles, reconstruct missing subtitles or introduce new punctuation. The full cleaned-title string was checked against the saved literal-colon flag for every measurement-resolved record during the measurement-sensitivity stage. Counts refer to this versioned metadata representation, not to an independent transcription from publisher PDFs.

Title-language screening used `fastText lid.176.bin` and CLD2, accessed through `fasttext-predict` 0.9.2.4 and `pyclid2` 0.42 [21, 22]. The `fastText` model SHA-256 was `7e69ec5451bc261cc7844e49e4792a85d7f09c06789ec800fc4a44aec362764e`. For language prediction only, titles were NFC-normalised, Unicode letters and combining marks were retained, other characters were replaced with spaces, and whitespace was collapsed. The colon outcome and stored title were unchanged. Unresolved, empty, overlength or letterless inputs were not forced into an English category.

The **reference** cohort consists of measurement-resolved records passing the original OpenAlex metadata filters. The **ft_top1** cohort additionally requires an English first `fastText` label. The main **agreement** cohort requires English predictions from both detectors, without a further confidence cutoff. The **supported** sensitivity additionally requires at least 20 letters and three input tokens, two `fastText` predictions, a first-label score of at least 0.8, a first–second score margin of at least 0.2, and CLD2’s reliable flag with at least 90% English byte coverage. The **strict** sensitivity raises the score, margin and coverage thresholds to 0.95, 0.5 and 95%, respectively. These are operational selection rules, not calibrated probabilities of a title being English. CLD2 was designed for longer web text rather than short titles[22]. Because detector

accuracy was not independently validated on scientific titles, the language cohorts are operational filters rather than accuracy-ranked samples. Language-detector agreement can retain shared errors and exclude valid short or technical English titles; the results therefore describe this filtered population rather than all English scientific titles.

2.4 Colon definition and structural sensitivity rules

The primary historical outcome is at least one ASCII colon, U+003A, in the cleaned title. It includes potential subtitles, ratios, times, biological notation, study labels and metadata title–subtitle concatenation, without assigning these functions. It does not substitute visually similar Unicode characters. Three nested sensitivity rules use the same cohort denominator:

1. H1: exactly one ASCII colon, at least one Unicode letter group on each side, and no recognised technical pattern.
2. H2: H1 plus whitespace immediately after the colon.
3. H3: H2 plus at least two letter groups on each side.

Here H1–H3 name *heuristic title-form rules*, not confirmatory hypotheses. A letter group matches a Unicode letter followed by zero or more letters or combining marks. It is not a linguistically validated word. Technical-pattern hints include URL schemes, a number–colon–number pattern with optional whitespace, and a double colon. Rejecting a heuristic candidate changes its binary outcome to zero; it does not remove the record from the denominator. These rules identify observable forms, not validated semantic title–subtitle boundaries.

2.5 Crossref matching and publication dates

The Crossref [23] comparison used full-file DOI matching. The March 2026 Crossref extraction contained 179,536,204 records in 35,908 derived parts. All working OpenAlex records were retained during matching, including unmatched and ambiguous cases. Analyses requiring a single Crossref record used the saved unique-match status rather than multiplying OpenAlex rows by multiple matches.

Stored comparisons distinguished agreement with the Crossref main title, agreement with a constructed main-title-plus-subtitle representation, other comparable titles, and noncomparable metadata. Joining the title and subtitle fields can itself introduce a colon, so agreement with the combined representation does not establish that Crossref deposited that punctuation. The main-title agreement subset avoids that construction when examining robustness. Agreement across databases remains metadata consistency, not an independent historical or semantic ground truth.

Crossref **issued**, **published**, **published-online** and **published-print** events were examined separately. Partial dates retained their available precision. Same-year comparisons and removal of OpenAlex dates on 1 January or on the first day of any month were sensitivity filters, not proof that the remaining dates were accurate. No months or days were imputed. The completed analyses use the original OpenAlex publication year. Alternative Crossref-year summaries are conditional on the same eligible OpenAlex-window records with usable event years, not complete independent

Crossref-year populations. The date checks support reporting at annual resolution; they do not validate quarterly comparisons.

2.6 Annual rates, fields and journal composition

Annual prevalence was computed as the number of qualifying titles with the relevant colon indicator divided by the number with resolved measurement in that cohort and year. Each assigned record contributed once through its primary-topic field among the 26 OpenAlex fields[24]. Missing field assignments remained visible rather than being reassigned. The analyses use the 26 field identifiers and display labels saved with the snapshot. These assignments describe the snapshot classification, not a separate historical taxonomy for each publication year. Automated field assignments can be wrong or differ in a later snapshot, particularly for interdisciplinary records, so field estimates are conditional on these saved labels.

Two field-standardisation specifications used weights proportional to pooled **agreement** counts in 2018–2022. The known-field specification included the 26 assigned fields; the all-record specification also included unassigned field as a separate stratum and used the same weights for all four colon definitions. Both hold publication composition fixed rather than assigning equal weight to fields. Figures and tables identify the specification used. A missing positive-weight stratum yields an unavailable estimate rather than a zero rate or silent within-year renormalisation.

The journal analysis first restricted sources to the upstream source-presence panel for 2010–2025, then required positive eligible counts in every year within the analysed cohort. The **agreement** panel contained 17,076 journals. Observed publication weights, fixed publication weights pooled over 2018–2022, and equal-journal weights were examined separately on each panel. Other cohorts can have different eligible journal sets. Requiring presence through 2025 selects surviving sources and does not establish complete journal coverage.

For 2022–2024 and 2022–2025, differences within the same journal panel were decomposed exactly as

$$\Delta p = \sum_j \bar{w}_j \Delta p_j + \sum_j \bar{p}_j \Delta w_j, \quad (1)$$

where bars denote the mean of the two endpoint values. The first term measures within-journal rate change and the second changing journal publication shares. This is additive accounting, not a causal decomposition. Journal breadth was summarised for all panel journals and for those with at least 20 or 100 eligible records in each endpoint year. Concentration summaries divide the largest positive within-journal contributions by the sum of positive contributions, not by the signed net change.

2.7 Descriptive reference trajectories and historical comparisons

Four ordinary-least-squares reference functions were fitted to annual prevalence: linear logit, quadratic logit and linear probability-scale trends over 2010–2022, and a linear-logit trend over 2015–2022. Each year received equal regression weight; article counts were not regression weights. The linear-logit 2010–2022 model was the main descriptive reference. Models were fitted only through the stated origin, then extrapolated to

2023–2025. The reported gap is

$$G_y = 100\{p_y - \hat{p}_y^{\text{reference}}\}. \quad (2)$$

Endpoint gaps for 2024 and 2025 and mean gaps over 2023–2024 and 2023–2025 are reported separately. These exploratory reference functions provide descriptive benchmarks rather than an interrupted-time-series effect estimate. Their spread reflects specification dependence, not a confidence or prediction interval. Linear-scale extrapolations outside $[0, 1]$ were flagged rather than clipped.

Historical rolling-origin comparisons preserved each model’s degree and training-window length, using comparison windows ending by 2022. Their ranks and empirical quantiles provide retrospective context only. Overlapping windows are dependent, and present-day snapshot coverage and retrospectively selected cohorts remain part of every calculation. The historical comparisons describe annual patterns and do not test a causal breakpoint or between-field differences. The millions of records within each year improve the precision of the recorded proportions but do not increase the number of annual observations.

2.8 Title length and sensitivity to selection

Title length was measured as Unicode code points of `title_clean`, including spaces and punctuation, not words or encoded bytes. Mean positive length was described over 2010–2025. Length-only and joint field-length standardisation used the same pooled **agreement** 2018–2022 weights across language variants. Coarse length bins were 1–50, 51–100, 101–150, 151–200, 201–300 and 301 or more characters. Fine bins covered 25-character intervals to 500 characters and a final 501-or-more category. Missing field was an explicit joint stratum.

The common-support specification retained strata present in every year of 2010–2025 and renormalised their reference weights once. It did not alter weights separately in each outcome year. Missing or nonpositive lengths were excluded from standardisation and means but remained in coverage accounting. Joint field-length adjustment and journal composition adjustment concern different populations; they are not one simultaneous adjustment for all characteristics. Length may itself reflect stylistic change, so conditioning on length is not an identified correction for confounding or an estimate of an LLM effect.

2.9 Title-generation experiment

The generation sample crossed publication years 2010–2025 with the 26 OpenAlex fields, yielding 416 year–field strata. From duplicate-group representatives with structurally valid reconstructed abstracts, a seeded hash selected up to 32 candidates per stratum without reference to title punctuation. Records were excluded if the complete sequence of lexical tokens from the published title occurred contiguously in the abstract. Model-facing files contained only a pseudonymous input identifier and abstract text; published titles, colon indicators, Work IDs, DOIs and source IDs were stored separately.

Of 12,974 candidates, 12,677 passed automated screening requiring English as the top fastText prediction and a rendered prompt length of no more than 4,032 tokens

in each model tokenizer. The first four eligible records in the hash-based order were selected from every stratum, producing 1,664 inputs. Published titles and their colon indicators were linked only after sample selection was complete.

Qwen2.5-7B-Instruct was the primary model, and Mistral-7B-Instruct-v0.3 and OLMo-2-1124-7B-Instruct provided secondary evaluations [25–27]. Checkpoint revisions and runtime versions are reported in Supplementary Methods. The checkpoints were downloaded before inference and run locally and offline using their native chat templates and 8-bit `bitsandbytes` quantisation. The system message was ‘You assist with scientific writing.’ The neutral prompt (P0) requested one title based only on the supplied abstract and instructed the model to return only the title. The field-convention prompt (P1) additionally stated ‘Follow the title conventions of [field name].’ Neither prompt mentioned colons, subtitles or the historical findings.

For every abstract–model–prompt combination, three stochastic repetitions were generated with base seeds 20260913, 20260914 and 20260915. Sampling used temperature 0.7, top- p 0.9, top- k 50, one beam and a maximum of 64 new tokens. Condition-specific sampling seeds were derived deterministically from the input and generation identifiers. Decoded responses were normalised to NFC after harmonising line endings and trimming outer whitespace; internal wording and punctuation were retained. The resulting panel contained 29,952 outputs (1,664 inputs $\times$ 3 models $\times$ 2 prompts $\times$ 3 repetitions).

The title-level outcome was the presence of at least one ASCII colon in the first proposed title, extracted by a deterministic rule defined after inspection of response formatting. The rule removed recognised introductory labels and outer formatting, retained the first nonempty title line, and excluded subsequent commentary or alternative suggestions and specified inline explanatory parentheses. It retained punctuation within the selected title and did not exclude any output. The leading labels `Title`, `Proposed title` and `Suggested title` accounted for 1,368 non-title-only colon positives; four further such positives occurred in other introductions or commentary, all in Mistral P1. The extraction rule and response-level audit are provided in Supplementary Methods and Data. This is an operational text-extraction rule, not an independently validated semantic annotation. Estimates based on complete, unedited responses are reported as a sensitivity analysis in Supplementary Table S5.

The primary comparison was the stratum-balanced paired risk difference between Qwen P0 titles and the corresponding published titles. The prompt comparison contrasted Qwen P1 with Qwen P0; analogous Mistral and OLMo comparisons were secondary. For each condition, the three repetition indicators were first averaged within the abstract–model–prompt cell. Differences were then averaged with equal weight across the 416 year–field strata.

Two-sided 95% intervals used 2,000 paired, stratified bootstrap resamples. Within each stratum, four input identifiers were sampled with replacement and carried with their published-title outcome and all generated conditions. Because each stratum contained only four inputs, bootstrap deviations from the observed contrast were multiplied by $\sqrt{4/3}$ before percentile limits were calculated, correcting the ordinary within-stratum bootstrap’s finite-sample variance shortfall. Analytic intervals based on unbiased within-stratum sample variances provided a comparison (Supplementary

Table S9). This small-stratum correction was adopted during analysis, not fixed before generation. For a broader eligible-abstract interpretation, the intervals assume exchangeability within year–field strata and retain equal stratum weights. They preserve pairing, do not treat repeated generations as independent articles, and do not cover variation across other models, prompts or new generation runs.

A sensitivity analysis used the archived strict-format criterion: responses had to be nonempty, consist of one nonempty line, have no leading title label and not be enclosed in matching quotation marks. Colon presence was measured in the extracted title for qualifying responses. For a given contrast, an input was retained when at least one qualifying repetition was available for every required generated condition. Because format compliance is itself model behaviour, these estimates describe a selected subset rather than an alternative primary population.

Repeat agreement, text identity and matching with the linked published title were assessed on the extracted title text. Matching was evaluated after NFC normalisation and, separately, after case-folding and whitespace normalisation. Failure to match a published title does not rule out exposure to the source article during model training.

2.10 Within-title punctuation retrieval experiment

The retrieval experiment asked whether retaining a colon changes the ability to recover a paper when the title words and their order are otherwise fixed. The primary contrast compared an unchanged H3-eligible title with the same title after replacing its single colon with a space. H3 requires exactly one ASCII colon, so titles with multiple colons were ineligible as targets; the intervention therefore replaced the sole literal colon in each eligible target. This target restriction did not apply to unchanged distractor titles. Secondary conditions replaced the colon with a dash, reversed the two components, or retained only the first component. The last two conditions changed information order or quantity and therefore did not isolate punctuation.

The primary index comprised one million title-only documents from the **agreement**, recognised-singleton-DOI population of 2010–2025. Abstracts were not required for distractors. Duplicate components linked records by shared normalised DOI, exact normalised full title or exact normalised abstract. One representative was selected per connected component before target groups were assigned to development and held-out test splits. For each query, only the designated target title changed; the other 999,999 titles remained original and identical across conditions. The original target was removed from competitor counts and reinserted in the tested condition, and alternative versions of one target never competed in the same ranking.

Each query was the earliest eligible sentence in the reconstructed source abstract under the prespecified sentence-selection rule: 12–60 lexical tokens after identifier cleaning, automatic English screening and rejection of a complete contiguous copy of the title. These are synthetic document-derived queries, not observed reader searches. Primary retrieval used `intfloat/e5-base-v2` revision `f52bf8ec8c7124536f0efb74aca902b2995e5bcd`, with query and passage prefixes, masked mean pooling, 768-dimensional L2-normalised `float32` embeddings and exact blocked inner-product ranking [28, 29]. TF32 was disabled. A target was excluded from all compared conditions

if its query or any compared title exceeded the 512-token limit; no condition was silently truncated.

BM25 with a fixed punctuation-separating, order-insensitive lexical analyser, $k_1 = 1.2$ and $b = 0.75$, served as a negative implementation/intervention control[30]. The original, space, dash and reversed-clause conditions yielded the same token multiset under this analyser and were required to produce identical rankings. A difference would have indicated an implementation error, not a scientific punctuation effect. The primary outcome was the paired mean difference in reciprocal rank truncated at rank 100,

$$\Delta\text{MRR@100} = \frac{1}{n} \sum_i [RR_{100,i}(\text{colon-to-space}) - RR_{100,i}(\text{original})], \quad (3)$$

where $RR_{100} = 1/r$ for $r \leq 100$ and zero otherwise. Throughout the manuscript and Supplementary Information, $RR@100$ differences are reported as colon-to-space minus original. Positive values favour colon-to-space, whereas negative values favour the original title. Results outside the first 100 were not assigned rank 101. Secondary outcomes were Success@10, Success@100, exact-rank summaries, the fractions improved, unchanged and worsened, all five title variants, and descriptive breakdowns by publication year, OpenAlex field and lexical title-length bin.

Development rankings used 5,000 groups. The prespecified precision rule, based on the development variance and an approximate half-width target of 0.005 MRR units, implied 2,947 groups; the prespecified minimum therefore fixed the held-out test at 20,000 groups. The test comprised the first 20,000 eligible held-out rows ordered by the prespecified target priority and then Work ID. Held-out rankings were examined only after the analysis plan was finalised. The sole prespecified confirmatory comparison was the E5 mean paired RR_{100} difference above. Its two-sided 95% interval used 2,000 paired percentile-bootstrap resamples with a fixed seed, retaining complete target groups. On the displayed scale, a confidence interval entirely above zero favours the colon-to-space title, one entirely below zero favours the original, and an interval containing zero is inconclusive. These intervals describe resampling variability across eligible target groups conditional on the observed index, query rule and encoder. The finite 20,000-target mean itself is fully observed; an interval for a broader same-frame mean requires exchangeability of target groups. It does not cover alternative encoders, indices, queries, indexing errors or real reader populations. Supplementary Data also provide a distribution-free Hoeffding interval for independent bounded target-group differences. All subgroup results and comparisons among the other variants were descriptive, with no confirmatory subgroup tests.

Reported paired direction counts refer to the $RR@100$ contrast. Equal $RR@100$ includes pairs for which both ranks exceed 100, and therefore does not imply an identical exact rank. This is distinct from equal similarity scores within one ranking. Candidates were ranked by decreasing similarity score, with ties resolved by increasing canonical OpenAlex Work ID.

2.11 Computational environment and validation scope

Historical analyses used Python 3.13.15, DuckDB 1.5.5, PyArrow 25.0.1, NumPy 2.5.3 and Matplotlib 3.10.8. Generation used Python 3.12.14, PyTorch 2.11.0, Transformers 5.17.0 and `bitsandbytes` 0.50.2 on an NVIDIA RTX 5080 Laptop GPU with CUDA 12.8. Retrieval used the same GPU class and the fixed local E5 checkpoint; semantic rankings used the GPU, whereas BM25 and final synthesis used the CPU. Validation during the experimental pipeline covered the complete generation panel, all 20,000 retrieval pairs and 100,000 target-variant ranking rows. Historical and retrieval summaries were checked against the exported tables. This pipeline validation is distinct from the distributed-data reproducibility workflow, whose scope is stated under Code availability.

ChatGPT and Codex (OpenAI) assisted manuscript preparation, code development and computational checks. These assistance tools are distinct from the three experimental model checkpoints evaluated in the generation study.

3 Results

3.1 Analytical populations and coverage

The working population contained 100,331,652 records. Measurement was resolved for 99,729,050 (99.40%), leaving 602,602 unresolved records outside the prevalence denominator. The main **agreement** cohort contained 94,527,936 records, 94.78% of the resolved population, with 20,843,509 colon-bearing titles over the full period. It included 94,145,521 records with a field assignment and 382,415 without one. The 2010–2025 portion contained 55,977,008 records. These are OpenAlex record counts, not confirmed unique-publication counts.

Within the working population, 561,488 records belonged to shared-DOI groups. Removing shared-DOI records from **agreement** retained 94,146,974 records; requiring a recognised singleton DOI retained 80,987,531. A single Crossref match was available for 80,082,722 main-cohort records. The substantially smaller supported and strict language cohorts are reported as sensitivity populations, not as independently established higher-accuracy samples (Table 1).

3.2 Recorded prevalence through 2024 and the provisional 2025 endpoint

Colon prevalence in the main cohort was 16.22% in 1976, 19.97% in 1990, 20.38% in 2000 and 19.07% in 2010, showing that the 50-year trajectory was not monotonic (Fig. 1; Supplementary Table S6). Prevalence subsequently reached 22.47% in 2015, 24.97% in 2022 and 26.30% in 2023. Between 2022 and 2024 it increased by 3.98 percentage points, from 963,660/3,859,498 records (24.97%) to 1,267,958/4,379,383 (28.95%). The provisional 2025 estimate was 1,480,108/4,352,758 (34.00%).

The 2022–2025 difference was +9.04 percentage points, including +5.05 points between 2024 and 2025. Because the 2025 denominator was smaller despite a larger colon count and may still change through indexing or deposit lag, we treat that endpoint and all associated contrasts as provisional.

Table 1 Historical analytical populations, 1976–2025.

Population	Records
Working records after metadata filters	100,331,652
Resolved measurement (reference)	99,729,050
fastText English top label	97,142,795
Two-detector English agreement (main)	94,527,936
Supported language sensitivity	41,900,455
Strict language sensitivity	1,718,725
Agreement, excluding shared DOI	94,146,974
Agreement, recognised singleton DOI	80,987,531
Agreement, single Crossref match	80,082,722
Agreement, same Crossref issued year	75,884,909

Rows overlap and must not be summed. All language variants require a resolved colon measurement. The unit is an OpenAlex Work record, not a verified unique publication.

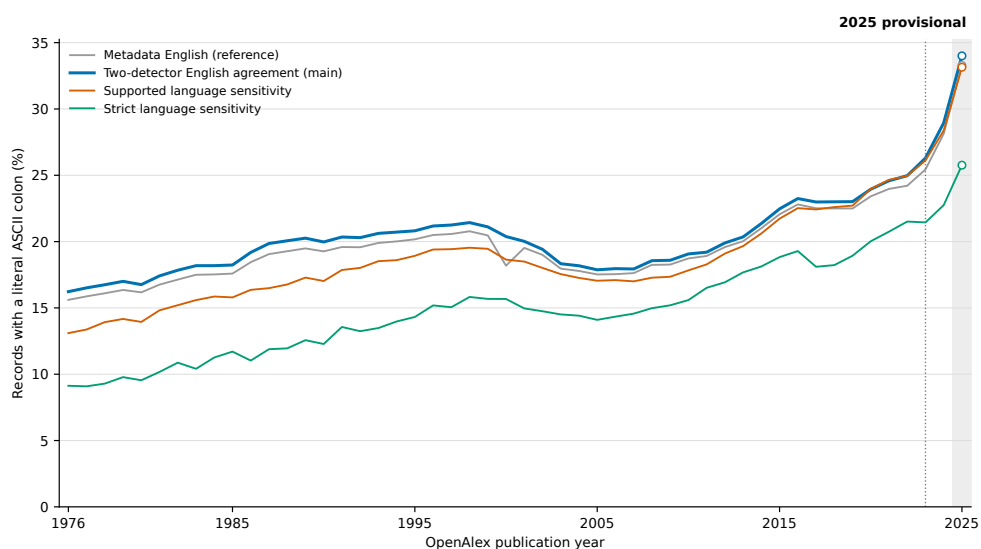


Fig. 1 Annual literal-colon prevalence across four language cohorts, 1976–2025. Each curve uses its own denominator and the OpenAlex publication year. The vertical line marks 2023; the shaded 2025 endpoint is provisional. In the agreement cohort, prevalence increased by 3.98 percentage points from 2022 to 2024 and by 5.05 points from 2024 to 2025.

3.3 Field and journal composition

All 26 assigned fields showed positive 2022–2024 differences, ranging from +0.49 points in Mathematics to +7.81 in Dentistry. Computer Science increased by +7.05 points. With fixed 2018–2022 field weights and unassigned field retained as a stratum, prevalence rose from 24.89% to 28.80%, a +3.91-point change. This was close to the unadjusted increase of +3.98 points, indicating that changing field publication shares alone did not account for the recorded increase under this standardisation. Figure 2 presents the known-field and continuing-journal series.

Extending the comparison to the provisional 2025 endpoint, all 26 fields again showed positive differences from 2022, ranging from +2.14 points in Mathematics to +13.38 in Dentistry. The field-standardised increase was +9.06 points, reaching 33.95%. Supplementary Table S1 reports the 2022–2025 field comparisons and the unassigned category. These field contrasts are descriptive; the study does not test differences between their magnitudes.

The continuing-journal analysis provides a separate composition check. In the balanced panel of 17,076 journals, prevalence increased from 26.77% in 2022 (2,310,085 records) to 30.27% in 2024 (2,440,077 records), a +3.50-point change. The exact decomposition assigned +3.82 points to within-journal rate changes and −0.32 points to changing publication shares. Among journals with at least 20 eligible records at both endpoints, 8,273 of 12,930 (63.98%) increased; among those with at least 100 records, 2,847 of 3,895 (73.09%) increased. Thus, the within-journal increase was also evident without using the provisional 2025 endpoint (Supplementary Table S10).

Extending the same panel to 2025 (2,540,102 records), prevalence reached 34.77%, an increase of +8.01 points from 2022. Within-journal changes contributed +7.80 points and changing publication shares +0.21 points. These decompositions concern the selected panel, not all main-cohort records, and do not identify the causes of changes within journals.

Among 12,584 panel journals with at least 20 records in both endpoint years, 9,489 (75.41%) increased, 2,946 decreased and 149 were unchanged. The median change was +5.89 points. Among 3,824 journals with at least 100 records at both endpoints, 3,320 (86.82%) increased. The 100 largest positive within-journal contributors accounted for 27.46% of the sum of positive contributions, not of the signed net increase. These summaries indicate broad participation within the selected journal panel; all involve the provisional 2025 endpoint.

3.4 Sensitivity to title form, length, language and dates

The 2022–2024 increase remained under all three nested form rules (Table 2). H1 prevalence increased from 24.01% to 28.00%, H2 from 23.86% to 27.87%, and H3 from 22.44% to 26.39%. Their increases of +4.00, +4.01 and +3.95 points, respectively, were close to the literal-colon increase of +3.98 points. For 2022–2025, the H3 increase was +8.78 points compared with +9.04 for literal-colon presence (Supplementary Fig. S10). Persistence under these rules shows that the descriptive result is not confined to the broadest punctuation definition, but does not validate a semantic title–subtitle interpretation.

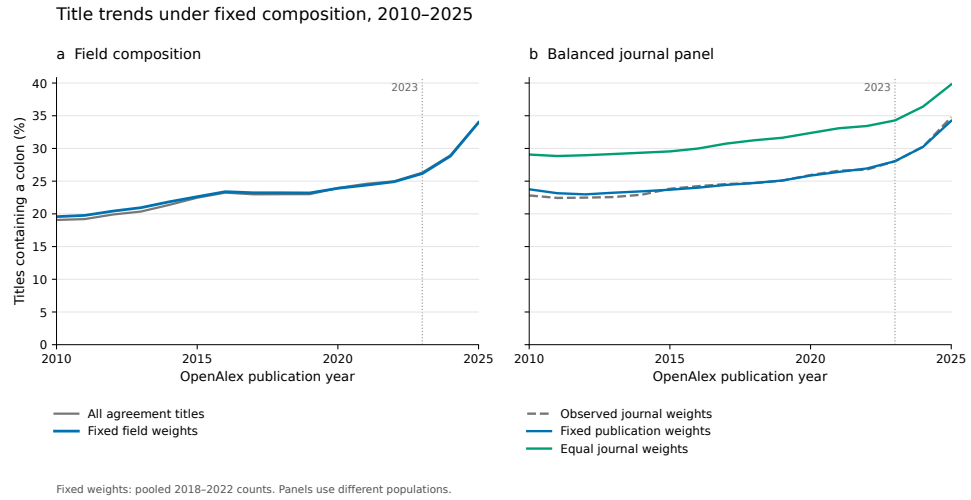


Fig. 2 Composition checks in the agreement cohort. **a**, Unadjusted and fixed-field prevalence for the 26 assigned fields. **b**, Observed-publication, fixed-publication and equal-journal estimates in 17,076 journals with eligible records in every year from 2010 to 2025. Reference weights were pooled over 2018–2022; 2025 estimates are provisional.

Table 2 Colon-definition sensitivity through 2024 and at the provisional 2025 endpoint.

Measure	2022 (%)	2024 (%)	2025 (%)	2022–2024 change (pp)	2022–2025 change (pp)
Literal ASCII colon	24.97	28.95	34.00	+3.98	+9.04
H1: one colon and letters	24.01	28.00	32.94	+4.00	+8.93
H2: H1 and following space	23.86	27.87	32.81	+4.01	+8.95
H3: H2 and two letter groups per side	22.44	26.39	31.22	+3.95	+8.78

Denominators are unchanged across definitions: 3,859,498 records in 2022, 4,379,383 in 2024 and 4,352,758 in 2025. Rejected candidates remain in the denominator with outcome zero. H1–H3 are structural heuristics, not statistical hypotheses or validated semantic categories. Changes use unrounded proportions.

Joint field-length standardisation on common 2010–2025 support yielded 2022–2024 increases of +3.32 points with coarse length groups and +3.23 with fine groups. The corresponding 2022–2025 increases were +7.75 and +7.54 points. Mean title length increased from 89.08 characters in 2010 to 103.37 in 2022 and 108.10 in 2025. Reweighting therefore attenuated, but did not remove, the recorded increase. Title length can itself be part of stylistic change, so these estimates should not be interpreted as a causal adjustment.

Language selection affected magnitude. For 2022–2024, the reference cohort increased by +3.92 points, the supported cohort by +3.43 and the strict cohort by +1.24; the corresponding 2022–2025 increases were +9.07, +8.20 and +4.25 points. The strict population contained only 1,718,725 records across 50 years, compared with 99,729,050 in the reference cohort. Its smaller increase reflects a strongly selected population and is not an independently validated correction to the main estimate.

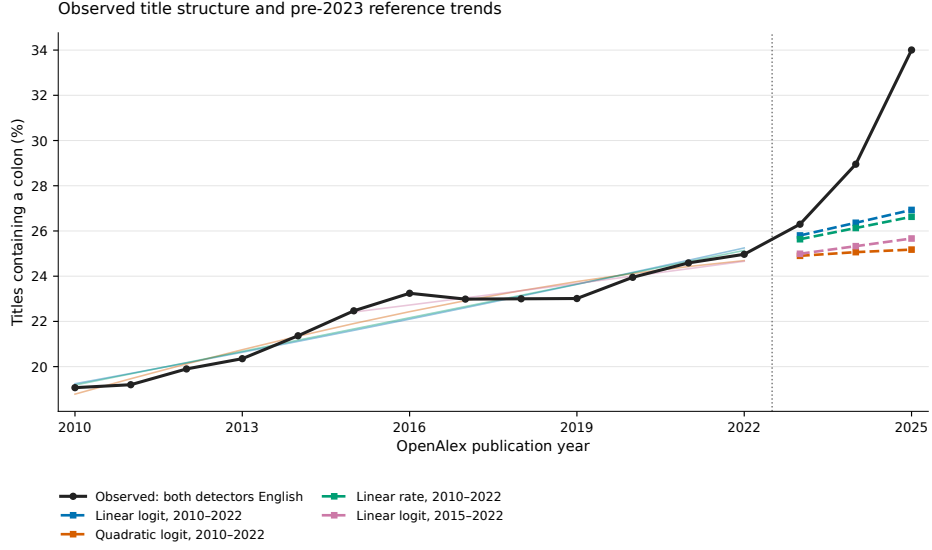
Restricting to exact agreement with the Crossref main-title representation yielded a 2022–2024 increase from 24.99% to 29.02% (+4.03 points). Requiring agreement between Crossref issued year and OpenAlex year yielded +3.99 points; agreement with Crossref online-publication year yielded +3.86 points; and excluding OpenAlex dates on the first day of a month yielded +4.02 points. These subsets also retained positive 2022–2025 differences (+9.11, +9.01, +9.11 and +9.09 points, respectively). Requiring a recognised singleton DOI gave a 2022–2025 increase of +9.04 points. The filters change the eligible population and test metadata consistency rather than establishing independent ground truth. Supplementary source tables report the full sensitivity series and their coverage.

3.5 Comparison with descriptive reference trajectories

For 2024, the main linear-logit reference fitted over 2010–2022 predicted 26.36%, compared with the observed 28.95%, a gap of +2.59 points. Across the four specifications, the 2024 gaps ranged from +2.59 to +3.89 points, and mean 2023–2024 gaps ranged from +1.54 to +2.64 points (Supplementary Table S7). For the main reference, 7 of 33 earlier comparison windows had mean gaps at least as large as the +1.54-point mean for 2023–2024, and four earlier windows had endpoint gaps larger than the +2.59-point gap in 2024.

At the provisional 2025 endpoint, the main reference predicted 26.93%, compared with the observed 34.00%, giving a gap of +7.07 points. The remaining specifications produced endpoint gaps of +7.38 to +8.83 points and mean 2023–2025 gaps of +3.62 to +4.70 points (Supplementary Table S7; Fig. 3).

Only when the provisional 2025 endpoint was included did the endpoint and three-year mean gaps rank first among the corresponding rolling-origin comparisons for all four models. The historical prominence of this departure therefore depends strongly on 2025; the increase through 2024 was not unprecedented under the main reference. Because the historical windows overlap and come from the same snapshot, these ranks provide descriptive context rather than a calibrated breakpoint test.



Faint lines: fitted past. Dashed lines: extrapolations. No causal effect or uncertainty interval is shown.

Fig. 3 Observed annual prevalence and four reference trajectories fitted through 2022, with extrapolations for 2023–2025. The longer models use 2010–2022 and the shorter linear-logit model uses 2015–2022. The 2025 endpoint is provisional.

3.6 Colon prevalence in generated titles

Published titles contained a literal colon in 448/1,664 cases (26.92%). After first-title extraction, Qwen 2.5 7B neutral-prompt titles contained colons in 2,058/4,992 outputs (41.23%). The paired difference from the published-title rate was 14.30 percentage points (95% CI 11.74–16.66; Supplementary Table S8; Fig. 4). With the field-convention prompt, the Qwen rate increased to 3,017/4,992 (60.44%), giving a P1-minus-P0 difference of 19.21 points (95% CI 17.48–21.04).

Neutral-prompt prevalence was 76.96% for Mistral 7B v0.3 and 64.50% for OLMo 2 7B, exceeding the published-title rate by 50.04 points (95% CI 47.59–52.49) and 37.58 points (95% CI 35.06–39.99), respectively. Under the field-convention prompt, prevalence was 80.43% for Mistral and 67.41% for OLMo; the corresponding prompt effects were 3.47 points (95% CI 2.22–4.67) and 2.90 points (95% CI 1.33–4.32). Results by source publication year and field appear in Supplementary Figs. S6 and S7.

Strict-format compliance was 90.3% under P0 and 95.9% under P1 for Qwen. Its strict-format P0–published estimate was 13.9 points, close to the 14.3-point full-panel result. Compliance was substantially lower for Mistral (4.8% under P0 and 7.7% under P1) and intermediate for OLMo (55.8% and 39.5%); the corresponding subset results and coverage are reported in Supplementary Table S2 and Fig. S5.

First-title extraction changed the Qwen neutral–published difference from 16.65 to 14.30 points and the P1–P0 difference from 17.45 to 19.21 points. For OLMo, the prompt difference changed from 11.32 to 2.90 points. All nine intervals remained above

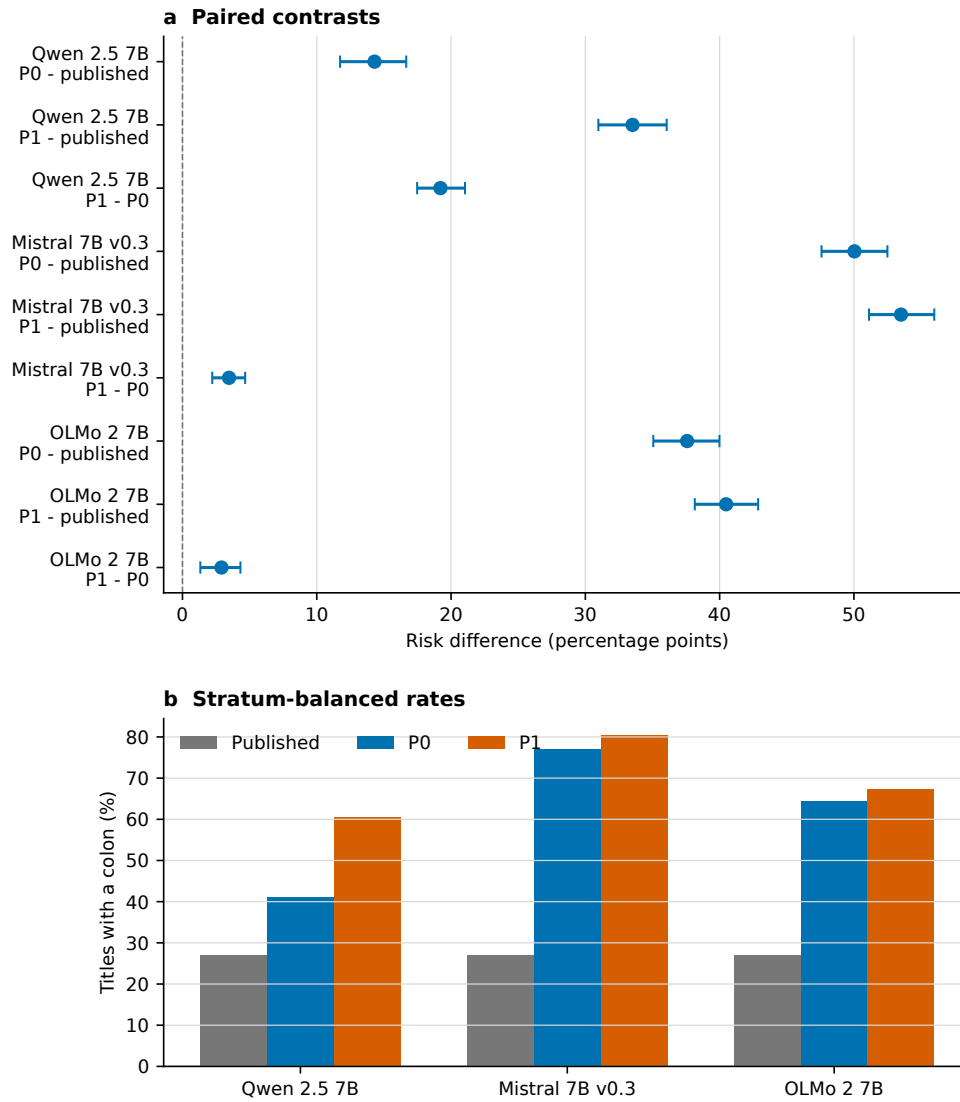


Fig. 4 Colon use in published and extracted generated titles. **a**, Paired risk differences with rescaled stratified-bootstrap 95% confidence intervals. **b**, Stratum-balanced rates. Each model-prompt estimate averages three repetitions per abstract; all 29,952 outputs were retained. The extraction rule separates the first proposed title from response labels and commentary.

zero under both outcome definitions and under the analytic uncertainty calculation (Supplementary Tables S4, S5 and S9).

Colon status agreed across all three repetitions in 75.2–78.7% of Qwen and Mistral cells and 62.6–65.0% of OLMo cells (Supplementary Table S3; Fig. S8). No generated title exactly matched its linked published title after NFC normalisation; five OLMo outputs matched after case-folding and whitespace normalisation.

3.7 Colon-to-space differences in E5 retrieval

The held-out test included 20,000 targets ranked against the fixed one-million-title index. Mean E5 RR@100 increased from 0.38021 for original titles to 0.40066 after replacing the colon with a space, a paired difference of +0.02044 (95% CI 0.01851–0.02243; Table 3; Fig. 5).

Colon-to-space produced higher RR@100 for 4,670 targets (23.35%), equal RR@100 for 13,046 (65.23%) and lower RR@100 for 2,284 (11.42%). Success@10 increased from 51.28% to 53.14%, Success@100 from 70.86% to 72.56%, the median exact rank improved from 9 to 8 and the 75th percentile from 178 to 140. The average improvement was therefore accompanied by unchanged RR@100 for most targets.

Table 3 Held-out E5 retrieval metrics for the primary title contrast.

Metric	Original	Colon to space	Colon-to-space minus original
Mean RR@100	0.38021	0.40066	+0.02044
Success@10 (%)	51.28	53.14	+1.86 pp
Success@10 (n/N)	10,256/20,000	10,628/20,000	
Success@100 (%)	70.86	72.56	+1.71 pp
Success@100 (n/N)	14,171/20,000	14,512/20,000	
Median exact rank	9	8	−1
75th percentile rank	178	140	−38

Differences are calculated before rounding as colon-to-space minus original. Positive RR@100 and success differences favour colon-to-space. For exact ranks, lower values are better, so negative differences favour colon-to-space. Success and rank summaries are secondary outcomes without separately estimated confidence intervals.

Mean E5 RR@100 was 0.39229 for colon-to-dash, 0.39026 for retaining only the first component and 0.38875 for reversing the components. The first-component variant increased mean RR@100 relative to the original but reduced Success@100 to 66.31%. Under BM25, the original, space, dash and reversed-component variants had identical rankings and mean RR@100 of 0.39539, confirming the expected behaviour of the punctuation-separating, order-insensitive analyser. The first-component variant yielded 0.31403.

The space-minus-original contrast was positive in all 16 publication years and in 25 of 26 assigned fields; the exception was Veterinary, with 21 targets. It declined from +0.03763 among titles with 1–10 lexical tokens to +0.00732 among titles with at least 21 tokens.

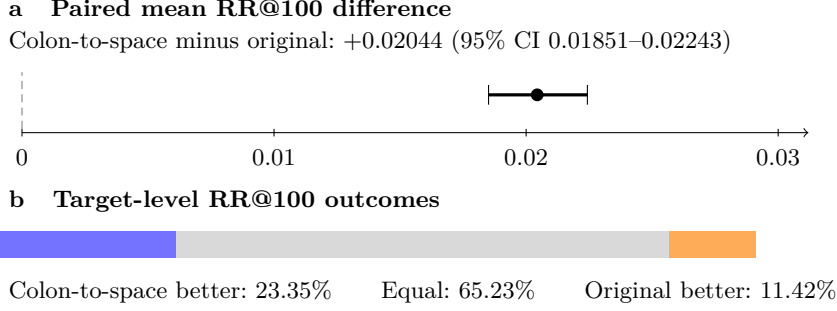


Fig. 5 Held-out E5 effect of replacing the target colon with a space. **a**, Paired mean RR@100 difference with a percentile-bootstrap 95% confidence interval. **b**, Target-level outcome shares. The analysis included 20,000 targets and one million candidate titles.

4 Discussion

Across 94.5 million OpenAlex records, colon prevalence increased through 2024; the evaluated language models used colons more frequently than the linked published titles; and replacing the colon with a space improved ranking in the controlled E5 experiment. These components concern different processes—historical publishing practice, model-generation behaviour and retrieval-system response. They do not form a tested causal sequence, but together show that colon use changes over time, depends on the model and prompt, and can affect ranking in a specified retrieval system.

The historical results place the recent increase within a much longer and non-monotonic trajectory. Colon use in titles attracted scholarly attention in the early 1980s [1], and the expression *titular colonicity* appeared explicitly in 1985 [2]. Later studies documented changes in title length and compound structures across periods and disciplines [3, 4, 31]. Our analysis extends this literature by measuring colon prevalence over 50 years in a large multidisciplinary corpus. Between 2022 and 2024, prevalence increased by 3.98 percentage points. The increase remained similar after fixed-field standardisation (+3.91 points), joint field-length reweighting (+3.23 to +3.32 points) and application of the three structural colon rules (+3.95 to +4.01 points). Crossref title and date subsets produced comparable changes. The provisional 2025 estimate added a further 5.05 points, but its magnitude may change as indexing and deposits mature. Neither this temporal pattern nor its departure from the fitted reference trajectories identifies an effect of LLM adoption.

All evaluated model-prompt conditions produced more title-text colons than the published-title baseline. This direction is consistent with earlier studies reporting frequent colons or two-part structures in titles generated under particular ChatGPT settings [11, 12]. However, the magnitude differed substantially among models and prompts: the field-convention instruction increased Qwen prevalence by 19.21 points but produced much smaller increases in Mistral and OLMo. Such variation is consistent with the broader observation that title-generation outputs depend on the model, task formulation and evaluation procedure [15, 16]. Separating the first title from response labels and commentary also changed several contrasts, most notably for OLMo, while leaving all nine confidence intervals above zero. Evaluations of generated titles should

therefore distinguish punctuation within the proposed title from punctuation introduced by response formatting.

Replacing the sole colon with a space increased E5 mean RR@100 by 0.02044 and improved Success@10 and Success@100 by 1.86 and 1.71 percentage points, respectively. Dense-retrieval performance is known to depend on the encoder, query construction, corpus and evaluation task [17, 28]. Accordingly, our result should be interpreted as a property of the evaluated E5 configuration rather than a universal retrieval advantage of removing colons. The effect was also heterogeneous: RR@100 was unchanged for 65.23% of targets, improved for 23.35% and worsened for 11.42%. Identical BM25 rankings for the original, space, dash and reversed-component variants confirmed the expected behaviour of the punctuation-separating, order-insensitive lexical control, rather than showing that punctuation is irrelevant to lexical retrieval generally.

Several limitations define the scope of these findings. Historical estimates refer to one OpenAlex snapshot, operational language filters and literal or structural punctuation rules; they do not identify semantic subtitle boundaries or AI-assisted publications. The 2025 endpoint remains provisional, and the annual reference functions are descriptive extrapolations rather than causal counterfactuals. Published titles are observational references rather than verified human-only controls, and prior model exposure to source articles cannot be excluded. Generation results depend on the specified checkpoints, prompts, chat templates, quantisation and decoding settings, while Mistral’s strict-format analysis represents a small selected subset. Finally, the retrieval experiment measured known-item recovery using one encoder, one index and abstract-derived queries, not reader behaviour, general-purpose search or title quality.

The findings nevertheless have direct methodological implications. Colon use alone should not be used to infer AI authorship because the convention long predates generative models and varies across fields and publication periods. Studies of generated titles should report model and prompt dependence and separate the proposed title from response labels and commentary before measuring its punctuation. Retrieval systems should test punctuation normalisation empirically rather than assume that separators are computationally neutral. These recommendations concern measurement and system evaluation; they do not establish that colon-bearing or colon-free titles are intrinsically better for authors or readers.

Future work should repeat the historical analysis using later frozen snapshots and equal-lag publication cohorts to determine how much of the provisional 2025 increase persists after metadata maturation. A stratified semantic annotation study could distinguish subtitles from technical notation, study labels and metadata concatenation, thereby testing how closely the structural H1–H3 rules represent rhetorical title divisions. Generation experiments should cover additional model families, sizes, prompt formulations and publication periods under preregistered outcomes. Retrieval robustness should be examined across dense, sparse and hybrid systems, multiple encoders, alternative query sources and operational tie-handling procedures. Finally, preregistered reader studies or search experiments could measure comprehension, recall, perceived relevance and selection behaviour, extending earlier work on title visibility and reader response [13, 14].

4.1 Conclusion

Across 94.5 million OpenAlex records, colon use followed a non-monotonic 50-year trajectory but increased by 3.98 percentage points between 2022 and 2024; the larger 2025 estimate remains provisional. In the controlled generation experiment, all three language models produced colon-bearing titles more frequently than the linked published titles after separating the first proposed title from response labels and commentary, although prevalence and prompt effects differed substantially among models. In held-out E5 retrieval, replacing the sole colon with a space increased mean $RR@100$ by 0.02044, but most targets remained unchanged and the result is conditional on the evaluated system. Together, these findings establish colon use as a longstanding, context-dependent title convention with measurable consequences for both language-model output and semantic retrieval. They do not show that LLM adoption caused the recent historical increase or that a colon identifies AI authorship. Instead, title punctuation should be measured explicitly and evaluated separately across historical, generative and retrieval settings.

Declarations

Funding. No funding was received for this study.

Supplementary information. Supplementary Information contains detailed methods, historical field and journal tables, generation robustness and quality tables, comparisons between complete-response and extracted-title outcomes, and retrieval and generation sensitivity figures. Supplementary Data and Code contain machine-readable results and intermediate count cells, generated response text and extraction decisions, retrieval ranks and query/index inputs, identifier linkage, source-provenance manifests, protocols, software specifications, computational modules and the workflow for validation, recalculation and figure generation.

Data availability. The OpenAlex and Crossref sources are identified in Methods and the references. The machine-readable tables in Supplementary Data and Code are enumerated in `00_file_index.csv`. Historical data include journal-by-year and field-by-length-by-year count cells for recalculating composition and length-standardisation analyses. Generation data include OpenAlex Work IDs and linked published titles for all 1,664 inputs, together with complete decoded responses, extracted first titles, extraction flags, text hashes and numerical outcomes for all 29,952 responses. Generation source abstracts can be reconstructed by Work ID from the identified fixed OpenAlex snapshot. Retrieval data include 20,000 paired target ranks, 100,000 target-variant ranking rows, target queries and title variants, and the one-million-title index with Work-ID and duplicate-group linkage. Source-provenance manifests identify the archived inputs. The record-level historical corpus, embeddings and model weights are not redistributed.

Code availability. Supplementary Data and Code provide public copies of 38 upstream Python modules covering OpenAlex and Crossref processing, language detection, cohort construction, model inference, generation analysis, embedding construction, retrieval ranking and result synthesis, together with public-run helpers. The archive also contains study protocols, software specifications and SHA-256 manifests. The

tested command `python run_all.py --require-complete` validates the distributed files and identifier linkage, reapplies title extraction to all 29,952 responses, recomputes generation contrasts and uncertainty, recalculates retrieval estimates and uncertainty from target-level ranks, reconstructs journal composition and field-length standardisation from count cells, and regenerates figures. Public-run integrity checks distinguish archived execution records from files produced by a new run. This verifies statistical reproduction from the distributed intermediate data; it is not an end-to-end rerun of source extraction, language detection, model inference or embedding-based ranking. Those stages additionally require the identified external datasets, pinned model revisions, substantial storage and compatible computational environments. The tested scope and upstream requirements are documented in `README.md`, `VERIFICATION.md` and `code/PIPELINE.md`.

Consent to Publish declaration. Not applicable.

Author contributions. Magdalena Igras-Cybulska: conceptualisation, methodology, analysis, data curation and writing; Artur Cybulski: conceptualisation, software, validation, analysis, data curation and writing.

Competing interests. The authors declare no competing interests.

Ethics and Consent to Participate declarations. Not applicable. The study analysed bibliographic metadata and locally generated model outputs and did not recruit participants or analyse person-level attributes or outcomes. OpenAlex can contain author names and affiliations, but these were not analysed as person-level variables. The study involved no intervention with human participants.

References

- [1] Dillon, J. T. In pursuit of the colon: A century of scholarly progress: 1880–1980. *The Journal of Higher Education* **53**, 93–99 (1982).
- [2] Neff, S. E., Lemons, J., Webster, J. R. & Burgess, R. L. Titular colonicity. *BioScience* **35**, 394–395 (1985).
- [3] Lewison, G. & Hartley, J. What’s in a title? numbers of words and the presence of colons. *Scientometrics* **63**, 341–356 (2005).
- [4] Jiang, F. K. & Hyland, K. Titles in research articles: Changes across time and discipline. *Learned Publishing* **36**, 239–248 (2023).
- [5] Jamali, H. R. & Nikzad, M. Article title type and its relation with the number of downloads and citations. *Scientometrics* **88**, 653–661 (2011).
- [6] Letchford, A., Moat, H. S. & Preis, T. The advantage of short paper titles. *Royal Society Open Science* **2**, 150266 (2015).

- [7] Kobak, D., González-Márquez, R., Horvát, E.-Á. & Lause, J. Delving into LLM-assisted writing in biomedical publications through excess vocabulary. *Science Advances* **11**, eadt3813 (2025).
- [8] Kousha, K. & Thelwall, M. How much are LLMs changing the language of academic papers after ChatGPT? a multi-database and full text analysis. *Scientometrics* (2026). Published online 3 April 2026.
- [9] Comas-Forgas, R., Koulouris, A. & Kouis, D. ‘AI-navigating’ or ‘AI-sinking’? an analysis of verbs in research articles titles suspicious of containing AI-generated/assisted content. *Learned Publishing* **38**, e1647 (2025).
- [10] Hao, J. & Dong, J. Mapping 100-year trajectories: Research article titles in four top general medical journals. *English for Specific Purposes* **83**, 222–240 (2026). URL <https://doi.org/10.1016/j.esp.2026.05.002>.
- [11] Sebo, P., Nie, B. & Wang, T. How ChatGPT writes scientific titles in medical research: structural and content differences compared to human authors. *Journal of the Medical Library Association* **114**, 222–237 (2026).
- [12] Ibrahim, S. K. M. & Mahmoud, Z. A. Z. Generative AI in academic writing: a comparison of human-authored and ChatGPT-generated research article titles. *Humanities and Social Sciences Communications* **13**, 394 (2026).
- [13] Shvets, A. Improving scientific article visibility by neural title simplification. In *Proceedings of the 8th International Workshop on Bibliometric-enhanced Information Retrieval*, CEUR Workshop Proceedings, vol. 2345, pp. 140–147. CEUR-WS.org, Cologne, Germany (2019). URL <https://ceur-ws.org/Vol-2345/paper13.pdf>.
- [14] Sebo, P., Nie, B. & Wang, T. Can ChatGPT write better scientific titles? a comparative evaluation of human-written and AI-generated titles. *F1000Research* **14**, 1470 (2026). Version 2.
- [15] Çano, E. & Bojar, O. Two huge title and keyword generation corpora of research articles. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 6663–6671. European Language Resources Association, Marseille, France (2020). URL <https://aclanthology.org/2020.lrec-1.823/>.
- [16] Chen, Y. & Eger, S. Transformers go for the LOLs: Generating (humorous) titles from scientific abstracts end-to-end. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pp. 62–84. Association for Computational Linguistics, Bali, Indonesia (2023). URL <https://aclanthology.org/2023.eval4nlp-1.6/>.
- [17] Thakur, N., Reimers, N., Rücklé, A., Srivastava, A. & Gurevych, I. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models.

- In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, vol. 1 (2021). URL <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/65b9eea6e1cc6bb9f0cd2a47751a186f-Abstract-round2.html>.
- [18] Priem, J., Piwowar, H. & Orr, R. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts (2022). URL <https://arxiv.org/abs/2205.01833>. arXiv:2205.01833.
 - [19] OpenAlex. Snapshot (2026). URL <https://help.openalex.org/access/snapshot/>. Documentation accessed 7 September 2026.
 - [20] OpenAlex. Works: Attributes (2026). URL <https://help.openalex.org/data/works/attributes/>. Documentation accessed 7 September 2026.
 - [21] fastText. Language identification (2026). URL <https://fasttext.cc/docs/en/language-identification.html>. Model documentation accessed 12 September 2026; executed model and checksum are reported in Methods.
 - [22] CLD2Owners. Compact language detector 2 (2026). URL <https://github.com/CLD2Owners/cld2>. Software documentation accessed 12 September 2026; analysis used pycld2 0.42.
 - [23] Crossref. REST API (2026). URL <https://www.crossref.org/documentation/retrieve-metadata/rest-api/>. Documentation accessed 7 September 2026.
 - [24] OpenAlex. Topics (2026). URL <https://help.openalex.org/data/topics/>. Documentation accessed 7 September 2026.
 - [25] Qwen Team. Qwen2.5-7B-Instruct: Model card (2026). URL <https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>. Model card accessed 16 September 2026; analysis used revision a09a35458c702b33eeacc393d103063234e8bc28.
 - [26] Mistral AI. Mistral-7B-Instruct-v0.3: Model card (2026). URL <https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3>. Model card accessed 16 September 2026; analysis used revision c170c708c41dac9275d15a8fff4eca08d52bab71.
 - [27] Allen Institute for AI. OLMo-2-1124-7B-Instruct: Model card (2026). URL <https://huggingface.co/allenai/OLMo-2-1124-7B-Instruct>. Model card accessed 16 September 2026; analysis used revision 470b1fba1ae01581f270116362ee4aa1b97f4c84.
 - [28] Wang, L. *et al.* Text embeddings by weakly-supervised contrastive pre-training (2022). URL <https://arxiv.org/abs/2212.03533>. arXiv:2212.03533.
 - [29] intfloat. e5-base-v2: Model card (2026). URL <https://huggingface.co/intfloat/e5-base-v2>. Model card accessed 7 September 2026; analysis used revision f52bf8ec8c7124536f0efb74aca902b2995e5bcd.

- [30] Robertson, S. & Zaragoza, H. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval* **3**, 333–389 (2009).
- [31] Hartley, J. Planning that title: Practices and preferences for titles with colons in academic articles. *Library & Information Science Research* **29**, 553–568 (2007).

Supplementary Files

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

- [Supplementary.pdf](#)
- [SupplementaryDataandCodePart2.zip](#)
- [SupplementaryDataandCodePart3.zip](#)
- [SupplementaryDataandCodePart1.zip](#)