"Crucial research findings remain": Vocabulary change in PubMed abstracts from 2015 to 2026

Joost C. F. de Winter
Department of Cognitive Robotics, Delft University of Technology, Delft, The Netherlands
Email: j.c.f.dewinter@tudelft.nl
Date: Sep 26, 2026

Abstract
Words associated with AI-assisted writing may become less frequent as other expressions take their place. This study examined such changes in 13,437,416 PubMed abstracts published from January 2015 to August 2026. Monthly word frequencies were compared with pooled frequencies from 2018–2021, with adjustment for abstract length. Words were assigned to three groups according to the timing and shape of their changes after 2022. From each group, 30 words were selected one at a time; each step added the word that most increased the percentage of abstracts containing at least one selected word, relative to 2018–2021. Screening, grouping and selection were repeated in two-way cross-validation. The 'early set' rose and then fell towards its 2018–2021 level, the 'intermediate set' peaked later, and the 'latest set' continued to rise through August 2026. Common words such as "across", "remains" and "remain" characterized this latest phase. In August 2026, 91.77% of abstracts contained at least one of the 90 selected words, compared with 70.33% in 2018–2021. A separate cross-validated analysis estimated that at least 67.5% of abstracts from August 2026 were AI-assisted, assuming that a model fitted to 2018–2021 correctly predicts word use in unassisted abstracts. The results show that, as early markers such as "delves" declined, other words continued to rise.

Introduction

Large language models (LLMs) favor particular expressions and can thereby change the vocabulary of scientific writing. De Winter et al. (2023), for example, noted words such as "delves" and "crucial" in ChatGPT-generated text. Such observations have led to methods that estimate AI use from changes in word frequency. Kobak et al. (2025) compared observed word frequencies in biomedical abstracts with frequencies expected from earlier years. The difference between the two gave a lower bound on the share of LLM-assisted abstracts, and the ratio revealed increases in less common words. Other studies have examined changes in different scientific fields (Liang et al., 2025) and in the vocabulary, syntax and tone of civil and environmental engineering abstracts (Sanger & Maurer, 2026). Holzwarth et al. (2026) refined the lower bound and applied it to full-text biomedical publications.

The words associated with AI-assisted writing may change as models and writing practices change. Bitton et al. (2025) found that text generated by Claude, Gemini, Llama and OpenAI models could be distinguished by its style. The adoption of different models could therefore alter the expressions appearing in scientific papers. Authors may also change their prompts or edit conspicuous wording. Monthly word frequencies show whether an early increase persists, subsides or is followed by a rise in other expressions. Such patterns need to be interpreted alongside changes in research topics and writing conventions, which also affect scientific language (Bizzoni et al., 2020).

The present study examined how academic vocabulary changed through August 2026, with particular attention to words that rose late in the observation period. Monthly frequencies in PubMed abstracts from 2015 onward were used to group words whose frequencies changed in similar ways. Each group was represented by a set of 30 words, so that the groups could be compared using sets of equal size. Individual-word rankings described the largest absolute and proportional increases. To check whether selecting words exaggerated the changes, words were selected in one half of the database and evaluated in the other half. Finally, the selected words were used to estimate a lower bound on the share of AI-assisted abstracts.

Methods

Abstracts

The database was assembled from the 2026 PubMed baseline files and subsequent update files (National Library of Medicine [NLM], 2026). The collection was frozen on 19 September 2026. The latest record for each PubMed identifier (PMID) was retained, and records marked for deletion by NLM were removed. Records had to contain an English abstract and a publication month from January 2015 to August 2026.

Publication months were assigned with an electronic-first date rule, and records without a resolvable publication month were excluded.

A final check removed 80,199 records that were not eligible abstracts, such as errata, retraction notices and placeholders. The final database contained 13,437,416 abstracts.

Analysis

Words were extracted as lowercased alphabetic sequences after removal of markup, and different grammatical forms were counted separately. Structured-abstract labels were retained, so the counts included headings as well as prose. A separate check showed that the increase in “findings” persisted after these labels were removed.

The analysis vocabulary consisted of 22,580 word forms of at least three letters that occurred in at least 0.01% of abstracts, and in at least 300 weighted abstracts, in at least one year of a preliminary sample of PubMed records retrieved through Europe PMC (Rosonovski et al., 2024). Stopwords and publication metadata, such as “background” and “conclusion”, were excluded, whereas “findings” and “aims” were retained. The sample was used only to define this vocabulary; all reported frequencies were counted in the full database.

Frequency was defined as the percentage of abstracts containing a word at least once. The reference period was 2018–2021, with 4,448,927 abstracts. Frequencies were adjusted for abstract length, because a longer abstract has more opportunities to contain a word. They were first calculated within 50-word length bands, with a final band of 500 words or more, and then combined using the length distribution of the reference period. In other words, every month was given the same proportions of short and long abstracts. Excess frequency was the difference from the reference-period percentage, and the frequency ratio was the observed percentage divided by the reference-period percentage.

The next step identified words whose use increased after 2022. A word from the 22,580-form vocabulary passed the numerical screen if, in any month from January 2023 to August 2026, its length-standardized frequency exceeded its reference-period frequency by at least 0.1 percentage points or by a factor of at least 2. To prevent ratios from being determined by very small counts, the ratio criterion required at least 20 abstracts containing the word in the target month and 20 in the reference period.

Words passing the numerical screen were then checked against an eligibility list of 4,742 forms classified as general academic wording. This list came from a preliminary analysis of the same database, in which AI-assisted judgments classified 13,519 screened forms and excluded topic-specific and technical terms, proper names, abbreviations and metadata. The present analysis used this list unchanged. To audit these judgments, Opus 5.5 Extra classified a random sample of 600 candidate forms (300 included and 300 excluded) as general, borderline or topic-specific in two independent runs that did not see the saved decisions. The first run also classified all eligible forms that passed the numerical screen. With borderline forms counted as not general, agreement with the saved decisions was 86% and 83% (Cohen’s $\kappa = 0.71$ and 0.66), and agreement between the two runs was 95% ($\kappa = 0.89$).

The retained words were grouped according to when their frequencies rose and fell. Each word was represented by its monthly excess frequencies from January 2023 to August 2026, scaled to remove differences in size. K-means clustering assigned the words to three groups, which were labeled early, intermediate and latest according to the timing of their average series.

A set of 30 words was selected from each group. Using the same number of words made the sets comparable and kept the lists short enough to inspect. The coverage of a set was defined as the percentage of abstracts containing at least one of its words, and excess coverage as the difference from its coverage in the reference period. Selection began with an empty set and, at each step, added the word that produced the largest increase in mean excess coverage over the 44 months from January 2023 to August 2026. The procedure stopped at 30 words, and the resulting set was applied to every month. For the combined set of 90 words, an abstract was counted once if it contained any selected word.

Individual-word rankings were used to describe which expressions changed most within each group. Words were ranked by their largest monthly excess frequency and by their largest monthly frequency ratio during January 2023–August 2026. Ranking words by their peaks captured brief surges as well as increases that persisted. A word's largest excess frequency and largest frequency ratio could occur in different months. The count thresholds used in the numerical screen also applied to the rankings by frequency ratio.

Two-way cross-validation tested whether increases found in one half of the database also appeared in the other half. The database was split into two halves by PMID. Screening, grouping, word selection and reference-period length weights were determined in one half (the training half) and applied to the other (the held-out half), after which the roles were exchanged. The eligibility list was the only exception: it had been compiled from screens of both halves, so the held-out half contributed to which words were eligible. Figure 1, Tables 1 and 2, and the word lists came from a separate analysis of the full database. The 95% confidence intervals allowed for dependence among abstracts from the same journal.

Conditional lower bound

A separate analysis estimated the minimum share of AI-assisted abstracts needed to explain the observed coverage, P, assuming that a historical model correctly predicts the coverage in unassisted abstracts, Q. The conditional lower bound was $B = \max[0, (P - Q)/(1 - Q)]$, adapted from Holzwarth et al. (2026). The denominator, $1 - Q$, is the largest possible increase in coverage, which is reached if every assisted abstract contains a selected word. To illustrate, if $Q = 80\%$ and $P = 90\%$, at least $(90 - 80)/(100 - 80) = 50\%$ of the abstracts must be AI-assisted. Each abstract was evaluated using the words selected in the other half.

The historical model allowed word use to change gradually over time and to differ between calendar months. Monthly coverage during 2018–2021 was fitted using binomial logistic regression, with a linear time trend in log-odds and eleven calendar-month indicators. A separate model was fitted for each word set, held-out half and abstract-length band. These models supplied projected coverage, Q, for every study month, including months before and after the fitting period.

The bound was calculated separately for each half and length band, with negative values set to zero, and then averaged with weights proportional to the numbers of abstracts.

The 95% confidence intervals of the bound allowed for uncertainty in the observed coverage and the fitted historical model, and for dependence among abstracts from the same journal. They exclude uncertainty from word selection and errors in extrapolating the historical model.

Estimates before 2023 served as a check on the historical model. The linear trend was chosen after a model with constant coverage produced positive estimates in those years; this choice was not

preregistered. January–October 2022 provided a check outside the fitting period and ended before ChatGPT was released in November 2022.

Results

In the full database, 6,602 words passed the numerical screen, of which 3,201 were on the eligibility list. K-means assigned these words to three groups: 567 early, 827 intermediate and 1,807 latest words. Tables 1 and 2 show the ten largest monthly changes in each group, measured as excess frequency or as frequency ratio.

Table 1.

Words with the largest monthly absolute increases during January 2023–August 2026 relative to 2018–2021. Entries give the word, its excess frequency in percentage points, and the month of the largest increase. Frequencies are standardized for abstract length.

Rank	Early	Intermediate	Latest
1	research (6.91) Mar 2025	findings (17.18) Aug 2026	across (15.88) Aug 2026
2	additionally (6.34) Mar 2025	while (13.37) May 2026	remains (14.12) Aug 2026
3	crucial (5.72) Mar 2025	potential (9.86) Jul 2025	remain (9.84) Aug 2026
4	various (4.54) Nov 2024	significant (9.59) Apr 2025	limited (9.55) Aug 2026
5	impact (4.27) Dec 2024	including (8.73) May 2026	framework (9.46) Aug 2026
6	aims (3.29) Apr 2025	key (8.47) Jan 2026	associated (8.17) Aug 2026
7	understanding (2.94) Jan 2025	strategies (8.24) Mar 2026	support (8.14) Aug 2026
8	due (2.71) Jul 2025	particularly (8.14) May 2026	whereas (7.16) Aug 2026
9	explore (2.65) Apr 2025	using (7.58) Jun 2026	yet (6.87) Aug 2026
10	furthermore (2.43) May 2024	through (7.54) Aug 2026	based (6.59) Jun 2026

Table 2.

Words with the largest monthly frequency ratios during January 2023–August 2026. Entries give the word, ratio, and month of the highest observed ratio. Ratios compare length-standardized frequencies with the pooled 2018–2021 reference. Reference n is the number of reference-period abstracts containing the word.

Rank	Early	Intermediate	Latest
1	delves (108.31) Apr 2024, ref. n = 190	underscoring (31.67) Feb 2026, ref. n = 2,499	reframes (89.49) Aug 2026, ref. n = 50
2	delved (41.42) May 2024, ref. n = 104	underscores (22.69) Sep 2025, ref. n = 4,449	resamples (63.14) Aug 2026, ref. n = 99
3	showcasing (25.43) Aug 2024, ref. n = 495	underexplored (19.40) Nov 2025, ref. n = 2,240	reshapes (44.31) Aug 2026, ref. n = 197
4	delving (25.22) Apr 2024, ref. n = 133	excels (17.97) Mar 2025, ref. n = 147	auditable (38.65) Aug 2026, ref. n = 58

Rank	Early	Intermediate	Latest
5	delve (18.57) May 2024, ref. n = 557	outperforming (17.20) May 2026, ref. n = 2,249	synthesizes (27.80) Jun 2026, ref. n = 1,663
6	boasts (14.35) Sep 2024, ref. n = 122	surpassing (16.20) Jul 2025, ref. n = 1,200	synthesises (21.76) Aug 2026, ref. n = 240
7	meticulously (13.39) Nov 2024, ref. n = 638	underscore (15.50) Oct 2025, ref. n = 7,067	explainability (21.13) Jul 2026, ref. n = 274
8	commendable (12.64) Mar 2024, ref. n = 218	framework's (14.52) May 2026, ref. n = 209	contrastive (20.12) Aug 2026, ref. n = 336
9	comprehending (10.91) Jun 2024, ref. n = 593	excelled (13.88) Mar 2025, ref. n = 148	clearest (19.08) Aug 2026, ref. n = 209
10	intricacies (10.76) Apr 2024, ref. n = 634	transformative (12.66) Sep 2025, ref. n = 2,182	learnable (18.63) Aug 2026, ref. n = 200

Note. Reference counts pool all 48 months of 2018–2021 and are unweighted abstract counts. Peak ratios are subject to sampling uncertainty and selection optimism.

In the latest group, the largest absolute increases were for “across”, “remains” and “remain” (Table 1), and the largest frequency ratios for “reframes”, “resamples” and “reshapes” (Table 2). These words peaked in the last observed months, whereas the largest ratios in the early group, such as for “delves” and “delved”, occurred in 2024.

Because selection took account of words occurring together in the same abstract, the 30-word sets differ from the individual-word rankings in Tables 1 and 2. The words are listed below in selection order.

Early: research, crucial, additionally, impact, aims, utilizing, holds, utilized, impacts, intricate, exploration, constructed, employing, utilization, impacting, exploring, alleviate, utilizes, experiencing, prospects, implementing, avenue, predicting, showcasing, aiding, encompassed, benefiting, complexities, harnessing, delves.

Intermediate: findings, while, challenges, highlights, insights, enhance, highlighting, enhancing, notably, learning, offering, offers, achieving, underscores, comprehensive, underscoring, notable, enhances, leveraging, emphasizing, advancements, managing, sustainable, mitigate, explores, incorporating, underexplored, necessitating, mitigating, introduces.

Latest: across, remains, remain, performance, framework, integrating, enabling, provides, alongside, rare, integration, supporting, driven, broader, substantial, thematic, achieves, preserving, gaps, favorable, integrates, narrative, enables, equitable, pronounced, equity, disrupts, proposes, prioritize, alleviated.

Figure 1 shows the coverage of each set over time. By August 2026, the early set had returned close to its reference-period coverage, whereas the intermediate and latest sets covered many more abstracts. From the reference period to August 2026, length-standardized coverage changed from 32.82% to 34.38% for the early set, from 39.43% to 70.59% for the intermediate set, and from 36.20% to 73.79% for the latest set.

The three sets peaked at different times. The early set peaked in March 2025, at 52.20%, and then declined. The intermediate set peaked in May 2026, and the latest set was still rising in August 2026, the last observed month.

Coverage of the combined set of 90 words, selected from the full database, increased from 70.33% in the reference period to 91.77% in August 2026, a difference of 21.44 percentage points (95% CI 21.00–21.89). In other words, fewer than one in ten abstracts from August 2026 contained none of the 90

words. Some increase preceded 2023: in 2022, coverage of the combined set exceeded its reference-period level by an average of 3.79 percentage points (95% CI 3.54–4.05).

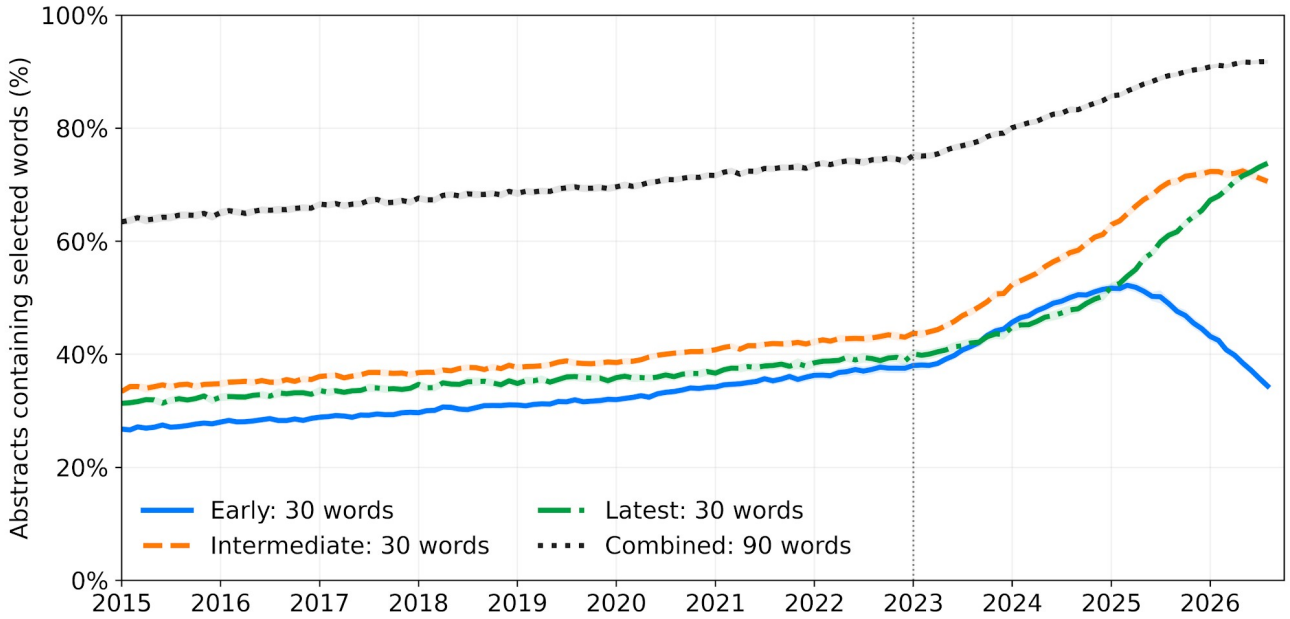

Figure 1. Monthly percentages of abstracts containing at least one word from each fixed 30-word set or the combined 90-word set (N = 13,437,416). Each abstract counts once per set. Word screening and selection used the full eligible database, and percentages are standardized to the pooled 2018–2021 length distribution. Shading shows conditional 95% journal-cluster intervals. The vertical line marks January 2023. Cross-validated results are reported in the text.

Cross-validation showed closely matched increases in the training and held-out halves. For the combined set, mean monthly excess coverage from January 2023 to August 2026 was 13.58 percentage points in the training halves and 13.56 in the held-out halves; in August 2026, the values were 20.97 and 20.99. These values differ slightly from the 21.44 points reported above, which were based on words selected in the full database. Of the 30 words per set, 24 (early), 26 (intermediate) and 26 (latest) were selected in both training halves. All 120 top-ranked words from the training halves (ten per group, ranking type and half) also increased at their peak month in the held-out half.

Figure 2 compares observed coverage with the coverage projected by the historical model. Before 2023, the two differed by at most 1.81 percentage points in any month. From 2023 onward, observed coverage of the intermediate, latest and combined sets rose well above the projection. The early set fell below its projection from February 2026 onward and was 10.23 percentage points below it in August 2026 (36.44% versus 46.67%).

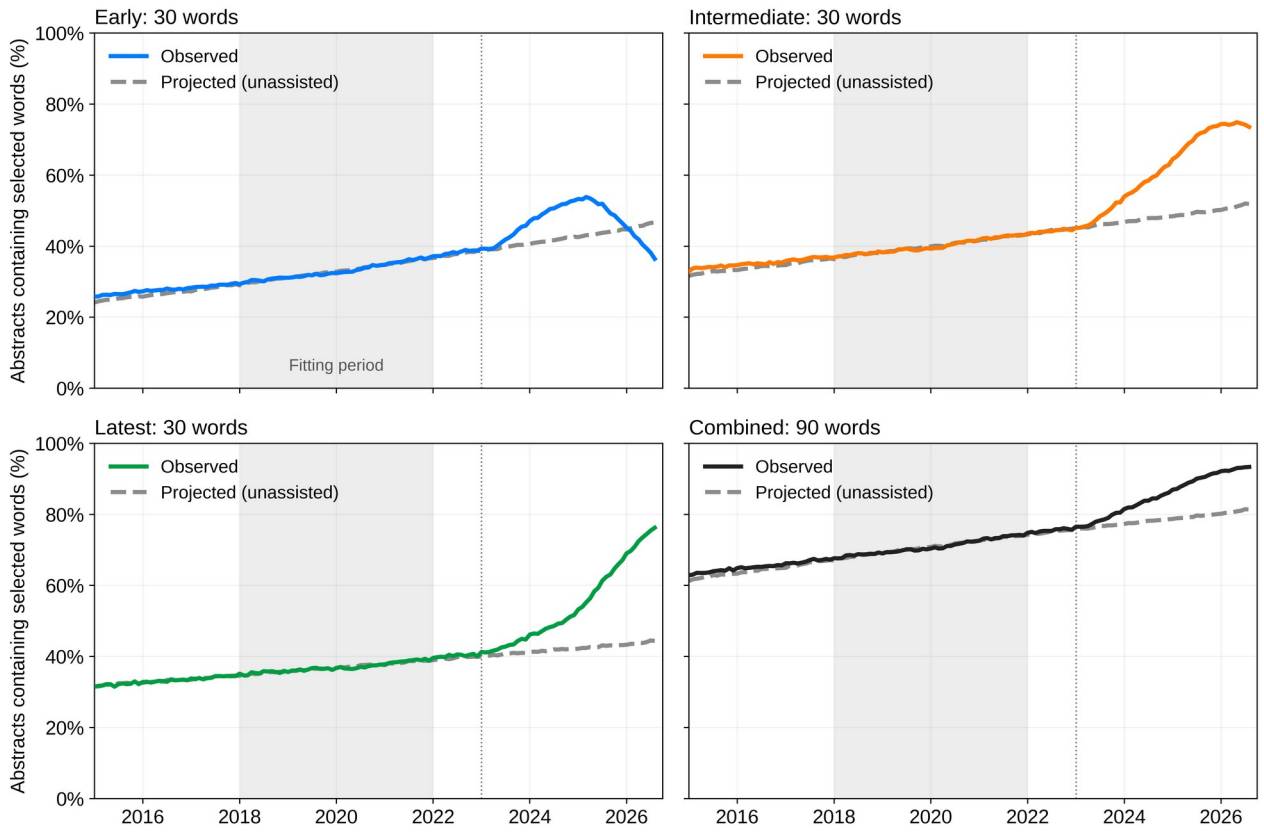

Figure 2. Observed and projected monthly percentages of abstracts containing at least one word from each 30-word set or the combined 90-word set. Solid lines show observed percentages, and dashed grey lines show the percentages projected for unassisted writing by models fitted to 2018–2021 (grey band). Each abstract was evaluated using words selected in the other half, and percentages use the actual mix of abstract lengths in each month (Figure 1 used words selected in the full database and a standardized length mix). The vertical line marks January 2023. The lower bounds in Figure 3 follow from the gap between the two lines.

Figure 3 shows the conditional lower bounds on the share of AI-assisted abstracts. For August 2026 (108,840 abstracts), the bounds were 0.05%, 45.97% and 57.98% for the early, intermediate and latest sets, and 67.52% for the combined set (conditional 95% CI 66.02–69.01%). The bound for the combined set was high because few abstracts contained none of the 90 words: 6.63% in August 2026, compared with a projected 18.59% under unassisted writing (both values pooled across length bands). The early set's coverage was below its projected coverage, so its bound was close to zero. Before 2023, the combined bound averaged 2.29% in January–October 2022 and reached its highest monthly value, 4.94%, in November 2015. These values reflect departures from the projected coverage and the effect of setting negative values to zero. Each word set gives a lower bound on the same share of AI-assisted abstracts, so the set-specific bounds cannot be added.

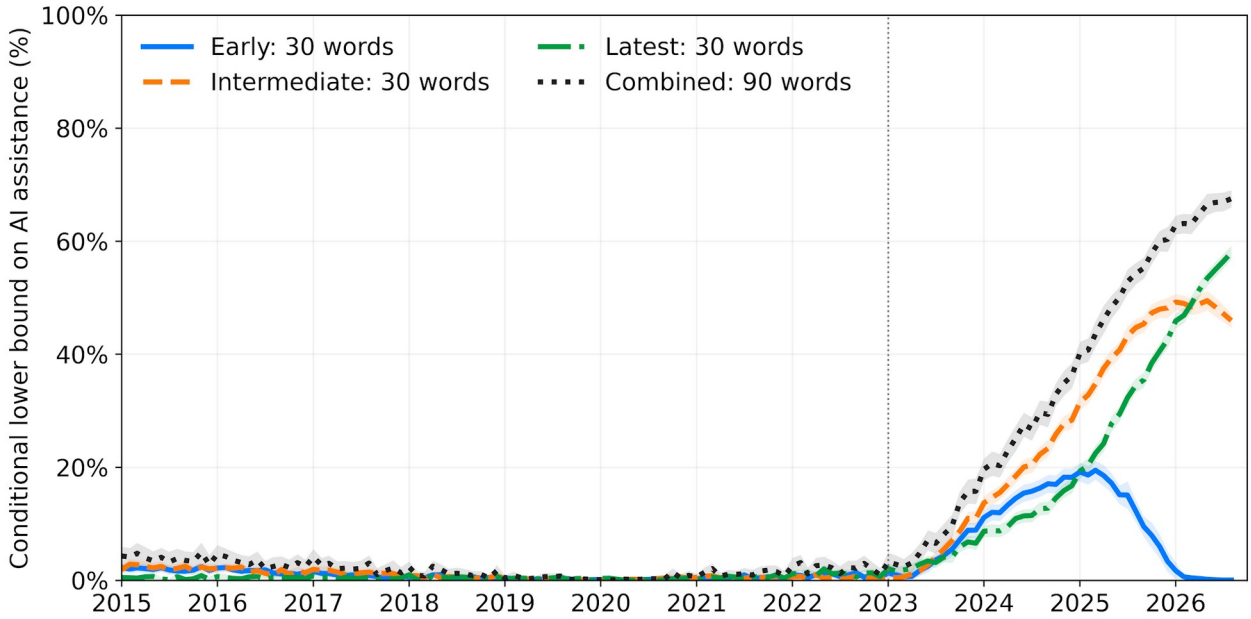

Figure 3. Monthly conditional lower bounds on AI-assisted abstracts. The early, intermediate and latest 30-word sets are shown by blue solid, orange dashed and green dash-dot lines, respectively. The black dotted line shows the combined 90-word set. Projected coverage under unassisted writing comes from models fitted to 2018–2021, with a time trend and calendar-month effects. Models were fitted in each held-out half and length band for words selected in the other half. Shading shows conditional pointwise 95% journal-cluster intervals. The vertical line marks January 2023. Values before 2023 serve as a check on the historical model. Interpretation as a lower bound on AI use requires projected coverage to represent unassisted writing.

Discussion

This study examined how the vocabulary of PubMed abstracts changed through August 2026. Coverage of the early set rose and then declined, the intermediate set peaked later, and the latest set continued to rise through August 2026. Common words such as “across”, “remains” and “remain” characterized this latest phase. Similar increases appeared when words were selected in one half of the database and evaluated in the other. The main result is the changing composition of the vocabulary: as earlier markers became less frequent, other expressions continued to rise.

Several changes in writing practices could explain this pattern. Model families differ in their characteristic wording (Bitton et al., 2025), so the adoption and replacement of models could change the expressions appearing in scientific abstracts. Authors may also revise their prompts, edit generated text or adopt expressions from other writers. Determining the contribution of these mechanisms would require information about how the abstracts were written and which models were used.

Limitations

Several methodological choices affect how the vocabulary changes are described. The reference period determines the comparison level, the eligibility list determines which words enter the analysis, and the choice of three groups determines how their histories are summarized. Pooling several reference years stabilizes the comparison level, although trends and seasonal variation can still affect the differences from it. Clearly topic-specific words, such as “covid”, “semaglutide” and “chatgpt”, were absent from the eligibility list, but the list was less consistent for method and computing terms. In the audit, 6% of the 3,201 words used in the full-database analysis were classified as topic-specific or technical and 14% as borderline; none of the 90 selected words was classified as topic-specific or technical. In Table 2, four of the ten words in the latest group were classified as technical (“resamples”, “explainability”, “contrastive” and “learnable”), so this ranking may partly reflect the growth of machine-learning research.

Cross-validation checked whether selecting words exaggerated their increases, but both halves covered the same periods, journals and topics, and shared the same eligibility list. Their agreement therefore shows replication within one database; prediction of future months is a separate question. The individual-word peaks may likewise overstate the changes, because the peak months were selected for having the largest differences.

Interpreting the lower bound as a share of AI-assisted abstracts depends on the historical model's ability to predict unassisted writing. Changes in research topics, journal composition or human writing conventions can alter coverage independently of AI assistance. The positive estimates before 2023 show how departures from projected coverage can contribute to the bound. The bound is especially sensitive to errors in Q when Q is high, because the denominator 1 − Q is then small. To illustrate, if unassisted coverage of the combined set in August 2026 had been 88% instead of the projected 81.41%, the bound would have been about 45%. The estimates are nonetheless in line with those of Holzwarth et al. (2026) for PubMed Central abstracts: 53% for 2025 and 68% for December 2025, compared with 51% and 60% here. The estimates describe the database as a whole. Assessing AI use in individual papers requires separate validation for the relevant models and editing conditions (Alshammari & Rao, 2025).

Finally, recent months are subject to indexing delays, and the length adjustment accounts only for abstract length, so changes in topics and journals can still affect the results.

Conclusions

The vocabulary associated with AI-assisted writing in PubMed abstracts changed between 2023 and 2026. Early markers such as "delves" rose and then declined, whereas common words such as "across", "remains" and "remain" continued to rise through August 2026. In August 2026, 91.77% of abstracts contained at least one of the 90 selected words and, assuming that the historical model predicts unassisted writing, at least 67.5% of abstracts were AI-assisted. The early set alone gave a bound close to zero, so a fixed list of early markers would miss the AI assistance that later words reveal. Markers of AI-assisted writing therefore need to be updated over time.

Data and code availability

The code, data and results are available in the code and feature-data archives. PubMed data: Courtesy of the U.S. National Library of Medicine. The released data reflect PubMed as of 19 September 2026 and do not reflect the most current data available from NLM.

Declaration of AI use

AI assistance was used for word classification, analysis programming and manuscript preparation. The analysis was done with GPT-6-Astra, with further support from Opus 5.5 Extra.

References

Alshammari, H., & Rao, P. (2025). *Evaluating the performance of AI text detectors, few-shot and chain-of-thought prompting using DeepSeek generated text*. arXiv. https://doi.org/10.48550/arXiv.2507.17944

Bitton, Y., Bitton, E., & Nisan, S. (2025). *Detecting stylistic fingerprints of large language models*. arXiv. https://doi.org/10.48550/arXiv.2503.01659

Bizzoni, Y., Degaetano-Ortlieb, S., Fankhauser, P., & Teich, E. (2020). Linguistic variation and change in 250 years of English scientific writing: A data-driven approach. *Frontiers in Artificial Intelligence, 3*, Article 73. https://doi.org/10.3389/frai.2020.00073

de Winter, J. C. F., Dodou, D., & Stienen, A. H. A. (2023). ChatGPT in education: Empowering educators through methods for recognition and assessment. *Informatics, 10*(4), Article 87. https://doi.org/10.3390/informatics10040087

Holzwarth, L., González-Márquez, R., & Kobak, D. (2026). *Most biomedical publications show signs of LLM-assisted writing*. arXiv. https://doi.org/10.48550/arXiv.2608.10715

Kobak, D., González-Márquez, R., Horvát, E.-Á., & Lause, J. (2025). Delving into LLM-assisted writing in biomedical publications through excess vocabulary. *Science Advances, 11*(27), Article eadt3813. https://doi.org/10.1126/sciadv.adt3813

Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W., Zhao, X., Cao, H., Liu, S., He, S., Huang, Z., Yang, D., Potts, C., Manning, C. D., & Zou, J. (2025). Quantifying large language model usage in scientific papers. *Nature Human Behaviour, 9*(12), 2599–2609. https://doi.org/10.1038/s41562-025-02273-8

National Library of Medicine. (2026). *PubMed 2026 baseline and update files* [Data set]. Retrieved September 19, 2026, from https://pubmed.ncbi.nlm.nih.gov/download/

Rosonovski, S., Levchenko, M., Bhatnagar, R., Chandrasekaran, U., Faulk, L., Hassan, I., Jeffryes, M., Mubashar, S. I., Nassar, M., Jayaprabha Palanisamy, M., Parkin, M., Poluru, J., Rogers, F., Saha, S., Selim, M., Shafique, Z., Ide-Smith, M., Stephenson, D., Tirunagari, S., . . . Harrison, M. (2024). Europe PMC in 2023. *Nucleic Acids Research, 52*(D1), D1668–D1676. https://doi.org/10.1093/nar/gkad1085

Sanger, M. D., & Maurer, B. W. (2026). *Have large language models enhanced the way civil & environmental engineers write? A quantitative analysis of scholarly communication over 25 years*. arXiv. https://doi.org/10.48550/arXiv.2602.03864