What Are LLMs Doing to Scientific Communication? Measuring Changes in Writing Practices and Reading Experience

Filip Miletić* **Neele Falk***
Institute for Natural Language Processing, University of Stuttgart, Germany
{filip.miletic, neele.falk}@ims.uni-stuttgart.de

Abstract

Has the style of scientific communication changed due to the growing use of large language models in the writing process? We address this question in the domain of Natural Language Processing by leveraging two data resources we create: a naturalistic corpus of over 37,000 papers from the ACL Anthology (2020–2024); and a synthetic dataset of 3,000 human-written passages and their LLM-generated improvements. We first implement a series of diachronic lexical analyses, showing that both word frequency and usage contexts have changed significantly over time, indicating semantic specialization in some cases and generalization in others. Broadening our perspective, we then model a range of more complex stylistic features and find that LLM-modified texts more frequently contain certain syntactic constructions, more complex and longer words and a lower lexical diversity. Finally, we connect these changes in writing practices to subjective reading experience through a pilot annotation study with 20 domain experts. They overall rate LLM-improved texts as more understandable and exciting, but also express negative qualitative attitudes towards LLMs, highlighting the strongly subjective effect of AI-assisted writing on reading experience.

Keywords: AI-assisted writing, scientific communication, language change

1. Introduction

Large language models (LLMs) are increasingly used to assist human writing, including in high-stakes domains such as scientific communication. The rapid and pervasive nature of these changes raises the question of the ways in which they may be altering prevalent writing practices (e.g., lexical and stylistic choices), and of the subsequent effect of those practices on reading experience (e.g., perceived clarity and trustworthiness of texts).

While evidence of these changes is beginning to emerge, prior work is limited in two major respects. First, several recent studies examine the distinctive linguistic properties of LLM-generated scientific text (Ma et al., 2023; Muñoz-Ortiz et al., 2024; Zanotto and Aroyehun, 2024; Zamaraeva et al., 2025), but they generally compare texts which are written by humans vs. entirely model-generated. This clear-cut distinction oversimplifies finer-grained practices: LLMs are typically used to improve human-written text rather than generate entire passages (Koller et al., 2024; Kobak et al., 2025); moreover, real-life documents tend to alternate between human-only and LLM-improved writing rather than contain a uniform amount of generated text (Lee et al., 2022; Richburg et al., 2024). The second major shortcoming is the focus on identifying distinctive properties of generated text without systematically measuring their effect on human readers. Even where included, such measurements are limited to broad patterns such as the ability to distinguish human vs. LLM writing (Gao et al., 2023; Ma et al., 2023). As a result, the link between objective differences in writing style and more subjective but vital dimensions of reading experience remains to be established.

This paper aims to provide a more realistic and comprehensive assessment of LLM use in scientific communication. We design our study so as to capture the collaborative nature of human–LLM writing and the uneven spread of such interventions within a document, as well as to explicitly connect the distinctive features of these writing practices with subjective reading experience. We conduct our analysis on NLP papers from the ACL Anthology, and define two periods of around two years each, respectively preceding and following the release of ChatGPT in November 2022. We view the two periods as reflecting community-level writing practices which do not include any vs. may include some writing by public-facing LLMs. Complementing this naturalistic scenario, we simulate the real-life use of LLMs in a more controlled synthetic setting: we sample 3,000 extracts from pre-ChatGPT papers and generate model-improved versions of those. We pose the following research questions:

RQ1 In what way did core linguistic choices change between these two periods?

RQ2 To what extent are more complex stylistic properties specific to each of the two periods?

RQ3 Do these differences in writing practices give rise to different reading experiences?

We first assess differences in linguistic choices by deploying a series of diachronic lexical analyses: statistical corpus measures to identify emergent terms; type-level word embeddings to characterize broad changes in their semantic properties;

*Equal contribution.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1811–1830
11–16 May 2026.

and token-level word embeddings to automatically retrieve their time-specific uses. Broadening our focus, we then investigate how different linguistic features — such as text length, sentiment, grammatical and lexical variability, and readability — contribute to explaining variation between human and LLM-assisted writings through regression analysis. Finally, we conduct an annotation study contrasting human-written texts with their LLM-produced improvements, and ask 20 domain experts for ratings of reading experience in terms of clarity, authenticity, trustworthiness, and excitement.

We provide the following contributions. **(1)** We show that post-ChatGPT papers are distinguished by more complex lexical choices (e.g., *enhance* rather than *improve*) and further stylistic properties (e.g., lower lexical diversity). By comparing naturalistic data from the ACL Anthology and synthetic data from text generation experiments, we confirm that these writing practices can be attributed to LLM use. **(2)** We further find that these stylistic changes are linked to differences in subjective reading experience, with LLM-improved texts perceived as clearer and more exciting. **(3)** We release an updated version of the ACL-OCL corpus (Rohatgi et al., 2023), containing PDF-extracted text of 99.3k papers from the ACL Anthology. We also provide a one-line script to ingest future papers. **(4)** We release 3,000 pairs of human-written texts and their LLM-produced improvements, and annotations of human reading experience for 200 pairs.[1]

2. Related Work

Detection of AI-generated content. Paralleling the rise in popularity of LLMs, recent years have seen growing interest in automatic detection of AI-generated content. This includes the release of datasets and benchmarks to train detection tools and evaluate different methods (e.g., Chen et al., 2023; Li et al., 2024; Guo et al., 2023; Dugan et al., 2024; Macko et al., 2023; Wang et al., 2024). Techniques for detecting AI-generated text include watermarking (Zhao et al., 2025), fine-tuning transformer-based classifiers (Guggilla et al., 2025), using model-related features (Wu et al., 2025) or linguistic features (Hamed and Wu, 2023).[2] While good results are generally achieved for AI-generated content, the detection of human–AI coauthored text remains a major challenge and requires adaptation of existing models (Richburg et al., 2024; Su et al., 2025). Early work includes the CoAuthor dataset (Lee et al., 2022), which includes essays augmented with GPT-3 suggestions, while more recent datasets focus on different variations of human–AI co-authored texts (e.g., human-written then machine-polished) (Wang et al., 2025).

Stylistic differences between human-written and AI-generated or AI-modified text. Several works more directly explore the difference in stylistic features of human and AI-generated text. Domains that are mostly covered are news articles (Muñoz-Ortiz et al., 2024; Zamaraeva et al., 2025), essays (Akinwande et al., 2024) and abstracts of scientific articles (Ma et al., 2023). Existing works examine features from all possible categories, such as the frequency of certain syntactic constructions, n-grams, hedging, lexical complexity, rhetorical properties and sentiment. Frequent linguistic peculiarities in AI-generated content include, e.g., lower lexical variation (Zanotto and Aroyehun, 2024; Akinwande et al., 2024; Yildiz Durak et al., 2025), more positive sentiment (Muñoz-Ortiz et al., 2024; Zamaraeva et al., 2025), fewer compounds (Zamaraeva et al., 2025), and the excessive use of certain verbs and modifiers such as *delve*, *crucial*, or *intricate* (Gray, 2024; Kobak et al., 2025; Reinhart et al., 2025).

Several works use these features to predict whether a text was human or AI-generated and to identify the strongest predictor (Ma et al., 2023; Desaire et al., 2023; Akinwande et al., 2024). Some works also investigate human perception of LLM-generated texts, e.g. Gao et al. (2023) and Hakam et al. (2024) find that human annotators struggle to distinguish between human and LLM-generated scientific texts. Russell et al. (2025) show that annotators with frequent LLM-writing experience better detect generated news. In Doru et al. (2025), participants classified scientific texts and rated their fluency, quality, and coherence. Lin and Zhu (2025) find that researchers use LLMs mainly to improve clarity and conciseness, leading to a more homogeneous writing style since ChatGPT's release.

Most prior studies focus on fully generated texts and rarely compare human–AI co-authored vs. human-only writing. For this reason, we compare articles published after ChatGPT's release with those from shortly before, expecting weaker but detectable linguistic shifts even if only a fraction were LLM-modified. Further, frequent exposure to LLM-generated language may lead researchers to unconsciously adopt its style. Unlike most prior studies, we analyze full papers rather than abstracts, since LLM use likely occurs across all sections. While LLM-related vocabulary and linguistic features have been studied, they are rarely examined together, and existing research often focuses on surface-level trends. Lexical choices, in particular, remain underexplored beyond frequency-based analyses, despite well-established methods

[1] Data and code are available at https://github.com/FilipMiletic/ScientificCommunication

[2] We refer to the survey by Wu et al. (2025) for a comprehensive overview on detecting AI-generated content.

for modeling semantic change (Tahmasebi et al., 2021; Schlechtweg, 2023). Finally, the subjective perception of LLM-generated content in scientific texts has hardly been studied, which is why we complement our data-driven analysis with a pilot study of reading experience with 20 domain experts.

3. Data

We now present our two English-language data resources: a naturalistic corpus of NLP papers from the ACL Anthology (henceforth *original* dataset); and a synthetic dataset of human-written passages and their LLM-generated improvements (henceforth *LLM* dataset).

3.1. ACL Anthology Corpus

Since the focus of our work is on scientific communication in the NLP community, we analyze the papers from the ACL Anthology,[3] the open-access publication repository of the Association for Computational Linguistics (ACL). As our starting point, we use the ACL-OCL corpus (Rohatgi et al., 2023) containing ca. 73,000 papers. They were obtained by crawling the Anthology website for PDF files and then extracting the full text using GROBID.[4]

The content in the original corpus ends in September 2022 and to our knowledge has not been updated since. We therefore implement an update to bring its temporal span closer to the present day. We also note two other recurrent problems. Some papers from the original time span are available in the Anthology but not included in the corpus, possibly due to coverage issues during crawling. Other papers are included in the corpus, but are associated with metadata without full text content due to file issues (e.g., failed extraction with GROBID or PDF missing from the Anthology at the time of the crawl, especially for very early conferences).

In our update, we do not crawl the Anthology website but instead use its BibTeX export as the most comprehensive structured record of available papers. We rely on BibTeX information to identify the missing papers based on citation keys, extract their metadata, and reconstruct their URLs. We download the corresponding PDF files and then use GROBID to extract paper text. This process also recovers the textual content for a subset of papers lacking it in the original corpus; we remove any remaining papers without textual content. We accompany the updated corpus with code (run as a one-line command) which checks the locally available papers against those in the Anthology and passes any missing papers through the full update pipeline.

[3] https://aclanthology.org

[4] https://github.com/kermitt2/grobid

	Time period	Papers	Paras	Tokens
t_1	Jan 2020 – Dec 2022	20,259	801,016	100.2 m
t_2	Jul 2023 – Dec 2024	17,501	831,077	103.6 m
	Total	37,760	1,632,093	203.8 m

Table 1: Distribution of papers across time periods

The updated corpus contains 99.2k papers published until the end of 2024. For the purposes of our study, we define a subcorpus structured into two time periods around a critical point in time regarding LLM use: the release of ChatGPT in November 2022. The first time period (t_1) contains papers published from 2020 to 2022. Its last major conference event is EMNLP 2022, which had a camera-ready deadline in October of that year. The second time period (t_2) contains papers published in the second half of 2023 and in 2024. It begins with ACL 2023, whose camera-ready deadline was in May of that year, i.e., six months after the release of ChatGPT. The gap between the two periods ensures a clear distinction between them while limiting the effect of changes in topic over time. We only retain papers from events that took place in both time periods.

While we assume that these sampling constraints ensure good comparability of t_1 and t_2, we also inspect it more closely by running a topic analysis. We find some topical shifts in line with the evolution of the field (e.g., a stronger focus on individual levels of linguistic structure in t_1, and prominence of more recent machine learning methods in t_2), but the analysis overall confirms broad topical comparability of the two time periods. Detailed results are reported in Appendix A.

We preprocess PDF-extracted text using spaCy[5] (model `en_core_web_sm`). We segment the text into sentences, which are then tokenized, lemmatized, and POS-tagged. We run a subset of analyses on paragraph-level, which we operationally define as non-overlapping windows of five sentences. Final corpus structure is shown in Table 1.

3.2. LLM-Assisted Paraphrases

The original sub-corpus described in Section 3.1 describes the realistic scenario in which t_2 contains a hybrid form of human and LLM co-authored text. To compare whether the patterns that emerge from the analysis on this are similar to those in texts explicitly modified by LLMs, we replicate this scenario in a controlled setup and construct a dataset with texts from t_1 paired with GPT-improved paraphrases, thus offering a clear gold label (human vs. LLM-modified).

We select a random sample of 3,000 publications from t_1 from 2022 and, for each paper in this,

[5] https://spacy.io

a random paragraph with a minimum length of 100 tokens. The selected paragraph is chosen from the initial paper paragraphs to make sure that it mostly spans text from the introduction. In the next step, we develop 10 different prompts that scientists frequently use during the writing process to refine their texts. To identify these prompts, we conducted an anonymous survey in which colleagues and students were asked to share the prompts they frequently use to improve their own writing. The 10 final prompts ultimately include both more general requests (*Improve the following*; *Please polish the following text*) as well as prompts which ask for the improvement of specific text dimensions (*Please improve the coherence of the text*; *Refine grammar, tone, and readability*). We then prompt GPT-3.5-turbo for each of the 3,000 original texts, randomly choosing one out of the 10 prompts which results in the final corpus. The final dataset (referred to as *LLM*) consists of 6k paragraphs, 3k human-generated and 3k LLM-modified.

Note that over the full period covered by our naturalistic corpus, researchers may have used a range of different LLMs, particularly toward the end of t_2 when newer models became available (e.g., GPT-4, Claude). However, at the time of publication of most papers in our dataset, the most widely available and commonly used system was ChatGPT based on GPT-3.5. Therefore, we used GPT-3.5-turbo to generate paraphrases in our experiments.

4. Characterizing Writing Changes via Lexical Choices

We begin by addressing **RQ1**: Are there core linguistic choices which changed between t_1 and t_2 in connection with LLM use? Focusing on the lexical level, we identify words with strongest changes in rates of use and then characterize them in terms of finer-grained patterns of semantic change.

4.1. Experimental Setup

Preprocessing Starting from preprocessed paper text (cf. Section 3.1), we examine content words (nouns, verbs, adjectives, and adverbs) in the shared vocabulary of t_1 and t_2. We lowercase all lemmas and retain those that are at least three characters long and contain only alphabetic characters.

Corpus statistics On the simplest level of analysis, we quantify the extent to which a word's rate of use has changed between t_1 and t_2. To determine the strength of the change, we compute the log-likelihood score (Rayson and Garside, 2000), which compares the observed frequencies of a given word in two corpora while accounting for the expected frequencies based on total corpus size. To determine the directionality of the change, we calculate the frequency ratio of a word's frequency in t_2 to its frequency in t_1 (normalized per million tokens). A value <1 is indicative of a term falling out of use over time, and vice versa. We calculate these statistics both for the original corpus and the LLM dataset.

Type-level word embeddings To better understand broad semantic patterns behind lexical choices, we train type-level word embedding models. We use word2vec (Mikolov et al., 2013) in the gensim implementation (Řehůřek and Sojka, 2010), and set the algorithm to skip-gram with negative sampling, window size to 5, vector dimensions to 100, minimum frequency to 10, and other hyperparameters to default values. We train separate models for t_1 and t_2 (original corpus), and run the process three times to account for randomness across word2vec runs (Pierrejean and Tanguy, 2018).

We characterize a word's usage in a given time period via its distributional neighborhood density, which reflects semantic features such as polysemy and is therefore long-established in research on language change (Sagi et al., 2009). To calculate a target word's neighborhood density, we select its 100 nearest neighbors, and take the mean of the cosine similarity scores between the target and each neighbor. We repeat the process for each of the three word2vec runs and take the average value. We then calculate the change in neighborhood density for word w between time periods t_1 and t_2 as $\Delta ND(w) = ND(w_{t_2}) - ND(w_{t_1})$. An increase in density is indicative of a restriction in usage contexts, typical of semantic specialization; a drop in density indicates diversification of usage contexts, which is typical of semantic generalization. We test the significance of changes in density using the Mann-Whitney–U test at 0.05 level.

Token-level word embeddings Connecting broad semantic trends to occurrence-level differences in usage, we implement an analysis using token-level embeddings. For a given target word, we collect the sentences in which it occurs in the original corpus; these are capped at 1,000 per time period and randomly subsampled if necessary. We then obtain contextualized embeddings of the target word in each sentence using ModernBERT (base model with 22 layers, 768 dimensions, 149m parameters; Warner et al., 2025). We feed the model with one sentence at a time, and retain the embedding corresponding to the target word in the last hidden state. The obtained embeddings for each target word are then clustered using using k-means, for which we rely on the scikit-learn implementation (Pedregosa et al., 2011) and set k to 8. For each cluster, we calculate the proportion

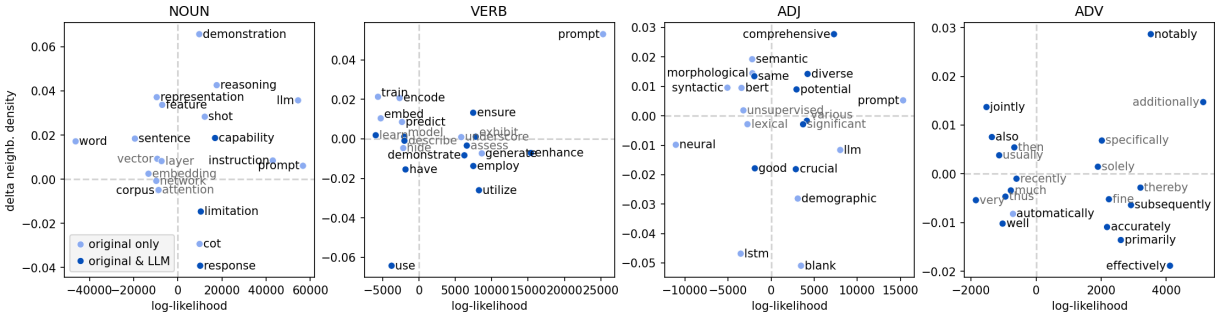

Figure 1: Target words with strongest differences in rates of use in t_1 vs. t_2 (top 10 per time period). X-axis: log-likelihood score, negative values indicate higher frequency in t_1. Y-axis: ΔND, higher values indicate an increase in neighborhood density over time (i.e., restriction of usage contexts); for grayed out targets, the difference in neighborhood density between t_1 and t_2 is not statistically significant. Color coding: dark blue targets also appear in the top 100 terms when contrasting original vs. LLM-paraphrased texts.

of examples from t_1 and t_2, and draw on this information to identify stable vs. shifting usages.

4.2. Results

On the most general level, we inspect changes in rates of use for the whole vocabulary based on the distribution of log-likelihood scores. This change is significant[6] for 13% (9,353 out of 71,993) words in the shared vocabulary as defined above. The significant changes are near-evenly distributed between drops (45%) and increases (55%) in frequency, matching the intuition that the obsolescence of most words is paralleled by a rise in prominence of their functional equivalents. We further analyze the words with significant changes in rates of use by correlating their log-likelihood score (to which we assign a negative value for words whose frequency drops over time) and the change in neighborhood density ΔND. We find a negative correlation (Spearman's $\rho = -0.26$, $p \ll 0.01$) suggesting that increased rate of use is associated with a lowering of neighborhood density, which typically reflects semantic generalization. However, the limited strength of the correlation indicates that the process does not apply to the whole vocabulary.

We now shift from vocabulary-level trends to the most relevant individual words. For each part-of-speech, we retain the words with the 10 highest positive and negative log-likelihood scores, respectively capturing the strongest increases and drops in rate of use over time. We plot these words in Figure 1, with log-likelihood on the x-axis and change in neighborhood density ΔND on the y-axis.

Changes in rates of use (x-axis) often reflect patterns of lexical replacement. Some reflect shifts in specialized topics, such as technical terms referring to core machine learning concepts dropping out of use (e.g., *embedding*), and those related to more recent approaches increasing their rate of use (e.g., *prompt*). Other patterns capture stylistic rather than terminological differences: general-language expressions which are semantically broad and stylistically neutral tend to fall out of use, while more formal and specialized words gain prominence. This trend is especially visible for verbs (e.g., *use* vs. *utilize*), adjectives (e.g., *good* vs. *comprehensive*), and adverbs (e.g., *then* vs. *subsequently*).

But can these changes be attributed to LLM-assisted writing? We inspect the log-likelihood scores calculated on our LLM dataset, comparing original t_1 texts and their LLM-assisted paraphrases. We select the words with 100 highest positive and negative log-likelihood scores (per part of speech). The overlap between this set and the top-10 sets from the original corpus is shown in Figure 1 using dark blue markers. The overlap is almost entirely limited to general-language and not technical terms. Given the topic-controlled nature of the LLM corpus, this finding indicates that (i) the changes in the original corpus reflect two parallel trends: a topical and a stylistic shift; and (ii) that the stylistic shift can be attributed to LLM use.

We now analyze the semantic change mechanisms associated with different rates of use, as measured by the change in neighborhood density ΔND (y-axis) and by the temporal distribution over clusters of token-level embeddings. As an example, we focus on the verbs *ensure* and *utilize*: they both show an increase in frequency over time, but differ in semantic mechanisms. The verb *ensure* shows an increase in neighborhood density, which is indicative of semantic specialization. This trend is supported by clustered examples like the following:

(1) Lastly, we aim to measure the semantic similarity between generated questions to **ensure** that the questions assess the same content.

(2) To **ensure** the high quality of the annotation procedure, we manually annotated a set of 200 control tasks.

[6] Based on the critical log-likelihood value of 15.13 recommended by Rayson et al. (2004).

Example (1) is from a cluster dominated by older data (65% t_1) where *ensure* conveys a broad sense of finality across diverse contexts. Example (2) comes from a more recent cluster (57% t_2). It is representative of the more specific meaning 'to guarantee', attested in a restricted set of transitive contexts usually referring to the quality of a scientific artifact.

In contrast, *utilize* shows a slight expansion of distributional neighborhood, typical of semantic generalization which is also borne out by these examples:

(3) The latter aims to **utilize** the multi-level interests to enhance both conversation and recommendation tasks when users chat with system.

(4) Finally, we **utilize** a ridge regression classifier to obtain final classification results.

Example (3) is from an older cluster (60% t_1) and conveys the specific meaning of 'use to the fullest potential'. Example (4) comes from a cluster with more recent data (56% t_2) where *utilize* is attested with the broad meaning 'make use of'.

Summarizing, we identified a substantial set of the vocabulary with a shift in the rate of use since the introduction of ChatGPT. As further validation, in Appendix B we also present complete corpus statistics for the strongest changes in rates of use, extend the same analysis to multi-word sequences, and analyze further clustering examples. We consistently find that some changes are due to topical shifts within the scientific community, while others are more clearly attributable to the adoption of LLMs. The changes are further reflected in shifting sense distributions, leading to semantic generalization in some cases and specialization in others.

5. Predicting Time Periods from Complex Linguistic Features

In the following section, we address **RQ2**: are there systematic stylistic differences between t_1 and t_2 that could be attributed to the use of LLMs as writing assistants?

5.1. Experimental Setup

We apply logistic regression as an analysis tool to find the linguistic markers that significantly contribute to explaining the outcome (a binary variable representing whether a text is from t_1 or t_2), after accounting for other relevant feature groups. The advantage of this approach is that it allows us to model the relationship between each feature and the target variable: positive odds indicate that higher values of a feature are characteristic of t_2 (human + LLM assistant), negative odds indicate that they are more characteristic of t_1 (humans only). We can also reveal which feature groups have the greatest impact.

Since many factors may influence stylistic change in our real-world data (original dataset), we repeat the same analysis in the controlled, synthetic setting (LLM dataset) to isolate stylistic effects specifically linked to LLM use.

Preprocessing We first extract around 1,000 linguistic features from six different features groups with `elfen` (Maurer, 2026) and `LFTK` (Lee and Lee, 2023) for all paragraphs from t_1 and t_2 in the original dataset, as well as all paragraphs in the LLM dataset. The features cover surface features (e.g., word and sentence length), morpho-syntactic features (e.g., occurrences of specific morphological constructions or part-of-speech tags), syntactic features (e.g., dependency relations and the complexity of dependency trees), psycholinguistic features (e.g., acquisition norms or average concreteness of words), lexical-semantic features (e.g., measures of lexical diversity or occurrences of specific named entities), and sentiment features (polarity and emotion-related words). All features are standardized and scaled.

Methodology As a first step, we apply a filtering method to select a meaningful set of features as predictors: We keep only features with variance >0 and perform a correlation analysis, retaining features with correlation <0.7 and selecting one representative feature per highly correlated group. This yields 145 initial features.

We then perform stability selection using logistic regression with elastic net regularization on a balanced subset of 200k paragraphs from the *original* dataset, applying 5-fold inner and 10-fold outer cross-validation to assess robustness. For our final selection we remove all features that (a) are not selected in all folds (a feature is selected when coefficient >0), (b) show inconsistent effects, (c) show less than $\pm$5% change in odds, and (d) were not significant (we bootstrapped CIs to approximate significance). This leaves 45 robust features. To further reduce the number, we rank them by absolute change in odds and select the top features per feature group, resulting in 24 final features. Finally, we refit the logistic regression on the full dataset for unbiased estimates. For the synthetic scenario, we use a generalized mixed-effects model with paper ID as a random effect to account for data dependencies.

5.2. Results

The regression model on the original dataset achieves a mean AUC of 0.65 $\pm$0.002 and a pseudo R^2 of 5.6% (McFadden) across the 10-fold cross-validation, which means that it explains not a large but meaningful amount of variance. Using the same

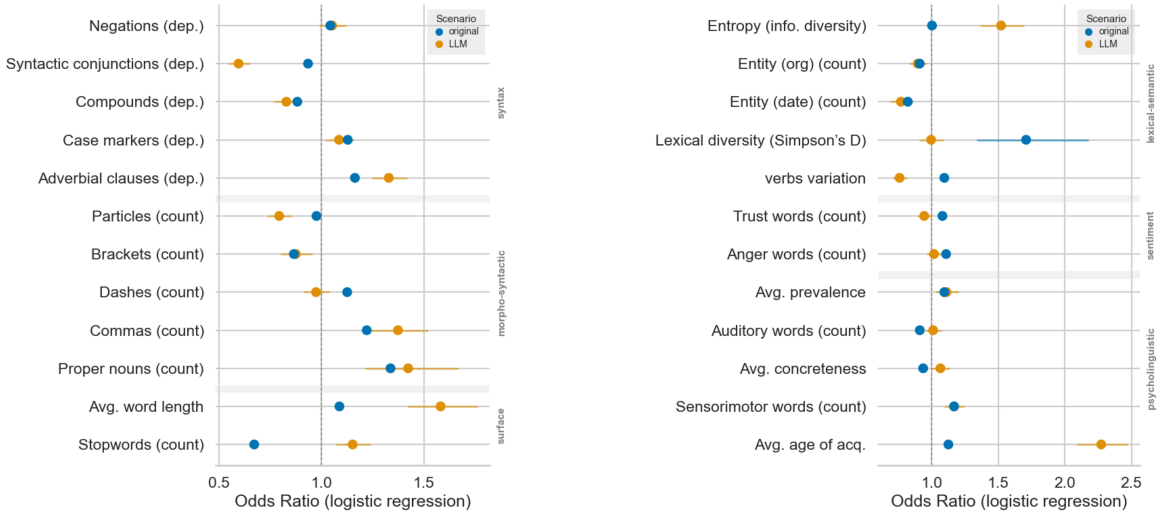

Figure 2: Comparison of odds ratios for linguistic features between the original and LLM-paraphrased datasets. Positive odds ratios indicate that higher feature values are more characteristic of the post-GPT / LLM-paraphrased texts. Horizontal bars represent 95% CIs (approximated with the Wald method for original). The vertical dashed line at 1 marks the point of no effect.

features as fixed effects in the mixed effects model explains 29% of the variance, which is a substantial portion and confirms that LLM-modified texts exhibit particular stylistic patterns that are characterized well by the linguistic features.

Figure 2 visualizes the direction and strength of association between each feature and the outcome variable (t_1 vs. t_2 or human vs. GPT-paraphrased), grouped by the different feature groups. Points to the right of the 1.0 decision boundary indicate that higher feature values are associated with LLM use, while points to the left correspond to human-written texts. When blue and orange points appear on the same side, the pattern is consistent across both the original and LLM datasets.

LLM use is characterized by longer and more complex words (higher average word length and age of acquisition), confirming the findings from the previous section. They show greater entropy while still containing many familiar, high-prevalence terms. In contrast, human-written texts use more stopwords, suggesting that LLM outputs are more semantically dense. Stopword use, however, presents a contrasting pattern: while it shows a strong positive association in the original corpus, the relationship is reversed in the LLM-modified texts. Human-written texts tend to be more varied, showing higher lexical diversity (Simpson's D) and greater verb variability, although this pattern is not entirely consistent.

Overall, the findings present a mixed picture of lexical variation and linguistic complexity: LLMs appear to produce more complex syntactic constructions and lexically dense content, yet with less lexical variety and a preference for familiar vocabulary. Whether this results in improved or reduced clarity will be examined in the following section.

In terms of syntactic and morpho-syntactic structures, there are distinct patterns between human and LLM-modified texts. LLMs tend to use more negations and adverbial clauses, indicating a preference for more elaborated or qualified sentence structures with additional modifiers. Human-written texts, in contrast, are more strongly associated with conjunctions – a construction often simplified or replaced with more complex connectives by LLMs – and contain more compounds.

We observe clear differences in punctuation use: human-written texts more often include brackets, whereas LLM-modified texts introduce more commas and possibly dashes. LLM-improved texts also contain more proper nouns, with the exception of organization entities, which occur more frequently in human-written texts. Similarly, date entities are more strongly associated with human-written texts, as LLMs are less likely to generate references and often remove them when prompted to improve human originals.

Sentiment patterns show a mixed picture: LLMs use more anger-related words, while original texts contain more trust-related words expressing confidence or credibility (e.g., *confirm*, *promise*, *support*) – although this trend is not confirmed in the paraphrased corpus. Another interesting pattern is the higher frequency of sensorimotor words in LLM-modified texts (e.g., *enhance*, *highlight*), possibly because LLMs were trained to make writing more dynamic and engaging. This could point to

a stylistic bias learned from human feedback and exposure to polished text.

Finally, we investigate the relative importance of each feature. Looking at the most important features, we find that dependency features (e.g., adverbial clauses) and punctuation have a major influence on both models, indicating a generalizable, strong pattern in terms of the difference between human and LLM-generated texts. Another strong feature in both datasets is the more frequent use of proper nouns in LLM-modified texts. The largest difference between the two datasets lies in their key predictive features: in the original corpus, variance is primarily explained by stopword use, whereas in the LLM dataset, it is driven by the average age of acquisition.

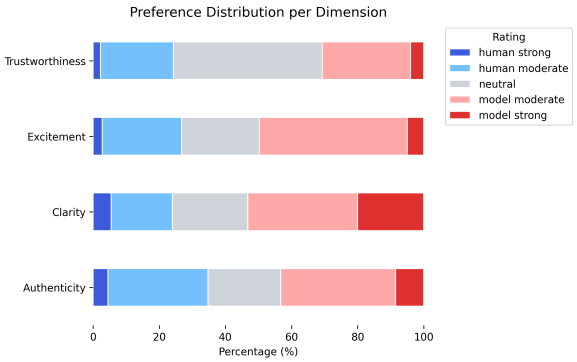

Figure 3: Distribution over Preferences in Human Raters. Higher red portion indicates a larger preference for LLM-improved texts, higher blue portion for the human originals.

6. Measuring Reading Experience

While prior research has examined lexical and stylistic differences between human- and LLM-generated texts, little is known about how readers perceive these changes. To address this gap, we conducted a pilot study to assess human perception of human-only and LLM-modified texts (**RQ3**).

To gain insights on that question, we look at four broader dimensions of reading experience: *clarity* measures whether a text is understandable and communicates the content clearly, *authenticity* captures to what extent the reader feels connected to the author and perceives the author as being genuine in their writing, *trustworthiness* captures the extent to which a text presents its arguments clearly, reliably, and credibly, while *excitement* assesses whether the reading experience is engaging and stimulates the reader's interest.

6.1. Annotation Setup

Study design To find out whether readers prefer human-written or LLM-improved texts and along which dimensions, we designed the annotation study as a pairwise comparison task. Given a pair of texts, text A and text B, annotators had to rate which one aligns more strongly with a certain statement. We measure each dimension with two statements, for example, *I read this text smoothly and fluently* to measure the dimension of *clarity*. Raters indicated their preference on a four-point Likert scale (strongly A, slightly A, slightly B, strongly B). We asked 20 different domain experts (NLP researchers) with varying backgrounds and levels of seniority to annotate 20 text-pairs. For each pair we collected annotations from two raters, resulting in 200 annotated instances. Appendix C provides the guidelines and detailed questionnaire.

Annotation data We use a subset of the LLM dataset to (a) control for topic effects and (b) ensure clear gold labels. To select good candidates, each original text and its paraphrase were converted into feature vectors capturing style (linguistic features) and semantics (using SBERT embeddings; Reimers and Gurevych, 2019, model `paraphrase-multilingual-mpnet-base-v2`). We calculated pairwise differences and ranked pairs by average distance. We then manually selected 200 high-quality pairs from the top 300, removing noisy instances, LLM artifacts, and formatting indicators.

6.2. Results

Quantitative ratings Figure 3 shows the overall preference distribution of human raters in our annotation study. We can see that raters tend to prefer LLM-improved texts, especially on the dimensions of clarity and excitement. However, there is also a substantial portion of texts where the human original was preferred, which shows that prompting an LLM to improve a scientific text does not always lead to the desired effect on the reader side. Additionally, a considerable number of texts showed no clear preference for either version.

For each dimension, we measure whether the ratings significantly deviate from 0 (neutral) and quantify the effect size. Table 2 shows that the strongest preference for LLM-improved texts can be observed in *clarity* and *excitement*. The clear preference for LLM-improved texts in *clarity* is unsurprising, as this aspect is often explicitly mentioned in the prompts. Interestingly, even for dimensions not directly prompted, such as *authenticity*, LLM texts are still preferred more often, though the effect size is small.

When comparing individual annotators, we also find notable differences. Some clearly favor human-

Measure	Mean	t	p (Wilcoxon)	Cohen's d
Clarity	-0.44	-7.53	**0.000**	-0.38
Authenticity	-0.12	-2.32	*0.019*	-0.12
Trustworthiness	-0.08	-1.93	0.054	-0.10
Excitement	-0.25	-5.22	**0.000**	-0.26

Table 2: Model preference scores (range: -2 to 2; 0 = no preference). Negative values indicate preference for LLM-paraphrased texts, positive values for human originals. LLM paraphrases are favored for clarity and excitement (strong direction, moderate effect size). Significance assessed using the Wilcoxon rank test.

written texts on the dimensions of *trustworthiness* and *authenticity*, while others strongly prefer the LLM-modified versions, indicating that these two dimensions are perceived more subjectively.

Qualitative remarks We conducted qualitative interviews with five participants after the annotation study. They all described the annotation task as challenging; with the exception of clarity-related ratings, they reported difficulties due to the subjective nature of most questions. They further noted lower confidence particularly for the ratings on the intermediate points of the scale, underscoring the need for a nuanced reading of the results.

Most participants made intuitive assumptions about which text in the pair was LLM-based, but they also noted a limited degree of certainty. They collectively reported a wide range of properties leading them to think a text was LLM-assisted: limited variability in lexical choices and sentence length; the amount of references and abbreviations; formatting, e.g., punctuation and bullet lists. The diversity of these strategies and their limited accuracy highlights possible interactions between personal attitudes and reading experience. A mismatch is further supported by generally negative expressed attitudes towards LLM writing (in this subgroup of participants) vs. overall positive collected ratings.

7. Conclusion

In this work, we investigated the influence of LLMs on writing style in scientific communication, specifically in the NLP domain. We relied on two data scenarios: a naturalistic corpus of over 37,000 scientific papers published in the two years preceding vs. following the release of ChatGPT, and a synthetic dataset of 3,000 human-written passages and their LLM-generated improvements.

With regard to word usage, we found that both word frequency and the contexts in which these are used have changed significantly, indicating semantic specialization in some cases and generalization in others. We also found specific stylistic features that distinguish LLM-modified texts from human texts. For example, LLM-modified texts more frequently contain certain syntactic constructions, more complex and longer words and a lower lexical diversity. Crucially, trends in word usage as well as stylistic features are broadly consistent across the naturalistic and synthetic data scenarios, indicating that many observed shifts in writing practices can be attributed to growing LLM use. Finally, in a pilot study, we measured the impact of these changes in writing style on the reading experience. The results indicate that LLM-improved texts are perceived as more understandable and exciting.

Our findings encourage further research along several dimensions. These include a larger-scale annotation study to verify the generalizability of the results; exploring prompt- or model-specific characteristics of writing style and word usage; and deploying the identified linguistic features in applied scenarios such as detecting AI-generated content.

8. Limitations

We note several limitations of our work. First, our analysis is framed as a binary comparison of language use across two time periods. This approach has important practical benefits and is underpinned by clearly stated assumptions (e.g., uncertainty regarding the precise degree of LLM use in the second time period). However, a finer-grained comparison of smaller time slices could provide a clearer picture regarding the development patterns of LLM-supported stylistic choices.

We also acknowledge that, prior to the release of ChatGPT, other AI-assisted writing tools such as Grammarly were already available and may have been used during the first time period, potentially influencing some publications. However, these tools primarily focus on grammar correction and minor stylistic suggestions rather than generating substantial new text or paraphrases. For this reason, their potential impact on the linguistic patterns we analyze is likely more limited compared to modern large language models. A larger empirical study comparing the stylistic differences between traditional writing tools and LLMs as writing assistants would be a valuable direction for future work.

More generally, our study is limited to a single scientific domain (Natural Language Processing) and one European language (English). Since academic writing conventions are highly specific to disciplines as well as languages, replicating this analysis on more diverse datasets would help assess the generalizability of the trends we report.

Finally, we ran a pilot annotation study with 20 participants, collecting two judgments per instance. A larger-scale setup with more annotations per instance would provide more robust results.

9. Acknowledgments

We are grateful to our 20 volunteer annotators for the time and effort they put into this study. Filip Miletić was supported by DFG Research Grant SCHU 2580/5-2 (*Computational Models of Semantic Variation in Multi-Word Expressions across Speakers and Languages*).

10. Bibliographical References

Mayowa Akinwande, Oluwaseyi Adeliyi, and Toyyibat Yussuph. 2024. Decoding ai and human authorship: Nuances revealed through nlp and statistical analysis. *International Journal on Cybernetics & Informatics*, 13(4):85–103.

Yutian Chen, Hao Kang, Vivian Zhai, Liangze Li, Rita Singh, and Bhiksha Raj. 2023. Token prediction as implicit classification to identify LLM-generated text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13112–13120, Singapore. Association for Computational Linguistics.

Heather Desaire, Aleesa E. Chua, Madeline Isom, Romana Jarosova, and David Hua. 2023. Distinguishing academic science writing from humans or chatgpt with over 99% accuracy using off-the-shelf machine learning tools. *Cell Reports Physical Science*, 4(6):101426.

Berin Doru, Christoph Maier, Johanna Sophie Busse, Thomas Lücke, Judith Schönhoff, Elena Enax-Krumova, Steffen Hessler, Maria Berger, and Marianne Tokic. 2025. Detecting artificial intelligence–generated versus human-written medical student essays: Semirandomized controlled study. *JMIR Medical Education*, 11:e62779.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12463–12492, Bangkok, Thailand. Association for Computational Linguistics.

Catherine A. Gao, Frederick M. Howard, Nikolay S. Markov, Emma C. Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T. Pearson. 2023. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *npj Digital Medicine*, 6(1).

Andrew Gray. 2024. ChatGPT "contamination": estimating the prevalence of LLMs in the scholarly literature. *arXiv*, abs/2403.16887.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv*, abs/2203.05794.

Chinnappa Guggilla, Budhaditya Roy, Trupti Chavan, Abdul Rahman, and Edward Bowen. 2025. AI generated text detection using instruction fine-tuned large language and transformer-based models. *arXiv*, abs/2507.05157.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. *arXiv*, abs/2301.07597.

Hassan Tarek Hakam, Robert Prill, Lisa Korte, Bruno Lovreković, Marko Ostojić, Nikolai Ramadanov, and Felix Muehlensiepen. 2024. Human-written vs AI-generated texts in orthopedic academic literature: Comparative qualitative analysis. *JMIR Formative Research*, 8:e52164.

Ahmed Abdeen Hamed and Xindong Wu. 2023. Improving detection of ChatGPT-generated fake science using real publication text: Introducing xFakeBibs a supervised-learning network algorithm. *arXiv*, abs/2308.11767.

Dmitry Kobak, Rita González-Márquez, Emőke-Ágnes Horvát, and Jan Lause. 2025. Delving into LLM-assisted writing in biomedical publications through excess vocabulary. *Science Advances*, 11(27).

Daphne Koller, Andrew Beam, Arjun Manrai, Euan Ashley, Xiaoxuan Liu, Judy Gichoya, Chris Holmes, James Zou, Noa Dagan, Tien Y. Wong, David Blumenthal, and Isaac Kohane. 2024. Why we support and encourage the use of large language models in NEJM AI submissions. *NEJM AI*, 1(1).

Bruce W. Lee and Jason Lee. 2023. LFTK: Handcrafted features in computational linguistics. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 1–19, Toronto, Canada. Association for Computational Linguistics.

Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a human-AI collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024. MAGE: Machine-generated text detection in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–53, Bangkok, Thailand. Association for Computational Linguistics.

Cong William Lin and Wu Zhu. 2025. Divergent LLM adoption and heterogeneous convergence paths in research writing. *arXiv*, abs/2504.13629.

Yongqiang Ma, Jiawei Liu, Fan Yi, Qikai Cheng, Yong Huang, Wei Lu, and Xiaozhong Liu. 2023. AI vs. human – differentiation analysis of scientific content generation. *arXiv*, abs/2301.10416.

Dominik Macko, Robert Moro, Adaku Uchendu, Jason Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko, and Maria Bielikova. 2023. MULTITuDE: Large-scale multilingual machine-generated text detection benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9960–9987, Singapore. Association for Computational Linguistics.

Maximilian Maurer. 2026. elfen: A python package for efficient linguistic feature extraction for natural language datasets. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, Rabat, Morocco. Association for Computational Linguistics.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings*, Scottsdale, Arizona, USA.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting linguistic patterns in human and LLM-generated news text. *Artificial Intelligence Review*, 57(10).

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337, Abu Dhabi, UAE. Association for Computational Linguistics.

Benedicte Pierrejean and Ludovic Tanguy. 2018. Towards qualitative word embeddings evaluation: Measuring neighbors variation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 32–39, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Paul Rayson, Damon Berridge, and Brian Francis. 2004. Extending the cochran rule for the comparison of word frequencies between corpora. In *JADT 2004 : 7es Journées internationales d'Analyse statistique des Données Textuelles*.

Paul Rayson and Roger Garside. 2000. Comparing corpora using frequency profiling. In *The Workshop on Comparing Corpora*, pages 1–6, Hong Kong, China. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Alex Reinhart, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. 2025. Do LLMs write like humans? Variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, 122(8).

Aquia Richburg, Calvin Bao, and Marine Carpuat. 2024. Automatic authorship analysis in human-AI collaborative writing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1845–1855, Torino, Italia. ELRA and ICCL.

Shaurya Rohatgi, Yanxia Qin, Benjamin Aw, Niranjana Unnithan, and Min-Yen Kan. 2023. The ACL OCL corpus: Advancing open science in

computational linguistics. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10348–10361, Singapore. Association for Computational Linguistics.

Jenna Russell, Marzena Karpinska, and Mohit Iyyer. 2025. People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5342–5373, Vienna, Austria. Association for Computational Linguistics.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 104–111, Athens, Greece. Association for Computational Linguistics.

Dominik Schlechtweg. 2023. *Human and Computational Measurement of Lexical Semantic Change*. Stuttgart, Germany.

Zhixiong Su, Yichen Wang, Herun Wan, Zhaohan Zhang, and Minnan Luo. 2025. HACo-det: A study towards fine-grained machine-generated text detection under human-AI coauthoring. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22015–22036, Vienna, Austria. Association for Computational Linguistics.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2021. Survey of computational approaches to lexical semantic change. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen, editors, *Computational Approaches to Semantic Change*, pages 1–91. Language Science Press, Berlin.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1369–1407, St. Julian's, Malta. Association for Computational Linguistics.

Yuxia Wang, Artem Shelmanov, Jonibek Mansurov, Akim Tsvigun, Nizar Habash, Alham Fikri Aji, Ekaterina Artemova, Zhuohan Xie, Jinyan Su, Rui Xing, Iryna Gurevych, and Preslav Nakov. 2025. PAN'25 generative AI detection (task 2): Human-AI collaborative text classification.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025. A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1):275–338.

Hatice Yildiz Durak, Figen Eğin, and Aytuğ Onan. 2025. A comparison of human-written versus AI-generated text in discussions at educational settings: Investigating features for ChatGPT, Gemini and BingAI. *European Journal of Education*, 60(1).

Olga Zamaraeva, Dan Flickinger, Francis Bond, and Carlos Gómez-Rodríguez. 2025. Comparing LLM-generated and human-authored news text using formal syntactic theory. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9041–9060, Vienna, Austria. Association for Computational Linguistics.

Sergio E. Zanotto and Segun Aroyehun. 2024. Human variability vs. machine consistency: A linguistic analysis of texts generated by humans and large language models. *arXiv*, abs/2412.03025.

Xuandong Zhao, Sam Gunn, Miranda Christ, Jaiden Fairoze, Andres Fabrega, Nicholas Carlini, Sanjam Garg, Sanghyun Hong, Milad Nasr, Florian Tramèr, Somesh Jha, Lei Li, Yu-Xiang Wang, and Dawn Song. 2025. SoK: Watermarking for AI-generated content. In *IEEE Symposium on Security and Privacy, SP 2025, San Francisco, CA, USA, May 12-15, 2025*, pages 2621–2639. IEEE.

A. Topic analysis

We inspect the comparability of the two time periods by analyzing their distribution across topics. We rely on BERTopic (Grootendorst, 2022): given a set of documents, it computes document embeddings using a pretrained transformer model, clusters those embeddings, and then represents the topics (corresponding to the obtained clusters) by identifying a set of distinctive keywords. We set the minimum topic size to 100 documents, representation model to Maximal Marginal Relevance, and use default values for other parameters. For each paper, we use the concatenation of its title and abstract. Papers initially classified as outliers are assigned to the best-fitting topic based on the probabilities computed in the soft-clustering step over document embeddings. We include the trained topic model in our corpus update pipeline to ensure consistency of topic labels for future papers.

For a given topic, we calculate what proportion of papers assigned to it come from t_1 vs. t_2 (after normalizing the counts by the total number of papers in the respective time period). We show sample topics with different temporal distributions in Table 3. Topics which are overrepresented in one of the two periods reflect general shifts in research trends within NLP. For example, papers focusing on individual levels of linguistic structure (e.g., word sense disambiguation, topic 41; dependency parsing, 45) and related methods (word embeddings, 19) are more frequent in t_1. Those concerned with LLMs (13, 54) and more recent methods such as reinforcement learning (62) are more prevalent in t_2. But beyond these rather intuitive differences, 39 out of 70 topics (62% of all papers) have a broadly balanced temporal distribution (normalized proportion of papers from the dominant time period $\leq 60\%$). Together with the fact that strictly all topics contain papers from both t_1 and t_2, we interpret these findings as confirming the overall comparability of the two time periods, without stark topical shifts likely to skew the outcome of our experiments.

Topic	% t_1	% t_2	Total
67_humor_humorous_offense_pun	79.8	20.2	117
51_sense_word_wsd_disambiguation	68.1	31.9	333
45_treebank_ud_universal_dependency	66.3	33.7	495
19_embeddings_similarity_word_sentence	65.5	34.5	637
61_annotation_nlp_annotators_and	52.8	47.2	342
15_question_questions_qa_answering	50.1	49.9	1,126
11_crosslingual_languages_multilingual_language	49.7	50.3	962
36_claim_claims_evidence_verification	49.3	50.7	236
43_text_machinegenerated_authorship_detection	26.4	73.6	208
13_reasoning_llms_mathematical_cot	24.3	75.7	792
54_hallucination_hallucinations_llms_detection	15.0	85.0	171
62_reward_preference_alignment_rlhf	4.6	95.4	152

Table 3: Sample topics with different temporal distribution. Percentages show the normalized proportion of papers from each time period within a topic; total shows topic size as the raw number of papers.

B. Additional Linguistic Examples

We present additional examples from our lexical analysis introduced in Section 4. We provide more detailed corpus statistics for single-word lexical choices; extend the same analysis to multi-word sequences, operationalized as word 5-grams; and then elaborate on our clustering analysis by discussing more target words and sample sentences.

B.1. Single-Word Lexical Choices

We present the strongest changes in single-word lexical choices for nouns (Table 4), adjectives (Table 5), verbs (Table 6), and adverbs (Table 7). Each table shows the 10 words with the strongest increase (top panel) and decrease (bottom panel) in frequency over time, as measured by the log-likelihood score. Examples are shown both for the naturalistic corpus (left panel) and the LLM-modified corpus (right panel). The following columns are shown:

- $Freq_{t_1}$ – frequency per million words in t_1
- $Freq_{t_2}$ – frequency per million words in t_2
- LL – log-likelihood score
- ND_{t_1} – neighborhood density in t_1
- ND_{t_2} – neighborhood density in t_2
- ΔND – change in neighborhood density in t_2
- U – Mann-Whitney–U test statistic for comparison of neighborhood densities
- p – significance levels for the Mann-Whitney–U test: ***< 0.001; **< 0.01; *< 0.05; ns ≥ 0.05

Naturalistic corpus targets marked with an asterisk also appear in the top 100 strongest changes in the LLM-modified corpus (for the respective part of speech and increase/decrease direction).

B.2. Multi-Word Lexical Choices

We perform a follow-up to our core analysis of single-word lexical choices, aiming to understand if comparable patterns can also be observed in longer sequences of words. We operationally define these as word 5-grams (lemmatized and part-of-speech tagged), retaining only those where each of the five constituent lemmas is composed of alphabetic characters only and is at least two characters long. We then collect frequency counts in t_1 and t_2 for both the naturalistic and the LLM-modified corpus, and compute the log-likelihood score using the same procedure as for individual words. We restrict our analysis to the 5-grams which appear at least 10 times in both t_1 and t_2 (for the corresponding corpus). The results are summarized in Table 8, which shows the 40 strongest rises and falls in use for the naturalistic and the LLM-modified corpus.

Some of the word sequences identified by the analysis point to the general evolution of the NLP community in terms of dominant methods (e.g., *language model such as bert* falling out of use) and writing conventions (e.g., *name or uniquely identify individual* growing in use, presumably as part of the Responsible NLP Checklist). But many other sequences clearly reflect the stylistic patterns already observed for single words. For example, the naturalistic corpus shows a decrease in meta-narrative devices relying on simple vocabulary (e.g., *we can see that the*, *result show that our model*) and an increase in more formal equivalent expressions (e.g., *it be evident that the*, *provide valuable insight into*). The LLM-modified corpus shows the same trend, with straightforward expressions falling out of use (e.g., *paper be organize as follow*, *it have be show that*), and their formal equivalents becoming more prominent (e.g., *paper be structure as follow*, *it have be observe that*). Like in the single-word analysis, we overall see that the naturalistic and the LLM-modified corpus overlap in most prominent stylistic changes, and that these generally involve more complex lexical choices. These findings further support the view that the stylistic changes in the naturalistic corpus can be at least partly attributed to LLM-assisted writing.

B.3. Clustering Examples

We provide further examples from our clustering analysis to illustrate fine-grained usage differences for the following target words: *ensure* (Table 9), *utilize* (Table 10), *crucial* (Table 11), and *notably* (Table 12). For each target word, we include two sample clusters capturing distinct uses, with four representative sentences manually selected for each cluster. For ease of reading, the examples are shown in a keyword-in-context format.

C. Human Annotation

In the human annotation, we measured four different dimensions of reading experience, each recorded with 2 items. The items for all dimensions are listed in Table 13. The annotation guidelines are presented in Figure 4 and an example of one item and the annotator view is shown in Figure 5. Annotation was conducted using the Potato annotation tool (Pei et al., 2022). The full configuration, dataset and annotation results are provided via the repository: https://github.com/FilipMiletic/ScientificCommunication

	Naturalistic corpus								LLM-modified corpus			
	Target	$Freq_{t_1}$	$Freq_{t_2}$	LL	ND_{t_1}	ND_{t_2}	ΔND	U	p	Target	$Freq_{t_1}$	$Freq_{t_2}$	LL
	prompt	169.2	910.0	56679.7	0.620	0.626	0.006	4122.0	*	study	879.5	2249.8	303.0
	llm	1.9	403.9	54465.4	0.618	0.653	0.036	1834.0	***	approach	1389.7	2565.0	173.0
	instruction	94.9	613.6	43024.1	0.567	0.575	0.008	4038.0	*	challenge	397.9	1079.6	159.5
	reasoning	223.9	593.3	17605.8	0.632	0.675	0.043	2455.5	***	research	757.1	1514.3	125.6
*	capability	83.3	339.2	16869.6	0.629	0.647	0.019	3248.0	***	advancement	12.2	216.3	108.1
	shot	251.4	559.4	12201.6	0.661	0.689	0.028	3314.5	***	realm	12.2	171.0	80.2
*	limitation	138.3	360.5	10416.0	0.601	0.586	-0.015	6022.5	*	process	518.3	991.0	73.5
*	response	391.5	722.1	10131.6	0.628	0.589	-0.039	8138.0	***	utilization	12.2	156.6	71.4
	cot	0.1	71.5	9866.6	0.756	0.726	-0.029	7505.0	***	instance	546.9	956.0	55.0
	demonstration	23.5	143.5	9714.8	0.571	0.637	0.066	785.5	***	methodology	124.5	344.1	52.2
	word	2879.9	1485.3	46279.9	0.630	0.647	0.017	3028.0	***	work	2036.5	976.6	185.7
	sentence	2502.5	1619.7	19432.9	0.600	0.618	0.018	3073.5	***	problem	765.2	288.4	109.1
	embedding	812.4	416.1	13283.1	0.661	0.664	0.002	4901.5	ns	way	477.5	191.6	61.5
	network	526.5	255.8	9758.4	0.677	0.676	-0.001	5185.0	ns	kind	106.1	12.4	41.4
	representation	1190.7	764.5	9550.2	0.617	0.654	0.037	2025.0	***	addition	297.9	138.0	29.3
	vector	577.0	297.3	9292.2	0.685	0.694	0.009	4240.5	ns	setting	457.1	265.8	25.0
	corpus	706.2	398.1	8887.6	0.645	0.640	-0.005	5992.5	*	idea	167.3	61.8	24.6
	attention	767.4	447.1	8721.4	0.644	0.639	-0.005	5114.0	ns	respect	95.9	24.7	21.9
	layer	942.3	608.9	7366.7	0.695	0.703	0.008	4291.5	ns	intuition	38.8	2.1	19.6
	feature	1043.3	696.8	7084.2	0.603	0.636	0.034	1606.5	***	amount	279.6	152.5	18.5

Table 4: Strongest changes in **noun** usage.

	Naturalistic corpus								LLM-modified corpus			
	Target	$Freq_{t_1}$	$Freq_{t_2}$	LL	ND_{t_1}	ND_{t_2}	ΔND	U	p	Target	$Freq_{t_1}$	$Freq_{t_2}$	LL
	prompt	55.9	268.0	15324.2	0.603	0.608	0.005	3950.0	*	various	461.2	1473.1	271.7
	llm	0.6	61.4	8023.0	0.617	0.605	-0.012	7262.5	***	significant	212.2	684.0	127.6
*	comprehensive	51.5	175.7	7298.5	0.606	0.634	0.028	2586.0	***	distinct	65.3	304.9	82.2
*	diverse	130.0	255.5	4232.5	0.584	0.598	0.014	3564.0	***	comprehensive	104.1	379.1	81.2
*	various	304.9	483.4	4148.1	0.602	0.600	-0.002	5207.0	ns	primary	55.1	276.1	78.6
*	significant	258.7	414.8	3710.9	0.590	0.587	-0.003	5050.5	ns	crucial	112.2	387.3	78.2
	blank	11.2	57.6	3475.3	0.581	0.530	-0.051	8354.5	***	specific	834.6	1403.0	71.2
	demographic	28.8	86.8	3094.5	0.642	0.613	-0.028	7095.5	***	valuable	42.9	206.0	56.8
*	potential	108.8	203.2	2949.4	0.606	0.615	0.009	4021.5	*	superior	16.3	140.1	54.8
*	crucial	70.0	147.4	2869.0	0.618	0.600	-0.018	6768.5	***	notable	34.7	179.2	52.2
	neural	427.0	175.5	11057.2	0.666	0.657	-0.010	5946.0	*	well	363.2	109.2	70.2
	syntactic	265.0	127.0	5068.2	0.691	0.700	0.010	4106.5	*	above	146.9	14.4	61.6
	lstm	60.8	12.6	3525.7	0.737	0.690	-0.047	8028.5	***	different	1669.2	1083.7	61.2
	bert	146.8	63.8	3434.0	0.753	0.762	0.009	3733.5	**	good	542.8	241.0	58.1
	unsupervised	167.7	80.3	3213.0	0.639	0.641	0.002	4489.0	ns	able	208.1	43.3	57.3
	lexical	260.6	155.3	2744.9	0.642	0.639	-0.003	5406.5	ns	many	673.4	362.6	46.2
	semantic	613.7	461.9	2192.3	0.648	0.667	0.019	3263.5	***	hard	175.5	47.4	38.2
	morphological	107.9	50.5	2171.7	0.694	0.708	0.014	3817.0	**	useful	167.3	45.3	36.3
*	same	966.0	783.6	1943.4	0.507	0.520	0.013	2925.5	***	possible	351.0	158.6	36.3
*	good	743.9	587.2	1885.7	0.563	0.546	-0.018	6990.0	***	particular	291.8	127.7	32.1

Table 5: Strongest changes in **adjective** usage.

	Naturalistic corpus								LLM-modified corpus			
	Target	$Freq_{t_1}$	$Freq_{t_2}$	LL	ND_{t_1}	ND_{t_2}	ΔND	U	p	Target	$Freq_{t_1}$	$Freq_{t_2}$	LL
	prompt	33.1	303.0	25305.3	0.612	0.665	0.053	1399.5	***	utilize	257.1	1905.7	693.9
*	enhance	126.9	403.2	15384.0	0.643	0.636	-0.007	6045.5	*	enhance	197.9	1306.2	446.2
	generate	1248.9	1752.5	8640.3	0.605	0.598	-0.007	5817.5	*	involve	240.8	1291.8	386.6
*	utilize	222.1	454.0	8254.0	0.631	0.605	-0.026	7447.5	***	address	338.7	1112.5	212.2
*	exhibit	63.3	202.8	7813.6	0.568	0.569	0.001	5094.0	ns	introduce	691.8	1633.8	191.7
*	employ	218.3	435.5	7476.5	0.586	0.573	-0.014	6577.5	***	assess	122.4	587.2	161.6
*	ensure	139.9	321.5	7460.6	0.536	0.549	0.013	3791.5	**	employ	222.4	782.9	161.6
*	assess	119.0	276.9	6582.3	0.617	0.614	-0.003	5176.5	ns	demonstrate	351.0	953.9	141.3
*	demonstrate	313.2	540.9	6250.3	0.624	0.616	-0.008	6262.0	**	encompass	14.3	276.1	141.0
	underscore	2.5	53.5	5819.9	0.598	0.598	0.001	5082.5	ns	incorporate	242.8	729.3	124.4
*	learn	1037.9	719.1	5934.0	0.581	0.583	0.002	4519.0	ns	show	1552.9	661.3	180.0
	train	1838.0	1415.0	5624.7	0.610	0.631	0.021	2670.5	***	use	4395.5	2933.8	143.1
	embed	603.9	380.0	5245.9	0.662	0.673	0.010	4053.0	*	have	1283.6	665.5	97.2
*	use	5279.8	4674.5	3754.2	0.545	0.481	-0.064	8634.5	***	give	1169.3	593.4	93.4
	encode	253.5	151.4	2654.2	0.627	0.648	0.021	2854.5	***	propose	1763.1	1100.2	75.5
	predict	682.2	517.2	2322.6	0.570	0.578	0.009	3760.0	**	perform	640.8	272.0	74.7
	hide	147.7	79.1	2149.9	0.661	0.656	-0.005	5378.0	ns	obtain	518.3	206.0	67.8
*	model	232.8	146.6	2014.8	0.612	0.613	0.001	4611.0	ns	get	167.3	18.5	66.7
*	describe	519.1	386.4	1989.3	0.610	0.609	-0.001	4886.5	ns	ignore	120.4	8.2	57.0
*	have	1624.6	1391.3	1840.6	0.505	0.489	-0.016	6814.0	***	finetune	120.4	12.4	49.6

Table 6: Strongest changes in **verb** usage.

	Naturalistic corpus								LLM-modified corpus			
	Target	$Freq_{t_1}$	$Freq_{t_2}$	LL	ND_{t_1}	ND_{t_2}	ΔND	U	p	Target	$Freq_{t_1}$	$Freq_{t_2}$	LL
*	additionally	137.2	281.0	5136.8	0.605	0.620	0.015	4576.5	ns	additionally	140.8	815.9	257.3
*	effectively	98.9	209.3	4114.3	0.656	0.637	-0.019	7286.0	***	initially	12.2	255.5	132.7
*	notably	33.8	100.6	3524.0	0.600	0.629	0.029	3499.5	***	effectively	126.5	488.3	110.9
*	thereby	30.0	90.3	3206.5	0.564	0.561	-0.003	5295.5	ns	notably	36.7	269.9	97.8
*	subsequently	27.6	82.5	2915.3	0.574	0.568	-0.006	5861.5	*	subsequently	53.1	309.0	97.8
*	primarily	40.0	98.7	2602.8	0.586	0.573	-0.014	6504.5	***	primarily	57.1	274.0	75.4
*	fine	353.2	488.9	2236.6	0.660	0.655	-0.005	5173.0	ns	solely	38.8	224.6	70.8
*	accurately	41.5	94.8	2179.0	0.618	0.608	-0.011	6059.5	**	particularly	104.1	329.6	60.2
*	specifically	275.5	390.1	2018.9	0.557	0.564	0.007	4245.0	ns	furthermore	159.2	418.2	58.8
*	solely	30.4	73.9	1897.8	0.562	0.563	0.001	4791.0	ns	accurately	49.0	204.0	49.8
*	very	334.8	233.5	1853.2	0.628	0.623	-0.005	5329.0	ns	thus	487.7	98.9	136.9
*	jointly	91.5	46.5	1527.9	0.604	0.617	0.014	3152.5	***	first	540.8	177.2	94.0
*	also	1752.6	1542.8	1362.0	0.573	0.581	0.008	3816.0	**	usually	236.7	33.0	84.5
*	usually	161.1	106.7	1137.4	0.618	0.622	0.004	4429.0	ns	also	1540.7	898.3	83.5
*	well	981.1	845.2	1031.8	0.535	0.525	-0.010	6211.0	**	much	165.3	22.7	59.5
*	thus	451.1	364.2	945.7	0.624	0.620	-0.005	5090.5	ns	very	230.6	57.7	54.1
*	much	206.9	154.5	779.0	0.630	0.626	-0.003	5330.0	ns	finally	257.1	74.2	52.1
	automatically	173.8	127.9	713.6	0.626	0.618	-0.008	6221.5	**	far	487.7	236.9	43.2
*	then	942.1	834.0	671.6	0.586	0.591	0.005	4287.5	ns	only	1106.0	756.1	32.2
*	recently	152.1	112.6	605.2	0.555	0.554	-0.001	5316.5	ns	mainly	151.0	45.3	29.2

Table 7: Strongest changes in **adverb** usage.

	Naturalistic corpus				LLM-modified corpus			
	Target	t_1	t_2	LL	Target	t_1	t_2	LL
	consistent with their intend use	0.1	18.5	2440.5	be important to note that	8.2	72.1	28.6
	be consistent with their intend	0.1	18.3	2410.8	it be important to note	10.2	72.1	25.6
	identify individual people or offensive	0.1	17.8	2369.2	important to note that the	2.0	24.7	11.1
	information that name or uniquely	0.1	17.7	2365.9	it be worth note that	8.2	39.1	10.8
	that name or uniquely identify	0.1	17.7	2365.9	can be find in appendix	2.0	20.6	8.6
	or uniquely identify individual people	0.1	17.8	2364.7	paper be structure as follow	10.2	37.1	7.9
	uniquely identify individual people or	0.1	17.8	2364.7	to enhance the performance of	6.1	28.8	7.8
	name or uniquely identify individual	0.1	17.8	2362.0	the field of natural language	2.0	18.5	7.4
	individual people or offensive content	0.1	17.9	2357.6	lead to the development of	2.0	16.5	6.3
	the datum that be collect	0.1	17.4	2325.2	in the context of the	10.2	33.0	6.2
	that be compatible with the	0.4	17.5	2155.3	field of natural language processing	2.0	14.4	5.1
	report the number of parameter	0.2	15.8	2060.9	the key contribution of this	2.0	14.4	5.1
	of parameter in the model	0.3	16.0	1954.4	to address the issue of	4.1	18.5	4.9
	the number of example in	0.9	18.1	1945.2	can be summarize as follow	14.3	35.0	4.4
	the source of the datum	0.2	14.6	1847.9	the paper be structure as	8.2	24.7	4.3
	be the source of the	0.4	14.5	1738.0	the extent to which the	2.0	12.4	4.0
	number of parameter in the	1.0	16.7	1702.7	demonstrate that our model significantly	2.0	12.4	4.0
	the number of parameter in	2.0	17.5	1435.6	can be categorize into two	2.0	12.4	4.0
*	can be find in appendix	14.1	34.6	906.0	provide an overview of the	4.1	16.5	3.9
*	be important to note that	6.2	20.7	837.4	in the case of the	4.1	16.5	3.9
*	it be important to note	6.8	21.8	837.3	it have be observe that	4.1	14.4	3.0
	capability of large language model	0.1	4.4	530.5	that our model significantly outperform	4.1	14.4	3.0
	it be important to acknowledge	0.2	3.5	373.8	can be find in table	4.1	14.4	3.0
	provide valuable insight into the	0.1	3.0	318.8	result demonstrate that our model	2.0	10.3	2.9
*	it be worth note that	13.6	24.3	310.5	this paper be structure as	2.0	10.3	2.9
*	to enhance the performance of	1.3	5.0	225.9	to assess the quality of	2.0	10.3	2.9
*	can be attribute to the	6.0	12.3	218.8	experimental result demonstrate that our	2.0	10.3	2.9
	be important to acknowledge that	0.1	2.0	215.7	in the field of natural	2.0	10.3	2.9
*	to assess the effectiveness of	0.7	3.6	208.9	between the source and target	2.0	10.3	2.9
*	important to note that the	1.8	5.6	203.7	to mitigate the impact of	2.0	10.3	2.9
	grant fund by the korea	1.0	4.1	202.0	to assess the effectiveness of	2.0	10.3	2.9
	particularly in the context of	0.1	2.1	201.2	in order to address the	2.0	10.3	2.9
	natural science foundation of china	11.9	19.8	199.7	have show promising result in	2.0	10.3	2.9
	ability of large language model	0.1	2.0	194.8	enhance the performance of the	2.0	10.3	2.9
	it be evident that the	1.0	3.8	176.5	can be attribute to the	4.1	12.4	2.1
	prompt the model to generate	0.1	1.8	174.0	contribution can be summarize as	12.2	24.7	2.1
	national natural science foundation of	11.0	17.9	171.3	be worth note that the	2.0	8.2	2.0
	may be attribute to the	1.0	3.7	170.9	be commonly refer to as	2.0	8.2	2.0
	we use the same prompt	0.1	1.7	168.8	model be train use the	2.0	8.2	2.0
	do not align with the	0.3	2.2	152.9	result demonstrate that our propose	2.0	8.2	2.0
	the state of the art	11.4	3.9	390.6	the paper be organize as	22.4	2.1	9.7
*	paper be organize as follow	11.0	4.8	255.2	paper be organize as follow	30.6	6.2	8.6
	acknowledgment we would like to	6.9	2.6	205.0	have be show to be	20.4	2.1	8.5
	the encoder and the decoder	4.8	1.4	193.5	it have be show that	22.4	4.1	6.8
	we would like to thank	22.6	14.5	180.1	the good of our knowledge	114.3	68.0	5.8
	and the anonymous reviewer for	5.2	1.8	179.2	to the good of our	114.3	70.0	5.2
*	result show that our model	4.4	1.4	163.4	of the paper be organize	14.3	2.1	5.0
	the anonymous reviewer for their	24.4	16.4	161.1	this work be as follow	14.3	2.1	5.0
	in this paper we present	4.1	1.3	155.9	contribution of this work be	24.5	8.2	4.1
*	experimental result show that our	9.2	4.6	155.7	contribution of our work be	12.2	2.1	3.9
*	the paper be organize as	6.4	2.8	155.5	of this work be as	12.2	2.1	3.9
	our model be able to	4.3	1.4	152.8	the contribution of this paper	12.2	2.1	3.9
	we evaluate our model on	6.6	2.9	150.0	contribution be summarize as follow	16.3	4.1	3.8
	bidirectional encoder representations from transformers	5.7	2.3	149.1	contribution of this paper be	24.5	10.3	2.9
	the source and the target	3.9	1.2	147.3	result show that our model	14.3	4.1	2.9
	in this paper we propose	2.8	0.7	137.6	from the point of view	10.2	2.1	2.9
	rest of the paper be	5.6	2.4	133.8	the point of view of	10.2	2.1	2.9
	we use the adam optimizer	6.0	2.7	130.5	which be base on the	10.2	2.1	2.9
	state of the art in	4.2	1.6	126.4	it be not clear how	10.2	2.1	2.9
	language model such as bert	5.2	2.2	125.1	that can be use to	16.3	6.2	2.3
	the rest of the paper	8.0	4.1	124.8	main contribution of our work	12.2	4.1	2.1
	state of the art result	1.8	0.3	124.3	be one of the most	12.2	4.1	2.1
	we can see that the	15.8	10.3	119.8	we conduct experiment on the	12.2	4.1	2.1
	and future work in this	6.8	3.4	119.1	on the performance of the	12.2	4.1	2.1
*	of the paper be organize	4.6	1.9	118.4	contribution in this paper be	8.2	2.1	1.9
	of the word in the	5.1	2.2	118.1	we make the following contribution	8.2	2.1	1.9
	new state of the art	3.0	0.9	118.0	experimental result demonstrate the effectiveness	8.2	2.1	1.9
	on the basis of the	5.6	2.7	112.0	an answer to the question	8.2	2.1	1.9
	future work in this paper	5.2	2.4	108.0	for the purpose of this	8.2	2.1	1.9
	reviewer for their helpful comment	6.1	3.1	104.2	this be the first attempt	8.2	2.1	1.9
	thank the anonymous reviewer for	17.7	12.2	101.9	be the first attempt to	8.2	2.1	1.9
	state of the art on	3.1	1.1	99.5	we be the first to	34.7	20.6	1.8
	the hidden state of the	6.5	3.4	98.5	the main contribution of our	14.3	6.2	1.6
*	this paper be organize as	4.3	1.9	98.0	experimental result show that our	10.2	4.1	1.3
	we compare our model with	5.6	2.8	96.0	to encourage the model to	10.2	4.1	1.3
	show that our model outperform	2.5	0.8	96.0	similar to the one use	10.2	4.1	1.3
	anonymous reviewer for their helpful	7.6	4.3	95.6	evaluate the performance of the	10.2	4.1	1.3
	would like to thank the	13.3	8.8	94.3	main contribution of this work	10.2	4.1	1.3
	acknowledgment we thank the anonymous	2.7	0.9	92.5	the main contribution of this	24.5	14.4	1.3
	for each word in the	3.4	1.4	87.8	effectiveness of the propose approach	6.1	2.1	1.0

Table 8: Strongest changes in the use of word 5-grams. Shown columns: t_1: frequency in t_1 per million words; t_2: frequency in t_2 per million words; LL: log-likelihood score.

measure the semantic similarity between generated questions to	**ensure**	that the questions assess the same content .
We add an adversarial objective to	**ensure**	that the model focuses on language - agnostic
sampling process , we store this data to	**ensure**	that every model processes exactly the same set
we diversify the output of each task to	**ensure**	that the model can provide a variety of
To	**ensure**	the high quality of the annotation procedure ,
To	**ensure**	the quality of data , they also calculated
Finally , to	**ensure**	training stability and prevent overfitting , we modify
the hyperparameters the same for different models to	**ensure**	the fairness of the experiment .

Table 9: Example uses of the verb *ensure*. Top cluster: broad sense of finality similar to conjunctions like *in order to* (65% t_1 vs. 35% t_2; 431 sentences). Bottom cluster: more specific sense 'to guarantee' (43% t_1 vs. 57% t_2; 363 sentences). The growing use of the more specialized second sense is consistent with an increasing neighborhood density ($\Delta ND = 0.013$).

The latter aims to	**utilize**	the multi - level interests to enhance both
of data augmentation method and proposed methods to	**utilize**	the augmented data in more effective ways (
introduced pivot - based transfer learning techniques to	**utilize**	the resources of the pivot language .
We first transform emoticons into textual information to	**utilize**	their rich emotional information .
Finally , we	**utilize**	a ridge regression classifier to obtain final classification
For both tasks , we	**utilize**	a publicly available implementation that has been trained
In the second setting , we	**utilize**	the MLM training objective inherited from PLMs to
Subsequently , we	**utilize**	a language model equipped with adapters to obtain

Table 10: Example uses of the verb *utilize*. Top cluster: more specific sense 'use to the fullest potential' (60% t_1 vs. 40% t_2; 281 sentences). Bottom cluster: more general sense 'make use of' (44% t_1 vs. 56% t_2; 354 sentences). The growing use of the broader second sense is consistent with a falling neighborhood density ($\Delta ND = -0.026$).

, word order difference is one of the	**crucial**	factors that impact cross - lingual transfer (
Thus , sarcasm detection might be the first	**crucial**	step in these systems .
Apart from these	**crucial**	extensions we made for providing the use of
of styles used makes measurements , a most	**crucial**	aspect of scientific writing , challenging to extract
Therefore , it is	**crucial**	to understand the speaker 's intentions and emotions
of the inherent individual differences , it is	**crucial**	to incorporate diversity in the data collection process
the annotation tasks automatically , we deem it	**crucial**	to expand our proposed framework to new scenarios
These challenges make it	**crucial**	to devise a new framework that can work

Table 11: Example uses of the adjective *crucial*. Top cluster: finality-connoted sense 'important in determining an outcome' (58% t_1 vs. 42% t_2; 146 sentences). Bottom cluster: more general sense 'important, significant' (38% t_1 vs. 62% t_2; 208 sentences). The growing use of the broader second sense is consistent with a falling neighborhood density ($\Delta ND = -0.018$).

a mixture of content and structural properties ,	**notably**	the systems in the 2016 document alignment shared
, the model shows diminished retrieval performance ,	**notably**	in terms of R@1 .
had been dominated by the statistical methods ,	**notably**	the phrase - based (Koehn et al
switching corpora in comparison with monolingual corpora ,	**notably**	for high - resource languages , e.g. ,
predictions for the labels of other pairs is	**notably**	lower .
and GPT - labels , however , is	**notably**	larger than the other models , with a
and XLM - RoBERTa - Large models are	**notably**	improved in both the full and partial benchmarks
DSP on most datasets , where it is	**notably**	more efficient to train .

Table 12: Example uses of the adverb *notably*. Top cluster: semantically broad usage typically introducing an example, similar to 'especially, particularly' (63% t_1 vs. 37% t_2; 171 sentences). Bottom cluster: intensifier-like sense 'very' (44% t_1 vs. 56% t_2; 309 sentences). The intensification use is restricted to a relatively small set of cooccurrents (typically gradable adjectives), which aligns with an increasing neighborhood density ($\Delta ND = 0.029$).

Dimension	Question
Clarity	I read this text smoothly and fluently. I was able to understand the paragraph without difficulty.
Authenticity	I felt a sense of connection with the researchers through their writing. The authors are genuinely engaged with the topic they studied.
Trustworthiness	The authors appeared competent and trustworthy. I remained critical of the authors' arguments while reading.
Excitement	I am excited to read this paper in more depth. I found the text enjoyable to read.

Table 13: The following questions are used to measure four different dimensions of reading experience.

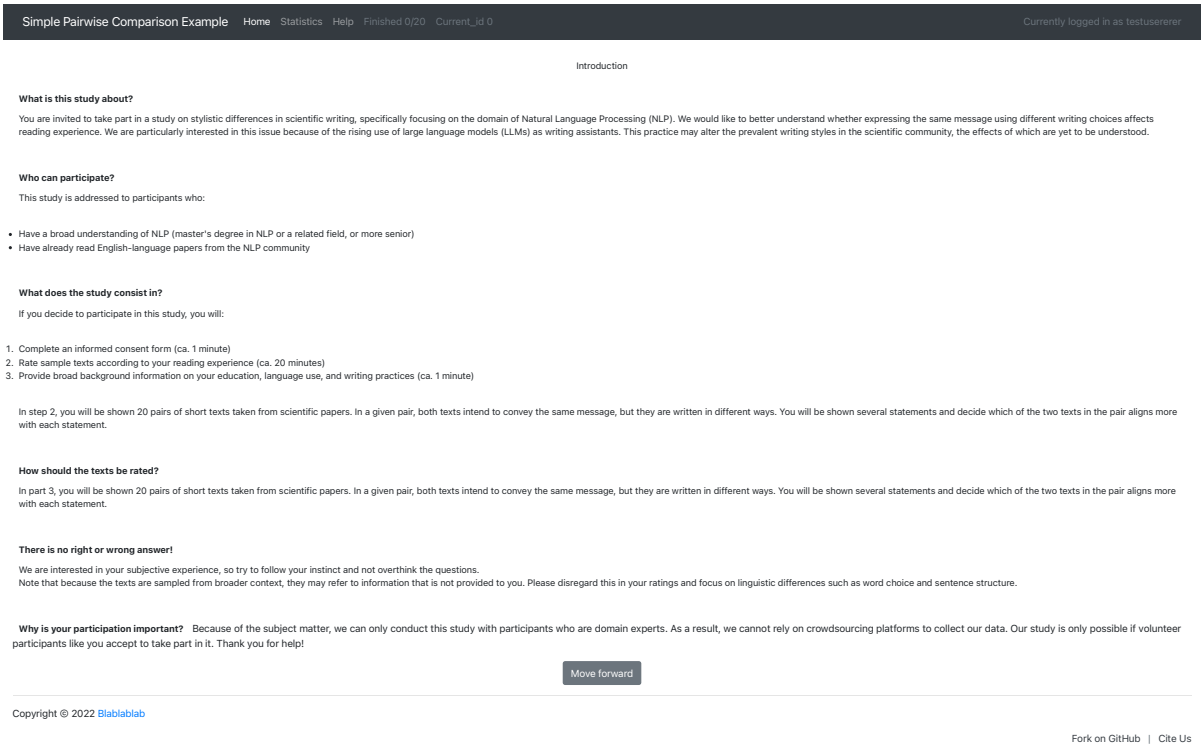

Figure 4: The introduction and study description annotators saw before rating the pairs of human vs. LLM-paraphrased scientific text.

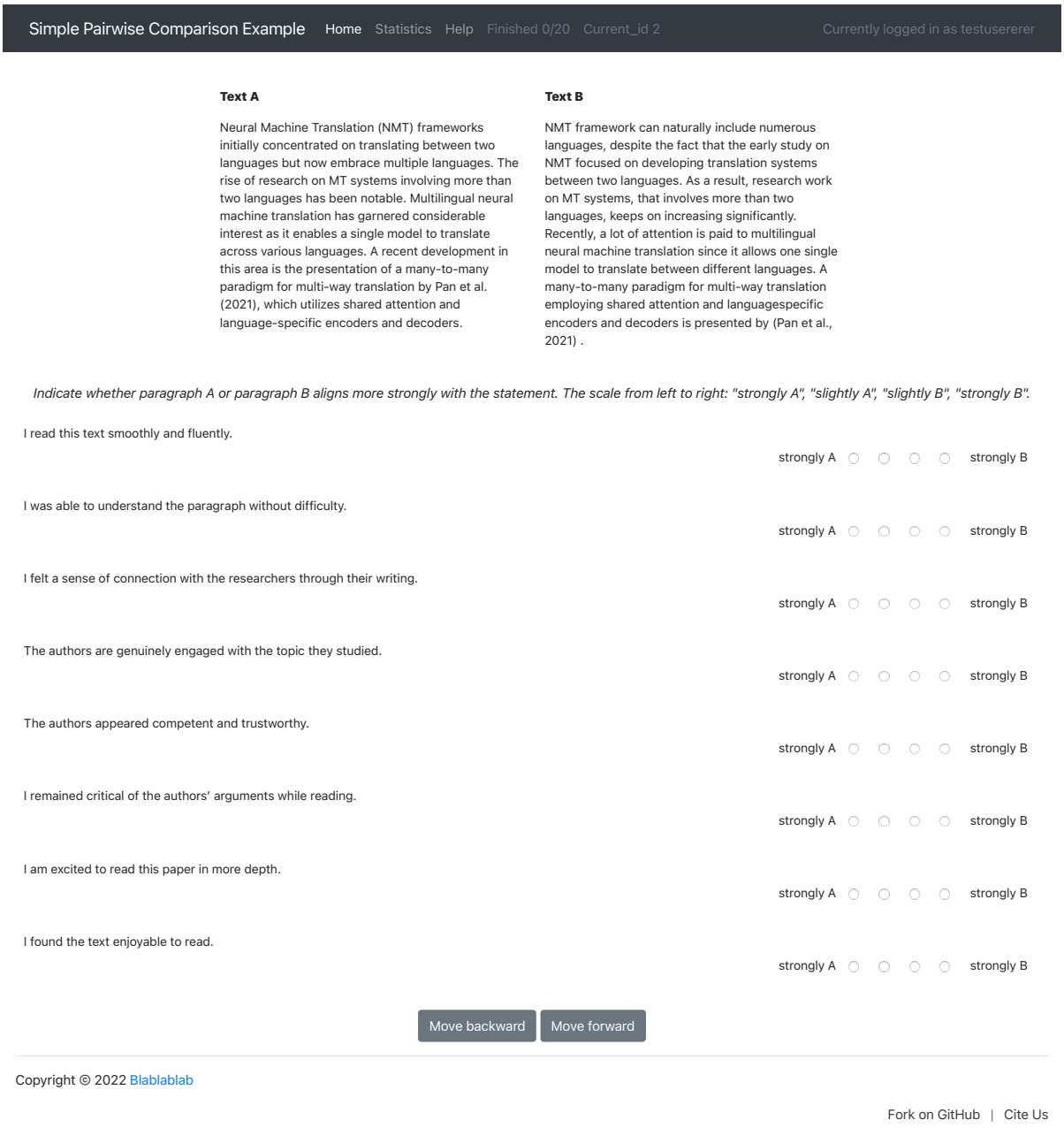

Figure 5: View of an example annotation item: pairwise annotation on four different dimensions of reading experience.