Has Scientific Writing Changed in the LLM Era? A Descriptive KLD Analysis of Language Change in Psychology Abstracts, 2020–2025

Chengze Chen
The University of Melbourne
chengzechen@student.unimelb.edu.au

Abstract

The rapid adoption of large language models (LLMs) has raised questions about whether scientific writing is changing in visible ways. The research examined this question using English OpenAlex psychology abstracts from 2020 to 2025. Kullback-Leibler divergence (KLD) was used to compare yearly language distributions at two levels: a primary unigram layer and a secondary part-of-speech (POS) trigram layer. The analysis included adjacent-year KLD comparisons and a tentative pre/post comparison between 2020–2022 and 2023–2025. The analysed corpus contained 112,689 abstracts. In the pre/post comparison, unigram divergence was detectable (KLD = 0.0256), but the largest adjacent-year unigram shift occurred between 2024 and 2025 (KLD = 0.0257) instead of between 2022 and 2023. The feature-level contributor plots were filtered using unpaired Welch's t-tests at $p < 0.05$. Top filtered unigram contributors include topic terms, function words, and possible academic style words, including `ai`, `findings`, `enhance`, and `insights`. POS trigram divergence was smaller, with a pre/post KLD of 0.0092, but it also increased in later adjacent-year comparisons. The findings suggest that there are measurable distributional signals during the period of LLM adoption, but they do not show that LLMs caused the observed changes.

Keywords: scientific writing; large language models; OpenAlex; psychology abstracts; Kullback-Leibler divergence; POS trigrams

1 Introduction

Scientific writing is one of the main ways scientific knowledge is circulated and evaluated. It is also a form of language that evolves rapidly. Sun et al. [1] described scientific writing as part of human culture, which has "evolved over centuries" to become its present form. Biber and Gray [2] hold a similar view, that academic writing has undergone major historical changes, specifically in terms of grammatical style. Recent computational research treats academic English as a measurable diachronic object by applying large corpora and information-theoretic methods [3]. Therefore, abstracts are useful units for analysing how scientific communities communicate.

The recent proliferation of large language models may have introduced new variables into this process. A Nature survey indicates that AI tools are becoming increasingly prevalent in science, and many researchers expect them to play a significant role in their respective fields [4]. Since the widespread adoption of ChatGPT in late 2022, several studies have attempted to examine whether LLM-assisted writing has become visible in academic contexts. One piece of evidence comes from lexical overrepresentation: Gray [5] estimated that at least 60,000 papers used LLM-assisted writing in 2023, with several keywords appearing unusually frequently in LLM-generated

text; Juzek and Ward [6] identified keywords such as "delve," "intricate," and "underscore" whose increased frequency in scientific abstracts may be related to LLM use. Another piece of evidence comes from larger-scale corpus studies. Liang et al. [7] reported a steady increase in LLM-modified content in scientific papers; Kobak et al. [8] reached similar conclusions in biomedical abstracts. However, this evidence requires careful interpretation. Beyond isolated words, Galpin et al. [9] suggest that recent changes may involve broader semantic and pragmatic patterns.

This study uses Kullback-Leibler divergence (KLD), since the research question is distributional. Kullback and Leibler [10] originally proposed this metric to describe distances or differences between statistical populations. In corpus studies, this concept can be applied by comparing the probability distributions of linguistic features across different periods. Degaetano-Ortlieb and Teich [11] used relative entropy to compare temporally adjacent periods in scientific English and identify features that led to changes. This is useful for this project because it does not begin with a manually selected list of AI-associated words. Instead, KLD analysis allows us to explore whether the overall feature distribution has changed and then examine which specific features contributed most to this change. Corpus comparison work also supports this feature-level perspective, as word-level visualisation helps examine changes across a broader corpus [12].

Therefore, the current analysis uses a unigram layer and a secondary POS trigram layer. Unigrams are individual word forms, so they provide a direct method for detecting lexical, topical, and stylistic changes. This is important because recent evidence related to LLMs largely focuses on increases in vocabulary. In contrast, POS trigrams provide a grammatical approximation by representing sequences of three part-of-speech tags. Degaetano-Ortlieb and Teich [11] used lemmas or unigrams for the lexical level and POS trigrams for the grammatical level. However, this does not mean that POS trigrams can fully capture syntactic structures. Sun et al. [1] explicitly state that POS trigrams cannot represent all grammatical structures. Therefore, in this study, POS trigrams are treated only as a cautious check of broader structural patterns.

Existing research provides strong impetus for examining LLM-era scientific writing; however, it remains unclear whether similar distributional features exist in areas such as psychology, and whether these features manifest primarily at the lexical level or also in coarse-grained syntactic patterns. This project utilises psychology abstracts from OpenAlex spanning 2020 to 2025 to conduct a small-scale, reproducible pilot study. The study design examines short-term distributional changes by comparing adjacent years and contrasts two phases: 2020–2022 and 2023–2025, representing the periods before and after the widespread adoption of LLMs. This division was chosen because ChatGPT began to be widely adopted after the end of 2022; however, this does not constitute a causal relationship. Other factors, including shifts in research focus, the composition of the database, and trends within specific fields, may also affect word and POS-trigram distributions.

This study aims to investigate whether OpenAlex psychology abstracts exhibit measurable changes in distribution during the widespread adoption of large language models (LLMs). Using English abstracts from 2020 to 2025, this project compares the Kullback-Leibler divergence (KLD) of adjacent years, as well as the tentative pre/post LLM-adoption split of 2020–2022 versus 2023–2025. The primary hypothesis is that, from 2023 onwards, there will be detectable differences in unigram distributions, with certain high-contribution features overlapping with the vocabulary or style discussed in recent literature on LLM writing. A secondary hypothesis is that POS trigram distributions may also change; however, such signals require more cautious interpretation and may not be as pronounced or direct as the differences observed in unigrams.

2 Methods

2.1 Research Design

The study employs a retrospective corpus design to examine whether English OpenAlex psychology abstracts exhibit measurable changes in linguistic distribution during the widespread adoption of LLMs. The analysis compares annual samples of abstracts from 2020 to 2025, as well as aggregated samples from the periods before and after the adoption of LLMs: 2020–2022 and 2023–2025. The design is descriptive rather than causal. As information regarding whether authors used LLMs is unavailable, the study aims to detect temporal distributional signals rather than to identify abstracts written with LLM assistance.

The primary layer used unigrams, since individual word forms are the most direct way to observe changes in vocabulary, topic, and style. The secondary layer used POS trigrams, which serve as a conservative syntactic check, since three-tag sequences can provide a compact view of repeated grammatical patterns.

2.2 Corpus and Sampling

The corpus was constructed using records from the OpenAlex database classified under the psychology field (`primary_topic.field.id = 32`). For each publication year from 2020 to 2026, up to 20,000 records of English-language articles with abstracts were initially collected using a fixed random seed. This represents only an upper sampling target, as data cleaning removed some records. This analysis reused the cleaned dataset and applied the analytical filters for 2020–2025 after cleaning, ultimately analysing 112,689 abstracts.

The 2026 sample has been excluded because this draft was written in June 2026, and the 2026 records represent only the first half of the publication year. This means that the results reported are current as of 2025 and should not be interpreted as describing the full 2026 publication year.

2.3 Text Cleaning

Abstracts were reconstructed from OpenAlex abstract text and processed using a cleaning workflow. This workflow was designed to remove obvious noise whilst retaining accessible scientific language. The workflow removed duplicate records, empty or extremely short abstracts, HTML tags and entities, URLs, and records flagged as suspicious non-abstract content.

For unigram analysis, the text was converted to lowercase and tokenised. A small amount of technical noise (such as HTML remnants) was excluded. We used a minimum abstract length threshold of 40 tokens to filter out excessively short abstracts. These rules are intended to improve the readability and reproducibility of the corpus.

2.4 Feature Extraction

We calculated unigram counts for each year of analysis. The final unigram vocabulary contained 10,008 features after filtering to features with a minimum total corpus count of 100. The unigram layer was expected to capture shifts in research topics, common academic terminology, and possible LLM-associated stylistic changes.

POS trigrams were extracted as sequences of three part-of-speech tags. The analysis utilised spaCy 3.8.11 and the `en_core_web_sm` POS model, and punctuation, special characters, spaces, and unknown tags were excluded when constructing trigrams. The final POS trigram vocabulary contained 1,731 features after filtering to POS trigrams with a minimum total corpus count of 100.

This layer was designed to be as simple as possible. Its purpose is to demonstrate whether broad grammatical patterns vary with changes in lexical distribution, not to support strong claims about syntax.

2.5 Kullback-Leibler Divergence Analysis

KLD was used to compare feature probability distributions across time. Drawing on the relative entropy method employed by Degaetano-Ortlieb and Teich [11], the method treats each time period as a probability distribution over a set of shared linguistic features and measures the degree of divergence between the later distribution and the earlier one. For two distributions P and Q over the same feature set F, KLD was defined as:

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{f \in F} P(f) \log_2 \left(\frac{P(f)}{Q(f)} \right) \tag{1}$$

In this formula, P is the target distribution to be encoded, and Q is the reference distribution. The logarithm was base 2, so this value can be interpreted as bits. KLD is asymmetric, meaning that $D_{\mathrm{KL}}(P \parallel Q)$ is not necessarily the same as $D_{\mathrm{KL}}(Q \parallel P)$. This study primarily reports on the direction from later to earlier periods, as the research question concerns whether the distribution of psychology abstracts from the later period differs from that of the earlier abstracts.

For each year or aggregated period, we calculated feature probabilities based on the retained vocabulary:

$$p_t(f) = \frac{\mathrm{count}_t(f)}{\sum_{g \in F} \mathrm{count}_t(g)} \tag{2}$$

This normalisation was performed separately for the unigram feature set and the POS trigram feature set. Since rare or absent features can cause zero-frequency issues in KLD, the analysis employed background corpus smoothing following the relative entropy method used by Degaetano-Ortlieb and Teich [11]. Let $B(f)$ denote the background probability of feature f in the full retained corpus. The smoothed probability for period t was:

$$p_t^*(f) = \lambda p_t(f) + (1 - \lambda) B(f) \tag{3}$$

The smoothing parameter was $\lambda = 0.995$, so the observed period distribution carried most of the weight, whilst the full-corpus background distribution supplied a small amount of probability mass. This setting was intended to preserve the observed yearly or pooled distributions while reducing instability caused by zero-frequency features. These smoothed probabilities are used in the KLD calculation.

Two experimental variants were subsequently computed. First, for adjacent-year change, each later year was compared with the immediately preceding year:

$$D_{\mathrm{adjacent}}(t) = D_{\mathrm{KL}} \left(P_t^* \parallel P_{t-1}^* \right), \quad t \in \{2021, \ldots, 2025\} \tag{4}$$

Second, for the main pre/post comparison, the 2020–2022 counts were aggregated into a tentative pre-LLM-adoption distribution and the 2023–2025 counts were aggregated into a tentative post-LLM-adoption distribution:

$$D_{\mathrm{prepost}} = D_{\mathrm{KL}} \left(P_{\text{2023-2025}}^* \parallel P_{\text{2020-2022}}^* \right) \tag{5}$$

The reverse direction was also saved by the pipeline for checking, but the Results section focuses on the later-versus-earlier direction. This follows the logic of asking which features became more distinctive in the later period relative to the earlier period.

This is a simplified adaptation of the research design by Degaetano-Ortlieb and Teich [11]. Their study employed sliding windows and different time periods to detect periods of diachronic change within a long-term historical corpus. This study, however, employs fixed year-on-year comparisons and sets the demarcation between pre- and post-LLM adoption based on theoretical grounds, as the analysed corpus covers the years 2020–2025 and the primary research question concerns recent changes in the LLM era.

For each comparison, the total KLD was decomposed into feature-level contribution terms:

$$d_f(P^* \parallel Q^*) = P^*(f) \log_2 \left(\frac{P^*(f)}{Q^*(f)} \right) \tag{6}$$

The sum of these feature-level terms equals the total KLD. Individual contribution terms can be positive or negative, but the top positive terms in the later-versus-earlier direction identify features that are most distinctive of the later distribution. The pipeline therefore saved both total KLD values and feature-level contribution tables. Total KLD was used to assess the overall magnitude of distributional change. Feature-level contributions were used to identify which unigrams or POS trigrams contributed most to the observed difference.

For feature-level reporting, a two-sided unpaired Welch's t-test was used to filter features that are involved in change. For a feature f, each abstract was first represented by its document-level relative frequency within the retained feature set. Let $\bar{r}_L(f)$, $s^2_L(f)$, and n_L denote the mean, sample variance, and number of valid documents for the later group, and let $\bar{r}_E(f)$, $s^2_E(f)$, and n_E denote the corresponding values for the earlier group. The Welch statistic was:

$$t_f = \frac{\bar{r}_L(f) - \bar{r}_E(f)}{\sqrt{\frac{s^2_L(f)}{n_L} + \frac{s^2_E(f)}{n_E}}} \tag{7}$$

The test did not assume equal variance between the earlier and later groups. Features shown in the unigram and POS-trigram contributor figures were required to pass an uncorrected threshold of $p < 0.05$. This filter was used only for feature-level display and trajectory selection; the total KLD values were still calculated over the full retained vocabulary. These feature-level results remain descriptive and require manual context checks before being interpreted as topic shifts, style shifts, or LLM-associated markers.

2.6 Figure Generation and Interpretation

Seven figures were generated based on the processed results: adjacent-year unigram KLD, pre/post top unigram contributors, selected unigram trajectories, adjacent-year POS trigram KLD, pre/post top POS trigram contributors, selected POS trigram trajectories, and a unigram/POS trigram comparison. These figures serve as a descriptive summary of the corpus. They are not interpreted as evidence that LLMs caused an increase in any individual word, topic, or syntactic pattern.

The Python code used to run the analysis process is released through GitHub via the link `https://github.com/chengzechen/language-shift-in-LLM-era-research`.

3 Results

3.1 Corpus Characteristics

After excluding 2026 from the reported KLD analysis, the final corpus for analysis comprised 112,689 abstracts. The sample sizes for each year were broadly similar, ranging from 18,109 abstracts in 2020 to 19,085 abstracts in 2025. The average length of abstracts increased from 205.85 tokens in 2020 to 215.84 tokens in 2025, whilst the median length rose from 190 tokens to 205 tokens. This trend is crucial for interpreting the results, as differences between years may partly reflect changes in abstract length and the composition of the corpus.

Table 1: Corpus characteristics by publication year.

Year	Clean abstracts	Total tokens	Mean tokens	Median tokens
2020	18,109	3,727,691	205.85	190
2021	18,572	3,866,537	208.19	195
2022	18,959	3,979,670	209.91	197
2023	18,900	3,974,499	210.29	198
2024	19,064	4,049,616	212.42	200
2025	19,085	4,119,291	215.84	205

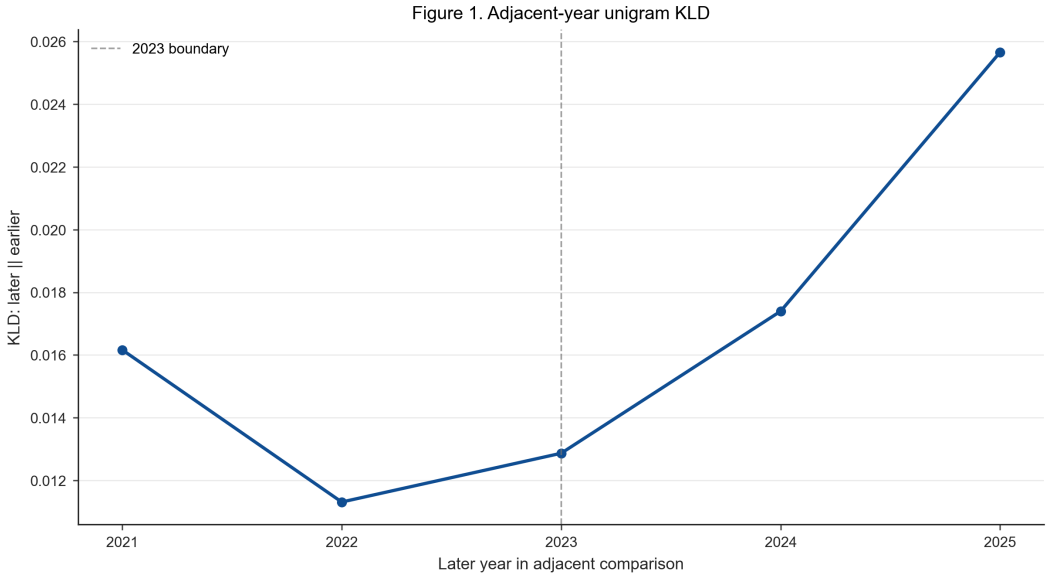

Figure 1: KLD of unigrams between adjacent years from 2020 to 2025. Each value represents the comparison between the following year and the preceding year. The greatest difference in unigrams between adjacent years was observed between 2024 and 2025, whilst the comparison between 2022 and 2023 showed a relatively small difference.

As shown in Figure 1, unigram KLD varied across different year-on-year comparisons. The minimum value occurred between 2021 and 2022 (KLD = 0.0113), whilst the maximum value occurred between 2024 and 2025 (KLD = 0.0257). Consequently, the comparison between 2022 and 2023 does not represent a peak between consecutive years (KLD = 0.0129). The unigram pattern exhibits a later increase, with the strongest fluctuation occurring between 2024 and 2025.

A comparison of the aggregated periods also reveals a measurable difference in unigram distribution between 2020–2022 and 2023–2025 (KLD = 0.0256). This supports the main hypothesis at the descriptive level, as the unigram distributions for these two aggregated periods are not identical. However, the results for adjacent years suggest that the timing of this shift is not so straightforward.

3.2 Pre/Post Unigram Contributors

In the pre/post unigram comparison, 3,117 of the 10,008 retained unigram features passed the Welch filter at $p < 0.05$, including 1,680 features with positive later-versus-earlier KLD contributions.

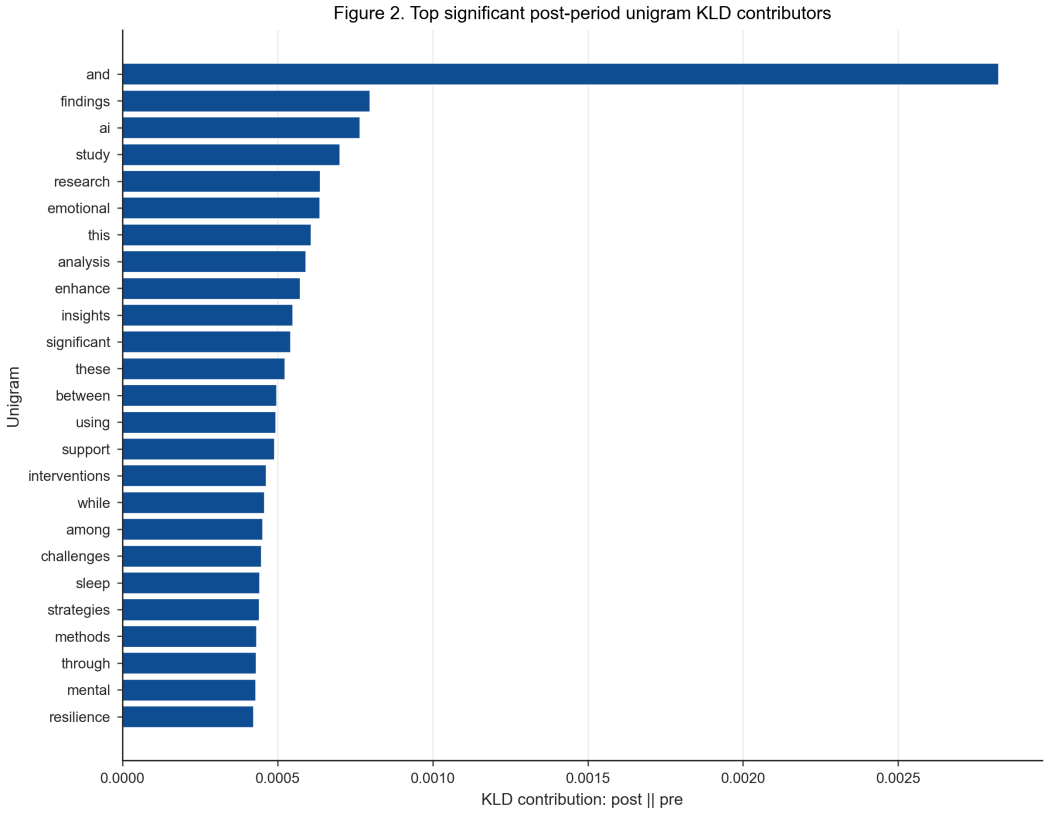

Figure 2: The unigrams making the greatest contribution to the KLD comparison between 2020–2022 and 2023–2025 after passing the unpaired Welch's t-test filter at $p < 0.05$. The bar chart illustrates the contribution to differences at the feature level. Some contributors may be topic terms, whilst others may reflect style or discourse organisation.

Figure 2 shows that the interpretation of the top Welch-filtered unigrams is rather complex. The feature with the greatest contribution is "and", whose frequency increased from 466,105 occurrences before the comparison to 512,197 occurrences after the comparison and passed the Welch filter. As "and" is a high-frequency function word, it should not be over-interpreted as a meaningful LLM marker until a more specific analysis of function words has been conducted.

The frequency of the term "ai" increased from 736 occurrences in 2020–2022 to 3,861 in 2023–2025. This may reflect the growing importance of AI-related research in the field of psychology. The meanings of some other high-contribution terms are more ambiguous. "Findings" increased from 14,366 to 20,637, "enhance" from 1,796 to 4,878, "insights" from 1,319 to 4,100, and "significant" from 17,696 to 22,577. These terms may overlap with broader academic stylistic patterns discussed

in recent LLM writing literature; however, manual contextual verification is required before a more precise interpretation can be made.

Some high-contribution terms are clearly related to psychological topics or methodologies. The frequency of the term "sleep" increased from 13,417 to 17,329, "emotional" from 11,671 to 16,694, and "interventions" from 5,789 to 9,159. These terms may indicate a shift in the subject matter of OpenAlex psychology abstracts, rather than merely a change in writing style. Consequently, the results suggest that shifts in subject matter, stylistic changes, and alterations in corpus composition collectively influence the content of the abstracts.

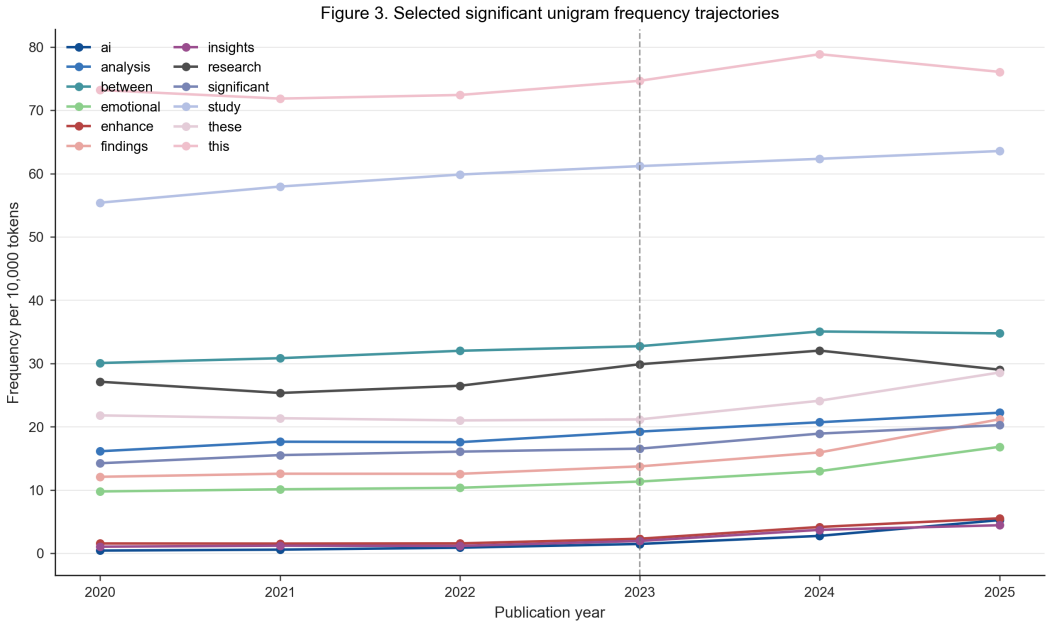

Figure 3: Selected unigram trajectories from 2020 to 2025, with frequency shown per 10,000 tokens. The selected vocabulary includes terms related to artificial intelligence, potential academic-style terms, and psychology-related keywords. The high-frequency word "and" is not shown in the figure, as its extremely high frequency would dominate the entire chart, making it difficult to analyse trends in lower frequency words.

Figure 3 clearly illustrates the temporal trajectories of these word-level changes. The frequency of "ai" increased from 0.43 per 10,000 tokens in 2020 to 5.24 in 2025. The frequency of "findings" increased from 12.09 per 10,000 tokens to 21.18, whilst that of "enhance" rose from 1.55 in 2020 to 5.53 in 2025. The frequency of topic-related vocabulary has also increased; for example, the frequency of "emotional" rose from 9.77 per 10,000 tokens to 16.83. These trends indicate that the unigram signal is not limited to a single step change in 2023, but extends across subsequent years in the corpus.

3.3 POS Trigram Divergence

As shown in Figure 4, in all comparisons of adjacent years, the differences in POS trigrams are smaller than those in unigrams. The POS trigram KLD is lowest for 2022–2023 (KLD = 0.0017), whilst the POS trigram KLD is highest for 2024–2025 (KLD = 0.0114). The comparison of aggregated POS trigram distributions also reveals a measurable difference (KLD = 0.0092). The POS trigram signal was therefore present but smaller than the unigram signal.

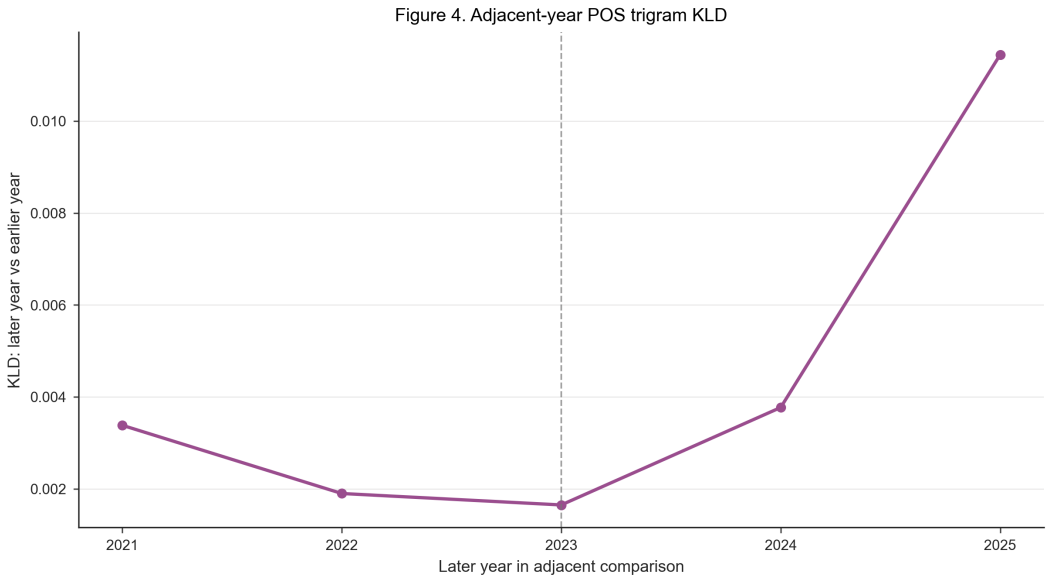

Figure 4: Adjacent-year POS trigram KLD from 2020 to 2025. The differences in POS trigrams are smaller than those in unigrams, but they also begin to increase after 2023, particularly in the comparison between 2024 and 2025.

In the pre/post POS trigram comparison, 886 of the 1,731 retained POS trigram features passed the Welch filter at $p < 0.05$, including 258 features with positive later-versus-earlier KLD contributions.

Figure 5 shows that several high-contribution POS trigrams involved adjective and noun patterns after the Welch filter. `VERB ADJ NOUN` increased from 93,143 occurrences in the pre period to 123,398 in the post period. `ADJ NOUN NOUN` increased from 182,555 to 212,013, `NOUN VERB DET` increased from 62,755 to 83,511, and `DET NOUN VERB` increased from 91,050 to 112,766. These patterns may suggest that later abstracts make greater use of descriptive or refined academic expressions. However, these POS trigrams are broad categories. For example, `ADJ NOUN NOUN` could represent many different phrase types, topics, and local contexts. Therefore, one cannot simply state that a precise change in syntax has occurred. The results of the POS trigrams are best viewed as a signal indicating that certain structural patterns warrant further inspection.

Figure 6 shows that the selected POS trigram contributors generally increased over the study period. `VERB ADJ NOUN` rose from 83.60 per 10,000 POS trigrams in 2020 to 124.02 in 2025. `ADJ NOUN NOUN` rose from 163.42 to 198.95, `NOUN VERB DET` rose from 56.49 to 80.94, and `NOUN VERB ADJ` rose from 33.12 to 56.98. These changes suggest a possible shift in grammatical patterns; however, given the relatively simple design of the POS trigram layer, these changes should be regarded as exploratory.

3.4 Comparison of Lexical and POS-Trigram Signals

Figure 7 compares the two feature layers directly. Unigram KLD was consistently higher than POS trigram KLD, suggesting that the clearest signal in this corpus is lexical rather than grammatical. This pattern is consistent with the design of this study: unigrams are more sensitive to topic and style words, while POS trigrams compress many different contexts into a smaller sequence of part-of-speech tags.

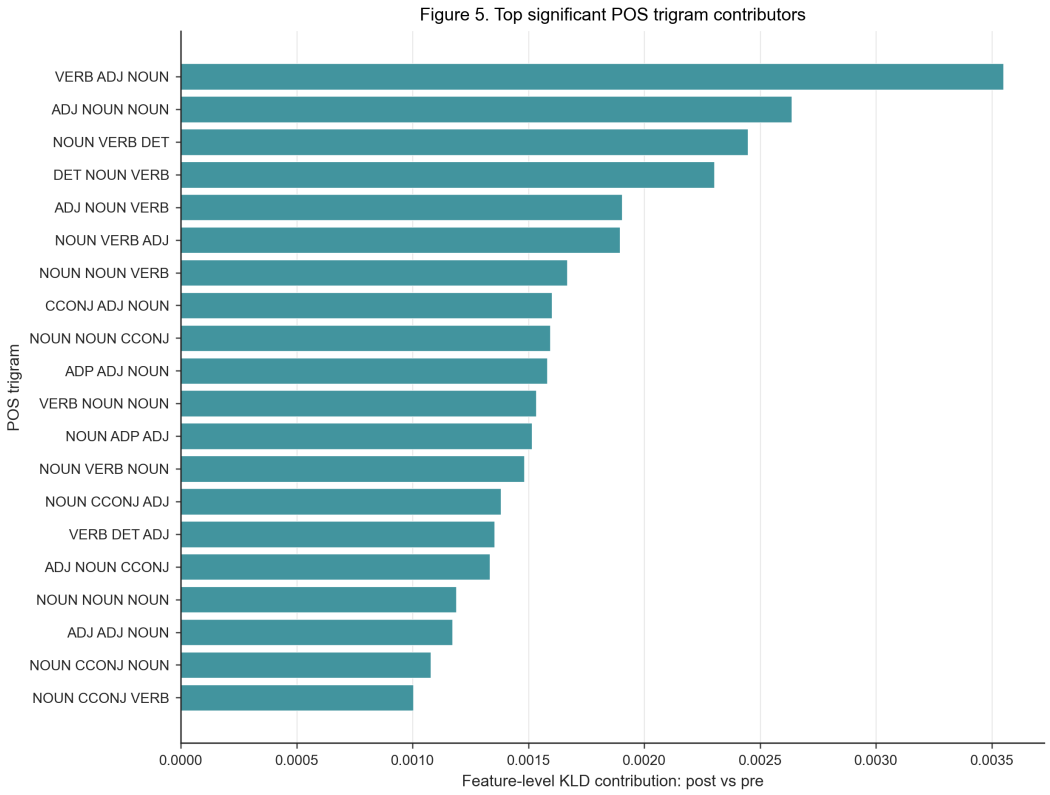

Figure 5: Top POS trigram contributors to the pooled 2020–2022 versus 2023–2025 KLD comparison after passing the unpaired Welch's t-test filter at $p < 0.05$. The strongest contributors include adjective-noun and noun-heavy patterns.

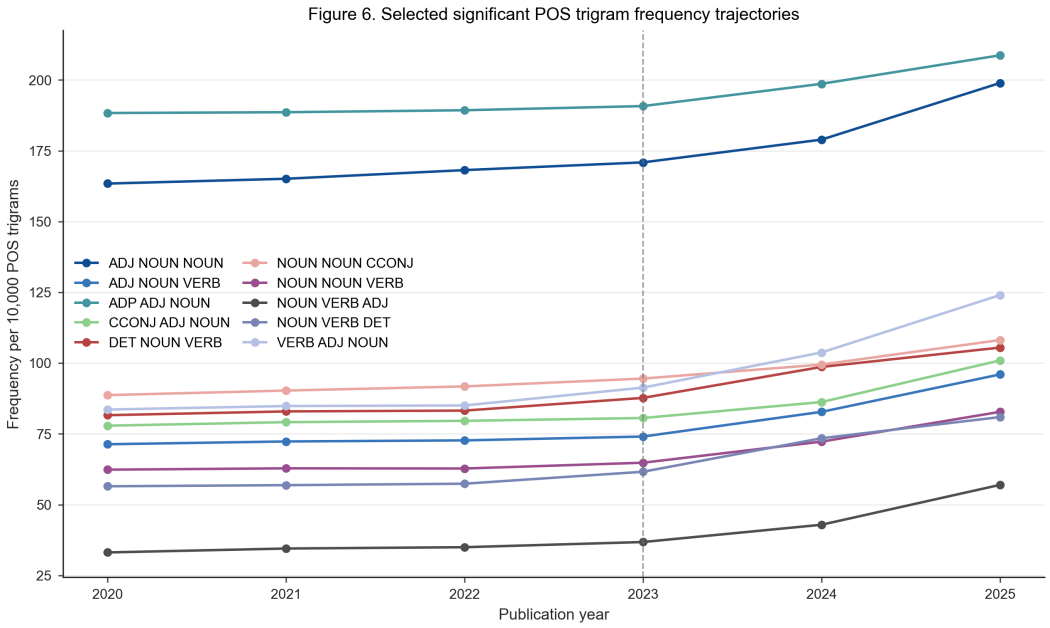

Figure 6: Selected POS trigram trajectories from 2020 to 2025, shown as frequency per 10,000 POS trigrams. Several selected adjective-noun and noun-heavy patterns increase over time.

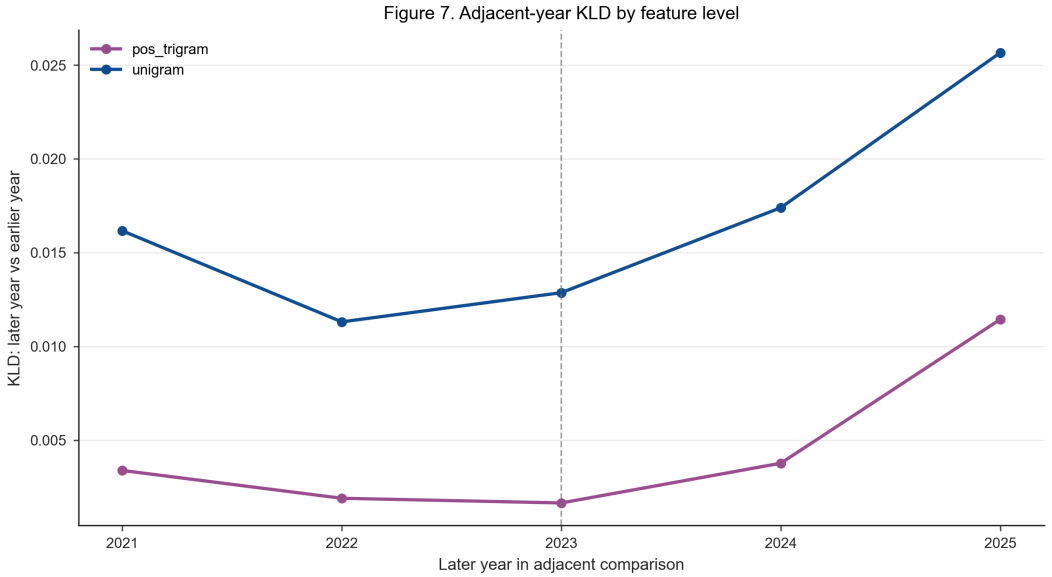

Figure 7: Comparison of adjacent-year KLD for unigram and POS trigram feature layers. Unigram divergence is consistently larger, suggesting that lexical, topical, and style-related shifts are more pronounced than the coarse POS trigram signal.

Overall, the hypotheses were partially supported. The primary unigram hypothesis was supported descriptively because the pooled pre/post comparison showed measurable divergence. The simple abrupt-break expectation was not supported, because the strongest adjacent-year shift did not occur at the 2022-to-2023 boundary. The secondary POS trigram hypothesis was partially supported because POS trigram distributions also changed, but the signal was smaller and less directly interpretable than the unigram signal.

4 Discussion

Overall, the results indicate that OpenAlex psychology abstracts changed measurably between 2020–2022 and 2023–2025. This was most evident in the unigram analysis, where the pre/post comparison produced a KLD of 0.0256. The POS trigram analysis also revealed a pre/post difference (KLD = 0.0092), albeit to a lesser extent. Consequently, the primary contribution of this pilot study is a map of distributional language change observable in the OpenAlex psychology corpus from 2022 onwards. These findings are consistent with the possibility of a distributional signal during the gradual adoption of LLMs, but do not indicate that LLMs caused the observed changes.

The results of the unigram analysis are best interpreted as a combination of shifts in both subject matter and style. The increase in "AI" is one of the most prominent thematic signals, as psychological research itself is increasingly focusing on artificial intelligence. Other contributing terms, such as "sleep", "emotional" and "interventions", also appear to be closely related to psychological themes or research domains. These terms help to distinguish between the two merged distributions, but cannot be regarded as LLM style markers.

At the same time, some unigram contributing terms may reflect shifts in abstract style or the language of scientific reporting. Terms such as "findings", "enhance", "insights" and "significant" may indicate changes in how abstracts organise their communication. However, this interpretation remains at a preliminary stage. The Welch filter reduces the risk of foregrounding features that differ

only through random fluctuation, but feature-level KLD contribution terms still merely indicate that a specific word helps distinguish between two distributions. They do not reveal why a word has a high contribution rate, or whether it was generated by LLMs. The next step should be to identify these high-contribution words through manual inspection of concordance lines or abstract contexts.

The results of the POS trigram analysis provide more detailed supporting evidence. Several high-contribution trigrams involved adjective-noun or noun-heavy structures, including `VERB ADJ NOUN`, `ADJ NOUN NOUN`, and `ADJ ADJ NOUN`. This may suggest that academic terminology in subsequent abstracts is more refined or descriptive. However, the POS trigram layer has been deliberately simplified and does not identify complete syntactic structures. Therefore, findings regarding POS trigrams should be viewed as prompts for further analysis.

The timing of the results is also crucial. If the release of large language models (LLMs) directly led to a shift in abstract language, the contrast between 2022 and 2023 should have produced the clearest distributional break. However, unigram KLD between 2022 and 2023 was not pronounced; the period with the greatest unigram change was 2024 to 2025. The period showing the largest POS trigram change was also the comparison between 2024 and 2025. This does not rule out the influence of large language models, as various factors, such as publication cycles and gradual AI adoption, may exert greater effects. However, it does undermine the persuasiveness of the simplistic assertion that "psychology abstracts suddenly changed after the release of ChatGPT".

Interpretation of this study should take the following limitations into account. First, the data for 2026 covers only the first half of that year; consequently, the KLD analysis presented in this report does not include data from 2026. This implies that further research is required before conclusions can be drawn regarding linguistic changes in 2026 and beyond.

Second, the contributors at the feature level require manual contextual validation. It is not yet possible to reliably attribute the observed KLD patterns to writing style. This limitation is particularly pronounced for high-frequency function words, as even subtle shifts in the distribution of very common features can lead to significant KLD effects. The current Welch test uses an uncorrected $p < 0.05$ threshold as an display filter; it should not be interpreted as confirmatory evidence after many simultaneous feature tests. Consequently, future research should compare the results of full-lexical analysis with those of robustness analyses in which function words are treated separately or excluded, and should test whether the feature rankings remain similar after multiple-comparison correction.

Third, the current analysis does not control for subfields, journals, or article types within psychology. The psychology abstracts in OpenAlex may contain heterogeneous research from other psychology-related fields. If the dataset permits, future research should include control variables such as subfields, journals, and article types.

Fourth, the POS trigram design is relatively simple. Although POS trigrams can serve as a preliminary approximation of syntax, they do not constitute a complete syntactic model. Furthermore, they compress many different phrases into a single unit, which makes interpretation difficult. A more in-depth structural analysis could incorporate phrase-level features, dependency patterns, or manual examination of the local context underlying high-contribution POS trigrams.

Future work should therefore treat this report as an early complete loop. The most effective next steps would be to rerun the analysis once complete 2026 publication data is available, manually examine the context of the highest-contributing unigrams and POS trigrams, compare results for the full vocabulary against those controlled for function words, and verify whether the results remain similar after controlling for subfields, journals, or article types.

In summary, this pilot study found measurable differences in the distribution of lexical and POS trigram features in OpenAlex psychology abstracts during the period of LLM adoption, with the

strongest signal being lexical variation. The results are consistent with broader distributional shifts during the period of LLM adoption, but they remain descriptive and require further validation.

References

[1] Kun Sun, Haitao Liu, and Wenxin Xiong. The evolutionary pattern of language in scientific writings: A case study of philosophical transactions of royal society (1665–1869). *Scientometrics*, 126:1695–1724, 12 2020. doi: 10.1007/s11192-020-03816-8.

[2] Douglas Biber and Bethany Gray. *Grammatical Complexity in Academic English*. Cambridge University Press, 05 2016.

[3] Yuri Bizzoni, Stefania Degaetano-Ortlieb, Peter Fankhauser, and Elke Teich. Linguistic variation and change in 250 years of english scientific writing: A data-driven approach. *Frontiers in Artificial Intelligence*, 3, 09 2020. doi: 10.3389/frai.2020.00073.

[4] Richard Van Noorden and Jeffrey M. Perkel. Ai and science: what 1,600 researchers think. *Nature*, 621:672–675, 09 2023. doi: 10.1038/d41586-023-02980-0. URL `https://www.nature.com/articles/d41586-023-02980-0`.

[5] Andrew Gray. Chatgpt "contamination": estimating the prevalence of llms in the scholarly literature. *arXiv (Cornell University)*, 03 2024. doi: 10.48550/arxiv.2403.16887.

[6] Tom S Juzek and Zina B Ward. Why does chatgpt "delve" so much? exploring the sources of lexical overrepresentation in large language models, 2024. URL `https://arxiv.org/abs/2412.11385`.

[7] Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D Manning, and James Y Zou. Mapping the increasing use of llms in scientific papers. *arXiv (Cornell University)*, 04 2024. doi: 10.48550/arxiv.2404.01268.

[8] Dmitry Kobak, Rita González-Márquez, Emőke-Ágnes Horvát, and Jan Lause. Delving into llm-assisted writing in biomedical publications through excess vocabulary. *Science Advances*, 11, 07 2025. doi: 10.1126/sciadv.adt3813.

[9] Riley Galpin, Bryce Anderson, and Tom S Juzek. Exploring the structure of ai-induced language change in scientific english. *The International FLAIRS Conference Proceedings*, 38, 05 2025. doi: 10.32473/flairs.38.1.138958.

[10] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86, 03 1951. doi: 10.1214/aoms/1177729694. URL `https://www.projecteuclid.org/euclid.aoms/1177729694`.

[11] Stefania Degaetano-Ortlieb and Elke Teich. Using relative entropy for detection and analysis of periods of diachronic linguistic change. *ACL Anthology*, pages 22–33, 08 2018. URL `https://aclanthology.org/W18-4503/`.

[12] Peter Fankhauser, Jörg Knappen, and Elke Teich. Exploring and visualizing variation in language resources. *ACL Anthology*, pages 4125–4128, 05 2014. URL `https://aclanthology.org/L14-1191/`.