

Supplementary Information for Did LLMs colonize scientific titles? Fifty years of titular colonicity

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

¹Interactive Crafts sp. z o.o.

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

S1 Supplementary methods

Historical rates concern the filtered, automatically English-classified metadata cohort; all 2025 endpoints are provisional. Generation outcomes were evaluated at two levels: complete model responses and the extracted first proposed title. The title-text measure is used for the main results, while complete-response estimates provide a formatting sensitivity analysis. All 29,952 outputs remain in the analysis. All retrieval RR@100 differences are reported as colon-to-space minus original; positive values favour colon-to-space. Retrieval results apply only to the fixed E5 checkpoint, index and abstract-sentence query construction.

S1.1 Frozen models and execution environment

The generation panel contained three Apache-2.0 instruction-tuned checkpoints, each frozen locally before experimental generation. The primary model was Qwen/Qwen2.5-7B-Instruct, revision [a09a35458c702b33eeacc393d103063234e8bc28](#). Supportive evaluations used mistralai/Mistral-7B-Instruct-v0.3, revision [c170c708c41dac9275d15a8fff4eca08d52bab71](#), and allenai/OLMo-2-1124-7B-Instruct, revision [470b1fba1ae01581f270116362ee4aa1b97f4c84](#). Every downloaded file was inventoried and SHA-256-verified. All inference was local and offline.

The final generation runtime used Windows 11, Python 3.12.14, PyTorch 2.11.0+cu128, Transformers 5.17.0, Accelerate 1.15.0, Safetensors 0.8.0 and bitsandbytes 0.50.2 on an NVIDIA GeForce RTX 5080 Laptop GPU (16,303 MiB reported memory; CUDA 12.8; driver 592.00). Models were loaded one at a time with bitsandbytes LLM.int8 quantisation, threshold 6.0, no FP16 weight copy, and

bfloat16 for non-linear layers. Exact checkpoint identifiers and runtime versions were recorded before output aggregation.

S1.2 Input-frame construction and blinding

The candidate population consisted of duplicate-group representatives from the historical retrieval frame with structurally valid reconstructed abstracts, nonmissing OpenAlex field, publication years 2010–2025, and the frozen upstream eligibility rules. Within each publication-year-by-field stratum, up to 32 candidates were selected by ascending SHA-256 of a frozen seed and canonical Work ID. Sampling priority did not use the published title or its punctuation. Candidates were excluded if the complete sequence of lexical published-title tokens occurred contiguously in the abstract, because the task was title generation rather than extraction.

The initial balanced frame contained 13,312 candidates (32 in each of 416 strata). After the copy exclusion, 12,974 model-facing candidates remained. Their model-facing records contained only `input_id`, an arm label, abstract text and an abstract hash. Work IDs, DOIs, source IDs, published titles and published-title colon outcomes remained in a private linkage file. Automated screening read no published title or colon indicator. It required fastText `lid.176.bin` top-1 language to equal English and the model-native rendering of a predeclared screening prompt to contain at most 4,032 tokens in every tokenizer, leaving 12,677 eligible candidates. The first four eligible records under the frozen rank were selected in each stratum. All 416 strata were complete, yielding 1,664 inputs. Selection was committed before published-title outcomes were unblinded.

S1.3 Prompts and generation settings

All models received a system message and one of two user messages through their native chat template with a generation prompt appended. The exact messages were:

System: You assist with scientific writing.

P0: Write one title for a scientific article based only on the abstract provided. Return only the title.

```
<abstract>
{abstract_text}
</abstract>
```

P1: Write one title for a scientific article based only on the abstract provided. Return only the title. Follow the title conventions of `{field_name}`.

```
<abstract>
{abstract_text}
</abstract>
```

Neither prompt named the colon, subtitles or any result from the historical analysis. A frozen preflight rendered P0 and P1 for all selected inputs and confirmed that both fitted every model’s context without truncation. For each model and prompt, three stochastic repetitions used base seeds 20260913, 20260914 and 20260915. Each batch received a deterministic sampling seed derived from a SHA-256 namespace and its

frozen condition identifiers. Generation used sampling (temperature 0.7, top- p 0.9, top- k 50), one beam, one returned sequence, 1–64 new tokens, repetition penalty 1.0, cache enabled and `clean_up_tokenization_spaces=False`. The complete panel contained 29,952 outputs.

Decoded output was normalised by replacing CRLF with LF, applying Unicode NFC, and trimming outer whitespace. Internal wording and punctuation were not repaired. An output was instruction-compliant if it was nonempty, contained exactly one nonempty line, did not begin with a label such as “Title:”, “Proposed title:” or “Suggested title:”, and did not have matching outer quotation marks.

S1.4 Generation estimands and uncertainty

The complete-response outcome counts at least one ASCII colon (U+003A) anywhere in the complete normalised response. The title-text outcome counts a colon in the first proposed title, extracted using the deterministic `first-title-v1` rule. This rule was defined after inspection of response formatting and was not fixed before generation. It first removes a complete outer Markdown bold wrapper, if present. It then removes one recognised leading label, including the case-insensitive pattern:

```
^\s*(?:title|proposed\s+title|suggested\s+title)
\s*(?:|[\-\-\])\s*
```

The two displayed lines form one continuous regular expression. Additional recognised introductions are “A Title for the Scientific Article:” and the same three standard label phrases directly preceding an opening quotation mark. After removing a matching introduction and trimming whitespace, the rule retains the first nonempty remaining line; later commentary and alternative suggestions are not selected. An inline explanatory parenthesis beginning “this title” or “following the title conventions” is removed with the rest of that suffix. Matching outer quotation marks are stripped. Internal punctuation in the selected title is retained. Multiple suggestions, placeholders, unspecified titles and empty extractions are flagged rather than silently excluded.

The standard leading-label pattern matched 4,221 responses and accounted for 1,368 complete-response positives whose only colon belonged to the label. Four further non-title-only colon positives, all in Mistral P1, came from other introductions or commentary. Thus, 1,372 outcomes changed from 1 to 0 and none from 0 to 1, leaving 19,517 colon-positive extracted titles. The formatting screen included all 56 multiline responses, five inline explanatory suffixes and two extended introductions; these categories can overlap. The distributed response-level extraction audit covers all 29,952 outputs and records the selected title, rule and ambiguity flags, linked to the complete responses. It is an algorithmic measurement audit, not an independent human annotation. The panel retains all 1,664 abstracts, 416 strata, six model-prompt combinations and three repetitions. The rule does not assess semantic subtitle function or scientific title quality.

Published-title punctuation was measured separately in the linked metadata title; its authorship process was not observed. For each generated condition, three repetition outcomes were averaged within each abstract-model-prompt cell. Risk differences were formed within the linked abstract and then averaged within each of the 416 publication-year-by-field strata; strata received equal weight. Qwen P0 minus the

published title was the primary comparison, and Qwen P1 minus P0 was the secondary prompt comparison. Corresponding Mistral and OLMo comparisons and each P1-minus-published comparison were supportive.

For uncertainty, each bootstrap replicate samples four input identifiers with replacement independently within every stratum and carries each sampled input’s published-title outcome and all generated conditions together. Two thousand resamples use the fixed seed `colon-gen-historical-stratified-bootstrap-20260914-v1` and SHA256-counter draw algorithm. Sorted year-field strata and sorted input identifiers within strata determine draw order. The same resampling indices are used for complete-response and title-text outcomes. With four inputs per stratum, the ordinary within-stratum bootstrap variance is only three-quarters of the unbiased stratified variance estimate. We therefore rescale each bootstrap contrast $\hat{\theta}_b$ around the observed contrast $\hat{\theta}$ as

$$\hat{\theta}_b^{\text{scaled}} = \hat{\theta} + \sqrt{4/3} (\hat{\theta}_b - \hat{\theta}).$$

The 2.5th and 97.5th percentiles (type-7 quantiles) of the rescaled draws form the 95% intervals. This variance correction was specified during analysis rather than before generation.

For comparison, analytic intervals use $\hat{\theta} \pm 1.96 \text{ SE}$, where

$$\text{SE} = \frac{1}{H} \sqrt{\sum_{h=1}^H \frac{s_h^2}{n_h}}, \quad H = 416, \quad n_h = 4,$$

and s_h^2 is the sample variance of the paired abstract-level differences in stratum h , with denominator $n_h - 1$. Table S9 compares the two interval methods. Pairing is preserved, and repeated generations are not treated as independent articles. The realised 1,664-input panel mean is observed; interpretation of the intervals for broader eligible-abstract populations assumes exchangeability within the year-field strata and retains equal stratum weights. The intervals are conditional on the evaluated models, prompts, settings and realised repetitions, and do not quantify variation across other models, prompts or new generation runs.

The full-panel title-text analysis retains all 1,664 inputs. Per-protocol sensitivity analyses apply the archived strict response-format criterion to the complete response, measure colon presence in the extracted title, average only compliant repetitions, and retain an abstract for a contrast only when at least one relevant repetition is compliant in every required generated condition. Compliance is model behaviour and selects on emitted form, so these estimates describe a selected subset rather than replacing the full-panel analysis. Table S2 reports point estimates and coverage; uncertainty is not estimated for subsets containing singleton strata, whose within-stratum variances cannot be estimated. Table S4 distinguishes label-bearing responses, label-only colon positives and positives attributable to other non-title text.

Repeat stability and text identity were calculated on the extracted titles. Matching to the published title was evaluated after NFC normalisation and after case-folding plus whitespace normalisation; neither screen can exclude training-data exposure. Response-format compliance, generation-length and stopping diagnostics describe the complete

emitted response. The numerical outcome CSV includes complete-response and title-text colon indicators, title-text colon counts, response-format flags, text hashes and comparison flags. Its sequential input index links to the distributed response text and OpenAlex Work IDs. The package includes all complete responses, extracted titles and extraction decisions, allowing independent text-level checking and recalculation of title-text and complete-response contrasts, rescaled-bootstrap intervals and analytic uncertainty comparisons.

S1.5 Distributed-data reproducibility workflow

The accompanying package includes the machine-readable tables listed in `00_file_index.csv`, all 29,952 decoded model responses and extracted titles, and OpenAlex Work-ID linkage for the 1,664 generation inputs. It also supplies journal-by-year and joint field-by-length-by-year count cells, 20,000 paired retrieval ranks, 100,000 target-variant ranking rows, the target queries and title variants, and the one-million-title index. Numeric retrieval `doc_id` values are distinguished from canonical OpenAlex `work_id` values; the distributed mapping preserves index order and duplicate-group linkage.

Running `python run_all.py --require-complete` validates the indexed files, panel structure and identifier linkage; reapplies the title-extraction rule; recomputes the nine title-text and nine complete-response generation contrasts and their uncertainty estimates; recalculates retrieval metrics, paired bootstrap intervals and descriptive subgroup results from saved ranks; reconstructs journal composition and joint field-length standardisation from count cells; and regenerates figures. The journal analysis includes 2022–2024 and 2022–2025 comparisons on the same 2010–2025 panel (Table S10). Software requirements and machine-readable verification reports accompany the workflow.

The archive also contains 38 upstream analysis modules, public-run helpers, protocols and file manifests. Public-run integrity checks distinguish archived execution records from newly generated artifacts; a new run must verify its own files rather than match run-specific output hashes. The commands `python code/retrieval/test_new_run_integrity.py` and `python -m unittest discover -s tests -v` test receipt verification, tamper rejection, identifier mapping and statistical calculations with synthetic fixtures. These software tests are separate from recalculation using the distributed research data. The tested workflow establishes statistical reproduction from saved responses, ranks and count cells, not an independent rerun of OpenAlex or Crossref extraction, language detection, model inference, embedding construction or large-scale retrieval ranking. Full source-to-result reproduction additionally requires the external data, model checkpoints and computational environments described in `README.md` and `code/PIPELINE.md`.

S2 Supplementary tables

Table S1: Field-specific literal-colon prevalence in the agreement cohort. All 26 fields and the unassigned category are reported. Counts refer to each original OpenAlex publication year. The 2025 endpoint and every 2022–2025 change are provisional.

Field	N_{2022}	2022 (%)	N_{2025}	2025 (%)	Change (pp)
Agricultural and Biological Sciences	195,277	15.49	217,761	24.44	+8.95
Arts and Humanities	89,784	44.69	91,503	51.60	+6.92
Biochemistry, Genetics and Molecular Biology	225,789	18.20	251,045	27.33	+9.13
Business, Management and Accounting	102,483	36.73	125,560	48.01	+11.28
Chemical Engineering	13,811	15.28	18,020	20.67	+5.39
Chemistry	67,441	18.37	74,536	23.30	+4.93
Computer Science	209,882	19.01	252,464	31.18	+12.17
Decision Sciences	34,115	27.25	41,028	37.93	+10.68
Dentistry	22,783	32.69	26,882	46.06	+13.38
Earth and Planetary Sciences	67,105	23.55	68,581	27.88	+4.33
Economics, Econometrics and Finance	96,707	33.66	94,883	43.21	+9.55
Energy	44,122	14.80	51,529	22.00	+7.20
Engineering	551,932	11.37	645,508	18.33	+6.95
Environmental Science	212,885	22.72	238,326	32.12	+9.39
Health Professions	84,399	35.71	96,972	45.46	+9.76
Immunology and Microbiology	42,344	20.70	47,906	29.60	+8.90
Materials Science	156,518	14.03	175,656	20.09	+6.07
Mathematics	59,985	9.94	56,163	12.08	+2.14
Medicine	872,089	31.54	1,014,196	43.05	+11.51
Neuroscience	63,360	22.38	72,637	31.18	+8.81
Nursing	21,610	25.63	27,463	35.01	+9.38
Pharmacology, Toxicology and Pharmaceutics	16,252	20.34	19,049	29.71	+9.37
Physics and Astronomy	99,245	13.20	104,456	16.64	+3.44
Psychology	122,391	39.81	140,196	48.93	+9.12
Social Sciences	377,982	39.55	389,913	49.75	+10.20
Unassigned	2,868	0.24	3,315	0.84	+0.60
Veterinary	6,339	18.73	7,210	26.13	+7.41

Differences are percentage points calculated before rounding. No field-specific significance test was performed. Unassigned records do not constitute a twenty-seventh substantive field.

Table S2 Full-panel title-text estimates and response-format-restricted sensitivity estimates.

Model	Contrast	Full panel (pp)	Per protocol (pp)	Inputs retained	Strata retained
Qwen 2.5 7B	Neutral–published	14.3	13.9	1,581 (95.0%)	416 (100.0%)
Qwen 2.5 7B	Field–published	33.5	33.3	1,641 (98.6%)	416 (100.0%)
Qwen 2.5 7B	Field–neutral	19.2	19.6	1,566 (94.1%)	416 (100.0%)
Mistral v0.3	7B Neutral–published	50.0	41.2	166 (10.0%)	133 (32.0%)
Mistral v0.3	7B Field–published	53.5	43.1	245 (14.7%)	173 (41.6%)
Mistral v0.3	7B Field–neutral	3.5	1.5	110 (6.6%)	89 (21.4%)
OLMo 2 7B	Neutral–published	37.6	41.3	1,196 (71.9%)	412 (99.0%)
OLMo 2 7B	Field–published	40.5	47.3	910 (54.7%)	382 (91.8%)
OLMo 2 7B	Field–neutral	2.9	6.6	808 (48.6%)	370 (88.9%)

Full-panel estimates measure colon presence in the extracted first title and retain every output. Per-protocol estimates apply the archived strict response-format criterion to the complete response and measure punctuation in its extracted title. Point estimates are shown for all subsets; singleton strata prevent estimation of unbiased within-stratum variance for some subsets. Low coverage limits comparability with the full panel.

Table S3 Response-format compliance, title-text repeat stability and published-title matching.

Model	Prompt	Compliant (%)	Colon agree. (%)	Identical title texts (%)	Exact	Casefold
Qwen 2.5 7B	P0	90.3	76.7	9.6	0	0
Qwen 2.5 7B	P1	95.9	75.5	5.2	0	0
Mistral v0.3	7B P0	4.8	75.2	1.7	0	0
Mistral v0.3	7B P1	7.7	78.7	1.1	0	0
OLMo 2 7B	P0	55.8	62.6	0.5	0	2
OLMo 2 7B	P1	39.5	65.0	0.5	0	3

Colon agreement and identical-text rates use the extracted first titles across all three repetitions. Exact and casefold columns count matches among 4,992 extracted titles per condition; matching outer quotes are removed by the extraction rule. Compliance describes the complete emitted response under the archived strict-format criterion. Absence of an exact match does not establish absence from model training.

Table S4 Response-to-title extraction by model and prompt.

Model	Prompt	Complete response colon	Leading label	Label only	Other non-title only	Title-text colon	Title-text (%)
Qwen 2.5 7B	P0	2,175	215	117	0	2,058	41.2
Qwen 2.5 7B	P1	3,046	66	29	0	3,017	60.4
Mistral 7B v0.3	P0	4,040	852	198	0	3,842	77.0
Mistral 7B v0.3	P1	4,163	715	144	4	4,015	80.4
OLMo 2 7B	P0	3,450	542	230	0	3,220	64.5
OLMo 2 7B	P1	4,015	1,831	650	0	3,365	67.4
Total		20,889	4,221	1,368	4	19,517	65.2

Each model–prompt row contains 4,992 outputs; the total contains 29,952. Leading label refers to the three standard label patterns. Label only counts positives eliminated by that removal; other non-title only counts additional positives eliminated when separating the first title from introductions and commentary. A label can coexist with a title colon, so not all 4,221 labels change the binary outcome. Complete-response colon minus the two non-title-only columns equals title-text colon. The distributed response-level extraction audit provides the selected text and flags.

Table S5 Complete-response and title-text colon contrasts.

Model	Contrast	Complete-response difference, pp (95% CI)	Title-text difference, pp (95% CI)
Qwen 2.5 7B	Neutral–published	16.6 (14.1, 19.1)	14.3 (11.7, 16.7)
Qwen 2.5 7B	Field–published	34.1 (31.5, 36.6)	33.5 (31.0, 36.1)
Qwen 2.5 7B	Field–neutral	17.4 (15.6, 19.3)	19.2 (17.5, 21.0)
Mistral 7B v0.3	Neutral–published	54.0 (51.7, 56.4)	50.0 (47.6, 52.5)
Mistral 7B v0.3	Field–published	56.5 (54.1, 58.9)	53.5 (51.1, 56.0)
Mistral 7B v0.3	Field–neutral	2.5 (1.2, 3.7)	3.5 (2.2, 4.7)
OLMo 2 7B	Neutral–published	42.2 (39.8, 44.6)	37.6 (35.1, 40.0)
OLMo 2 7B	Field–published	53.5 (51.2, 55.7)	40.5 (38.1, 42.9)
OLMo 2 7B	Field–neutral	11.3 (9.8, 12.8)	2.9 (1.3, 4.3)

Both sets of estimates retain all outputs and use identical pairing, stratum weights and bootstrap draws. Intervals use the $\sqrt{4/3}$ rescaling of centred bootstrap contrasts described in Supplementary Methods. Title-text outcomes refer to the extracted first proposed title. Qwen P0–published is the primary comparison, and Qwen P1–P0 is the secondary prompt comparison.

Table S6 Selected annual counts and literal-colon prevalence in the agreement cohort.

Year	Records	Colon-bearing titles	Prevalence (%)
1976	443,393	71,940	16.22
1990	889,256	177,613	19.97
2000	1,388,564	283,020	20.38
2010	2,760,427	526,414	19.07
2015	3,154,672	708,816	22.47
2022	3,859,498	963,660	24.97
2023	4,005,631	1,053,431	26.30
2024	4,379,383	1,267,958	28.95
2025	4,352,758	1,480,108	34.00

Counts are exact. Percentages are rounded to two decimals; changes are calculated from unrounded proportions. All 50 annual observations are supplied in the source data. The 2025 endpoint is provisional because differential indexing and deposit lag have not been evaluated with a later snapshot or equal-lag design.

Table S7 Descriptive gaps from annual reference trajectories.

Reference function	2024 gap (pp)	Mean gap 2023–2024 (pp)	2025 gap (pp)	Mean gap 2023–2025 (pp)
Linear logit, 2010–2022	+2.59	+1.54	+7.07	+3.39
Quadratic logit, 2010–2022	+3.89	+2.64	+8.83	+4.70
Linear rate, 2010–2022	+2.82	+1.74	+7.38	+3.62
Linear logit, 2015–2022	+3.63	+2.47	+8.34	+4.42

Gaps are observed minus reference prevalence. Observed prevalence was 28.95% in 2024 and 34.00% in 2025. For the main linear-logit reference, 7 of 33 earlier windows had mean gaps at least as large as the 2023–2024 mean; four had larger endpoint gaps than in 2024. All four models ranked the recent endpoint and mean gaps first only when 2025 was included. All comparisons involving 2025 are provisional. Years receive equal fitting weight. The four specifications are not independent replications, and their spread is not a confidence interval.

Table S8 Literal-colon prevalence in published titles and extracted generated titles.

Model	Contrast	Baseline colon rate (%)	Comparison colon rate (%)	Difference, pp (95% CI)
Qwen 2.5 7B	Neutral–published	26.9	41.2	14.3 (11.7, 16.7)
Qwen 2.5 7B	Field–published	26.9	60.4	33.5 (31.0, 36.1)
Qwen 2.5 7B	Field–neutral	41.2	60.4	19.2 (17.5, 21.0)
Mistral 7B v0.3	Neutral–published	26.9	77.0	50.0 (47.6, 52.5)
Mistral 7B v0.3	Field–published	26.9	80.4	53.5 (51.1, 56.0)
Mistral 7B v0.3	Field–neutral	77.0	80.4	3.5 (2.2, 4.7)
OLMo 2 7B	Neutral–published	26.9	64.5	37.6 (35.1, 40.0)
OLMo 2 7B	Field–published	26.9	67.4	40.5 (38.1, 42.9)
OLMo 2 7B	Field–neutral	64.5	67.4	2.9 (1.3, 4.3)

The published-title rate was 448/1,664. Generated-title counts out of 4,992 outputs per condition were 2,058/3,017 for Qwen P0/P1, 3,842/4,015 for Mistral and 3,220/3,365 for OLMo. No output was excluded. Rates average the three repetitions within each abstract and give equal weight to all 416 strata. Confidence intervals use 2,000 paired bootstrap resamples within strata with the small-stratum variance rescaling described in Supplementary Methods. Supplementary Table S5 reports estimates based on complete responses.

Table S9 Comparison of rescaled-bootstrap and analytic uncertainty for extracted-title contrasts.

Model	Contrast	Difference (pp)	Rescaled-bootstrap 95% CI (pp)	Analytic 95% CI (pp)
Qwen 2.5 7B	Neutral–published	14.30	11.74, 16.66	11.77, 16.83
Qwen 2.5 7B	Field–published	33.51	30.97, 36.06	30.96, 36.07
Qwen 2.5 7B	Field–neutral	19.21	17.48, 21.04	17.39, 21.03
Mistral 7B v0.3	Neutral–published	50.04	47.59, 52.49	47.61, 52.47
Mistral 7B v0.3	Field–published	53.51	51.12, 55.98	51.10, 55.91
Mistral 7B v0.3	Field–neutral	3.47	2.22, 4.67	2.21, 4.72
OLMo 2 7B	Neutral–published	37.58	35.06, 39.99	35.12, 40.04
OLMo 2 7B	Field–published	40.48	38.15, 42.87	38.10, 42.87
OLMo 2 7B	Field–neutral	2.90	1.33, 4.32	1.40, 4.41

Both methods account for four inputs per year–field stratum. Bootstrap limits use 2,000 paired draws and the $\sqrt{4/3}$ rescaling of centred contrasts. Analytic limits use unbiased within-stratum sample variances and $\hat{\theta} \pm 1.96$ SE. All nine intervals remain above zero under both methods. The machine-readable uncertainty table also includes complete-response comparisons.

Table S10 Journal-composition decomposition through 2024 and at the provisional 2025 endpoint. Both comparisons use the same 17,076 journals with positive eligible agreement-cohort counts in every year of 2010–2025.

Measure	2022–2024	2022–2025
Eligible records at endpoint	2,440,077	2,540,102
Colon prevalence in 2022 (%)	26.77	26.77
Colon prevalence at endpoint (%)	30.27	34.77
Total change (pp)	+3.50	+8.01
Within-journal contribution (pp)	+3.82	+7.80
Publication-share contribution (pp)	−0.32	+0.21
Journals with ≥ 20 records at both endpoints	12,930	12,584
Of these, journals with an increase (%)	63.98	75.41
Median change in this subset (pp)	+2.66	+5.89
Journals with ≥ 100 records at both endpoints	3,895	3,824
Of these, journals with an increase (%)	73.09	86.82
Top 100 share of positive within-journal contributions (%)	29.13	27.46

The 2022 denominator is 2,310,085 in both comparisons. The within-journal and publication-share terms sum to the total change before rounding. This is an accounting decomposition, not a causal attribution. The panel still conditions on journal presence through 2025; using a 2024 outcome endpoint does not remove that selection.

S3 Supplementary figures

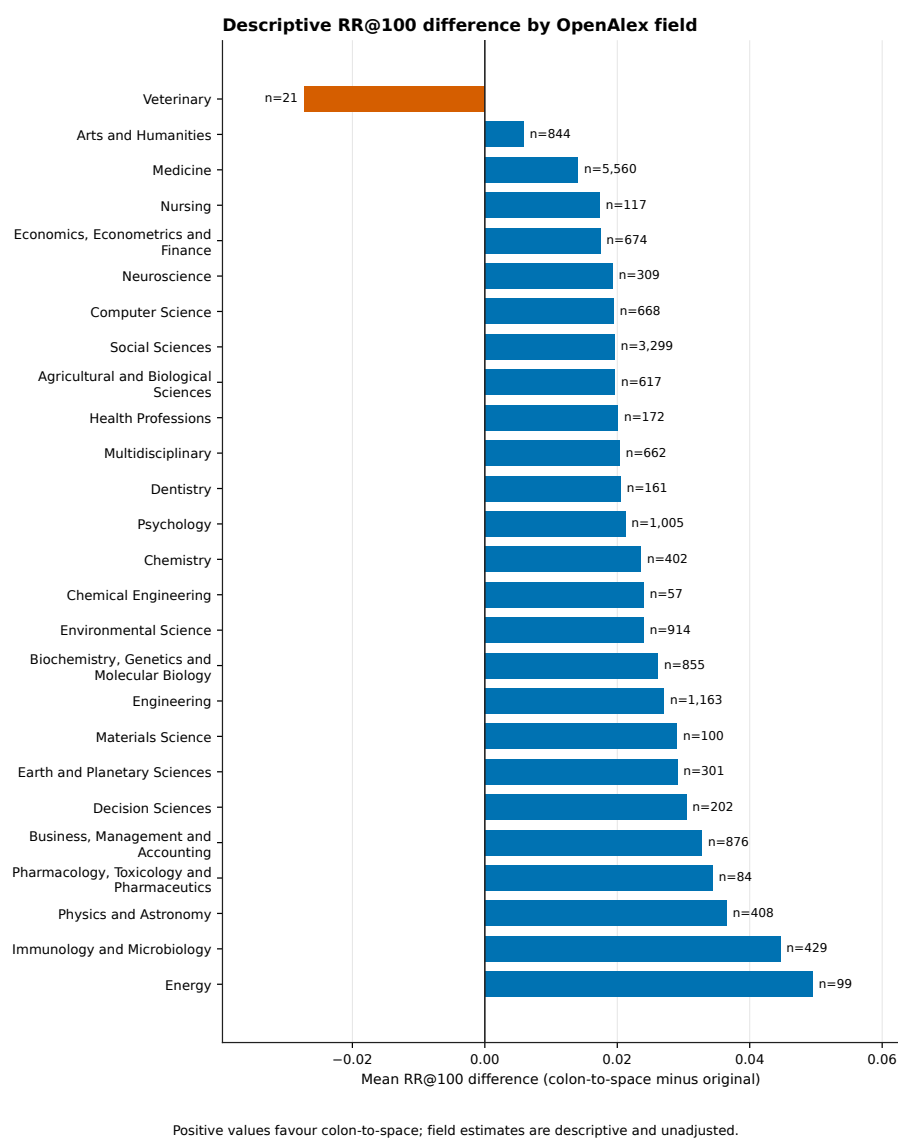
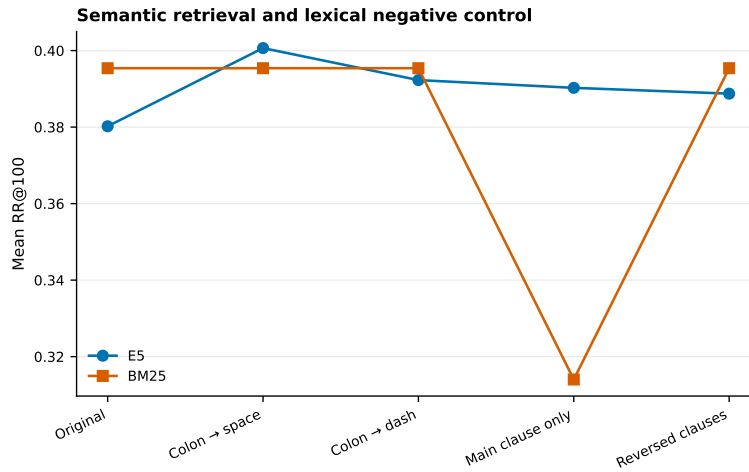
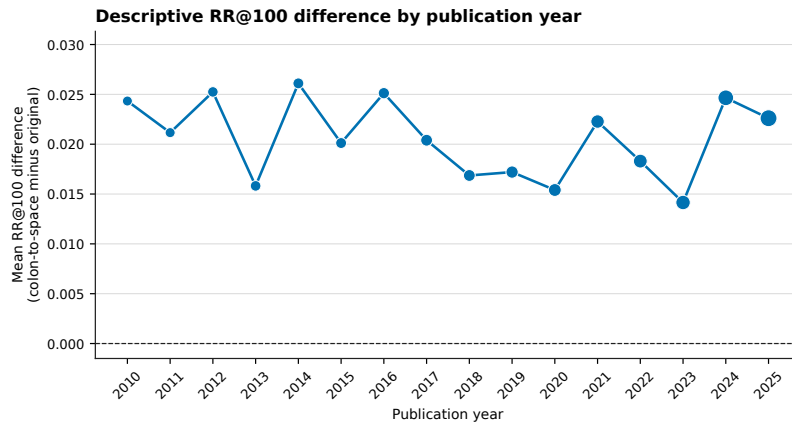


Fig. S1 Descriptive held-out RR@100 difference by OpenAlex field, calculated as colon-to-space minus original. Positive values favour colon-to-space. Field estimates are unadjusted, sample sizes are displayed beside the bars, and no confirmatory subgroup inference was specified.



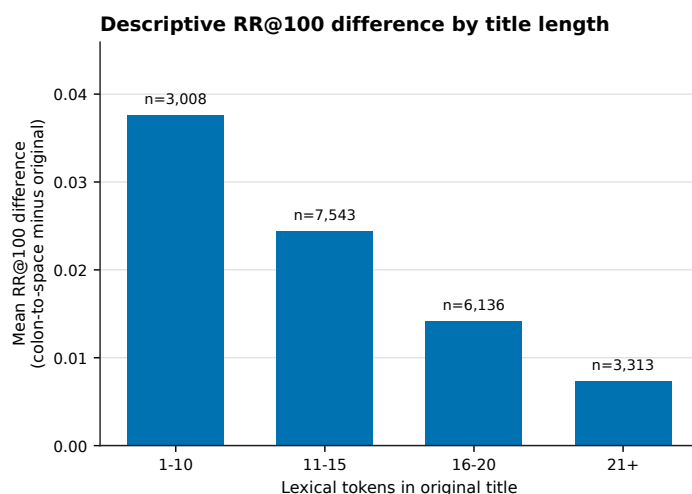
BM25 equality for punctuation-only variants follows the frozen analyzer; it is not a general claim about lexical search.

Fig. S2 Mean RR@100 across five frozen variants for semantic E5 retrieval and the BM25 negative implementation/intervention control. BM25 equality across the original, colon-to-space, colon-to-dash and reversed-clause variants follows from the frozen punctuation-separating, order-insensitive analyser; it is not a universal property of lexical search.



Positive values favour colon-to-space; point area reflects subgroup n.

Fig. S3 Descriptive held-out RR@100 difference by publication year, calculated as colon-to-space minus original. Positive values favour colon-to-space; point area reflects subgroup size. No confirmatory subgroup inference was specified. Results are conditional on the fixed E5 implementation.



Positive values favour colon-to-space; subgroup results are descriptive.

Fig. S4 Descriptive held-out RR@100 difference by lexical length of the original title, calculated as colon-to-space minus original. Positive values favour colon-to-space. No confirmatory subgroup inference was specified. Results are conditional on the fixed E5 implementation.

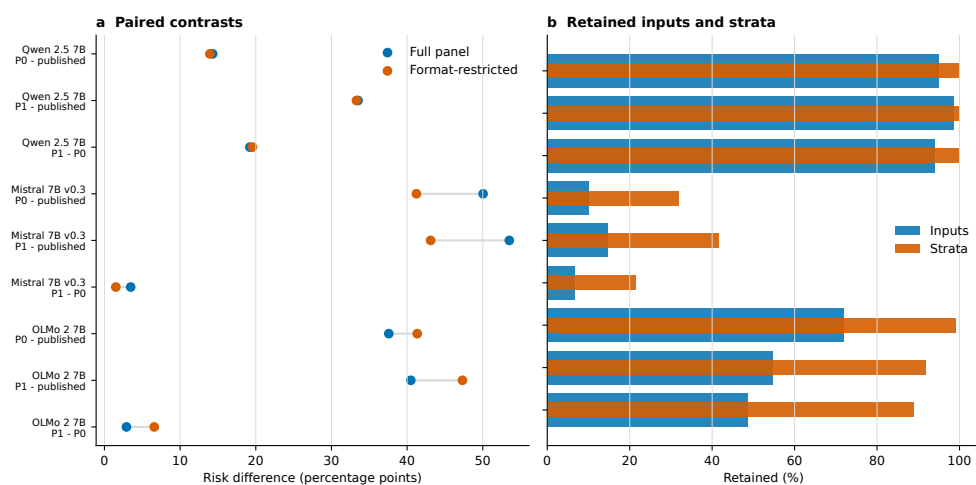


Fig. S5 Full-panel title-text and selected response-format-compliant contrasts. **a**, Full-panel versus per-protocol estimates. **b**, Fractions of inputs and year-by-field strata retained by the strict response-format criterion. Low coverage indicates selection and limits comparability.

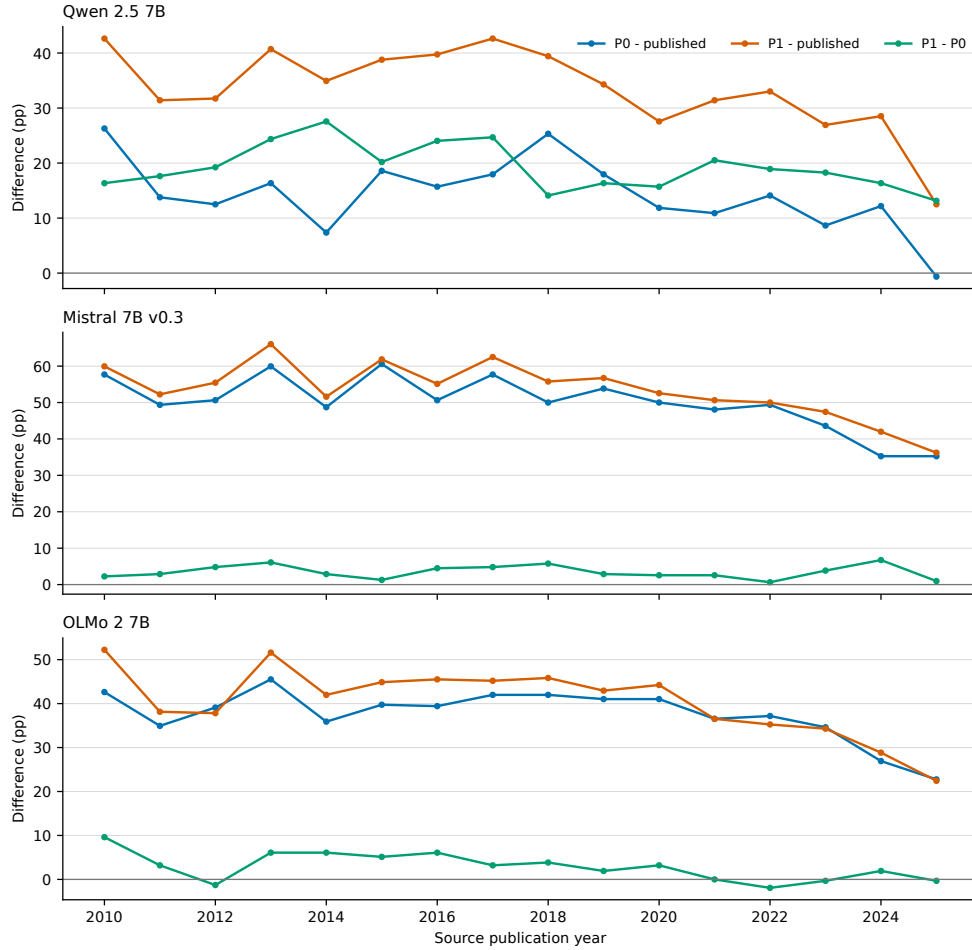


Fig. S6 Descriptive extracted-title colon contrasts by source publication year. All responses were generated in September 2026; the x-axis is not model execution time. No multiplicity-adjusted tests were performed.

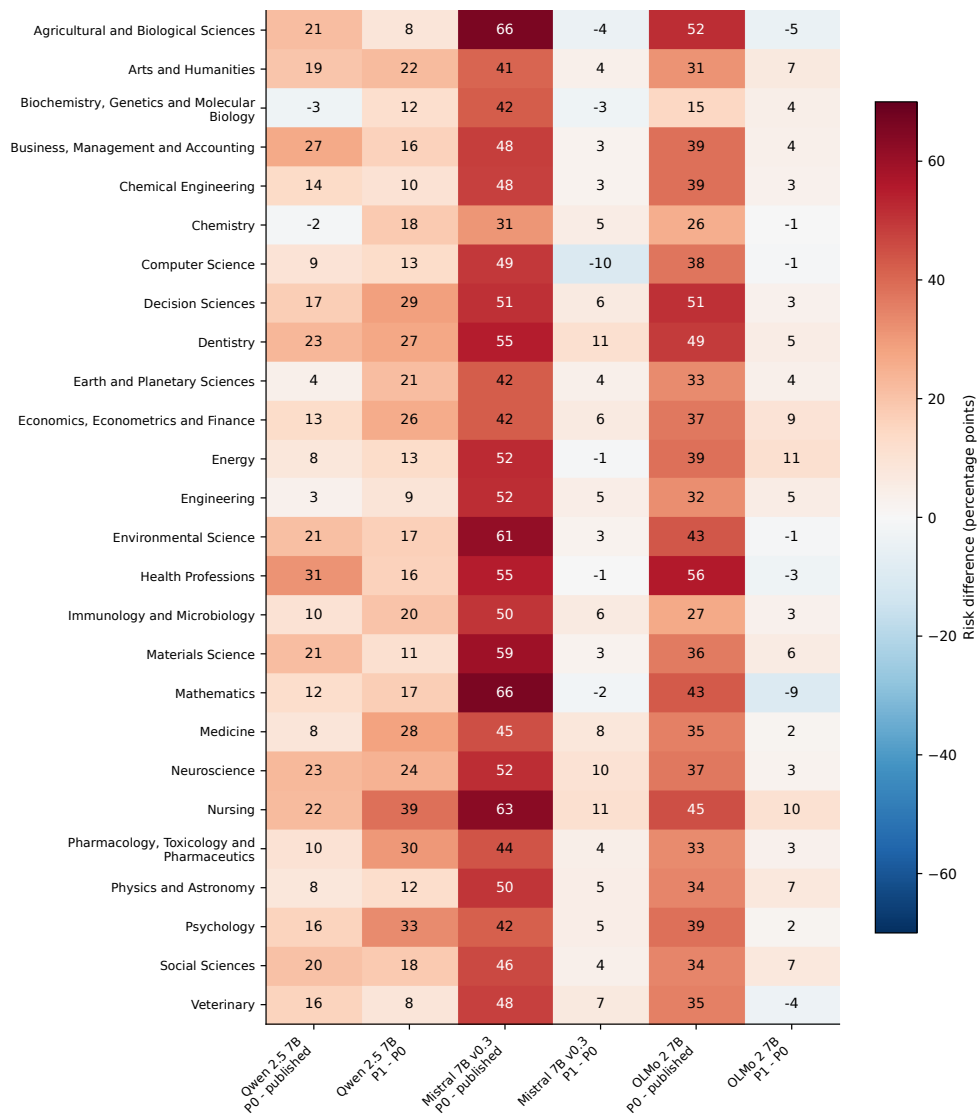


Fig. S7 Descriptive extracted-title colon risk differences across the 26 OpenAlex fields. Cell labels are percentage points; no multiplicity-adjusted field tests were performed.

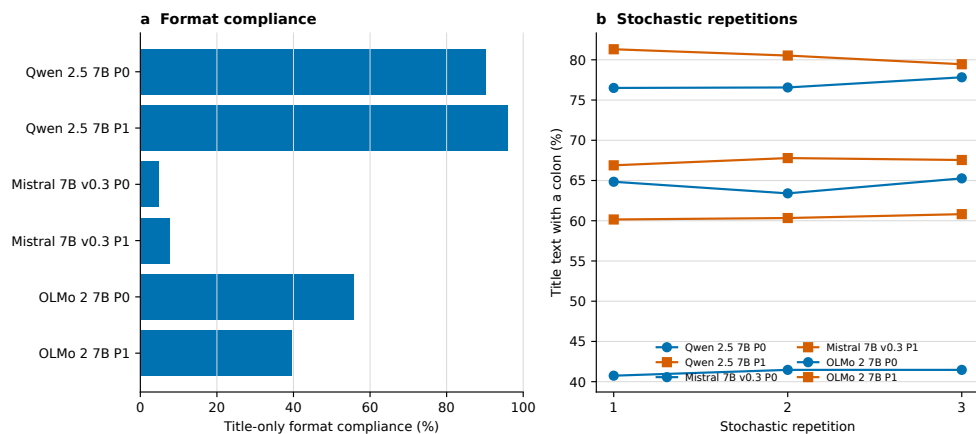


Fig. S8 Output quality and repeat stability in the generation experiment. **a**, Compliance with the archived strict-format criterion. **b**, Extracted-title colon rates in each stochastic repetition.

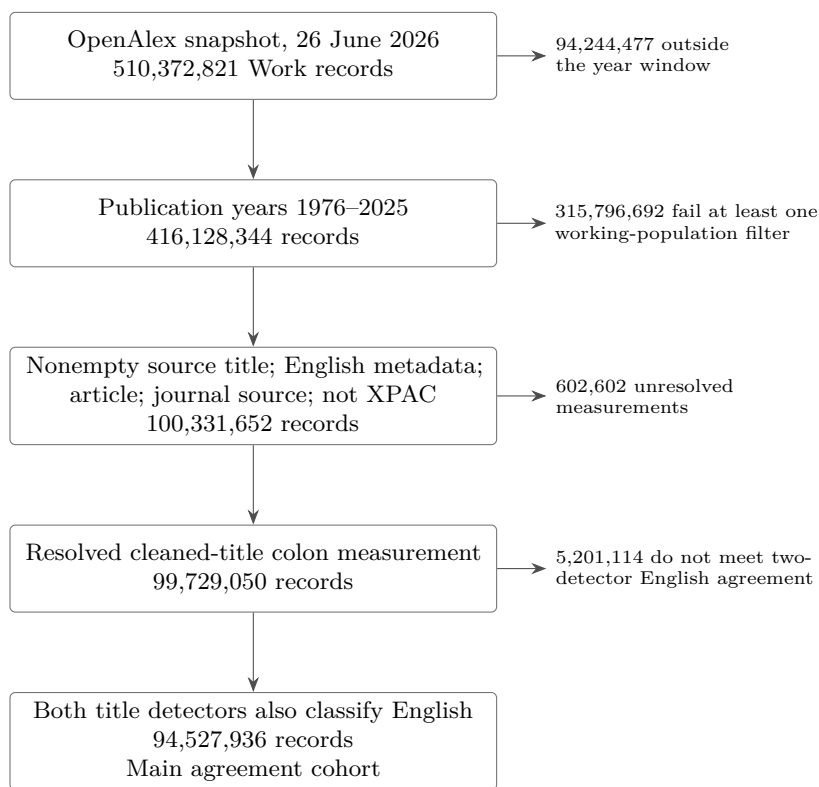


Fig. S9 Nested historical record selection. Exclusion counts are differences between successive recorded totals. The final cohort retains records without a DOI, abstract or field assignment; retractions were not excluded. Counts concern records, not independently deduplicated publications. Language agreement and resolved punctuation do not validate semantic title structure.

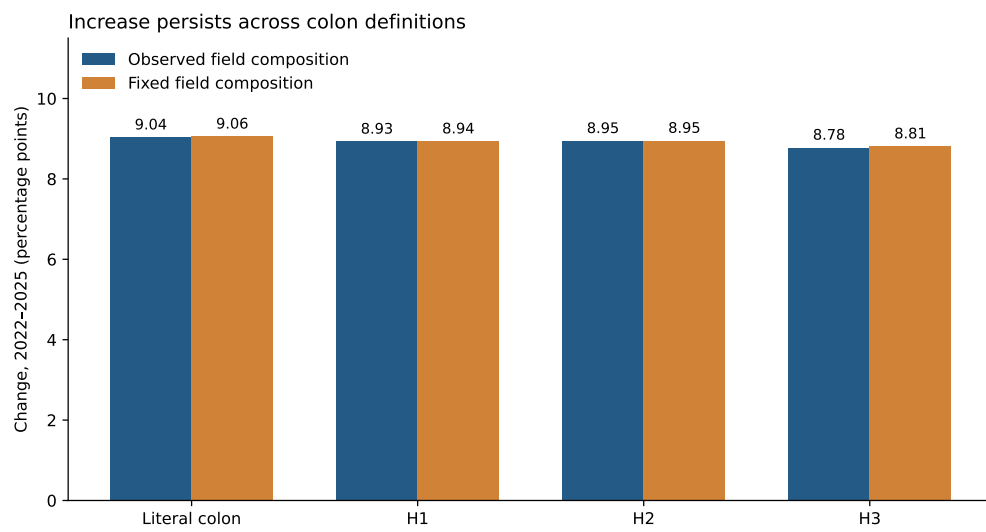


Fig. S10 Changes in colon prevalence from 2022 to 2025 under the literal and H1–H3 definitions, using observed and fixed 2018–2022 field composition. All estimates use the agreement cohort and common denominators; 2025 is provisional. The corresponding main-text table reports the comparison through 2024.