Why Does ChatGPT “Delve” So Much? Exploring the Sources of Lexical Overrepresentation in Large Language Models

Abstract

Scientific English is currently undergoing rapid change, with words like “delve,” “intricate,” and “underscore” appearing far more frequently in scientific papers than just a few years ago. It has been widely assumed that scientists’ use of large language models (LLMs) is responsible for these trends. We use a novel method to identify 21 focal words whose increased occurrence in recent scientific abstracts is likely the result of LLM usage. We then pose “the puzzle of lexical overrepresentation”: *why* are such words overused by LLMs? We look for but fail to find evidence that lexical overrepresentation is caused by model architecture, algorithm choices, or overrepresentation of the focal words in training data. To assess whether reinforcement learning from human feedback (RLHF) contributes to the overuse of focal words, we undertake comparative model testing and conduct an exploratory online study. Our results, though not conclusive, are consistent with RLHF playing a role. With LLMs quickly becoming a driver of global language change, investigating these potential sources of lexical overrepresentation is important. And yet, we note that lack of transparency surrounding model development remains an obstacle to such research.

1. Introduction

Like all human language, Scientific English has changed substantially over time \cite{degaetano2018using, degaetano2018information, bizzoni2020linguistic}. New discoveries have fueled (and in some cases been fueled by) the introduction of new lexical items into scientific discourse. For example, the terms ``electricity'' and ``electrify'' spiked in Scientific English during the 18th century \cite{degaetano2018using}. Changes in dominant methodological and explanatory frameworks – such as the rise of mechanical philosophy, or the mathematization of many scientific fields – have been accompanied by changes in word usage and syntactic structures as well. The nineteenth century, for instance, saw a rise in the use of certain nominal patterns with prepositions \cite{degaetano2018using}, as well as a general reduction in complexity of syntactic structures \cite{krielke2024cross}. And of course, such changes continue through the present. There is some evidence, for example, that the use of the passive voice in Scientific English has been declining in the last few decades in favor of active verbs with first person pronoun subjects \cite{banks2017extent}, although the time of change is under discussion \cite{leong2020passive}.

Over the last two years, however, Scientific English has witnessed increasing usage of certain lexical items at a seemingly unprecedented pace. As was first pointed out in discussions on social media (e.g.,\ \citealt{Koppenburg2024,Nguyen2024}), and then investigated more systematically by researchers \cite{cheng2024have,gray2024chatgpt,kobak2024delving,liang2024mapping,liu2024towards,matsui2024delving}, words such as ``delve,'' ``intricate,'' and ``nuanced'' have appeared far more frequently in scientific abstracts from 2023 and 2024 compared to earlier years. Unlike many previous changes in Scientific English, these trends do not seem to be explained by changes in the content of science or in wider language use. It is therefore widely assumed that the reason for the sharp increase is the use of large language models (LLMs) like ChatGPT for scientific writing. Evidence supporting this hunch has emerged in recent months (e.g.,\ \citealt{cheng2024have,liang2024monitoring}).

The goals of the present research were twofold. First, we aimed to provide a systematic characterization of this linguistic phenomenon. Some existing work has relied on informal methods, such as reading online discussion boards, to identify the words anecdotally observed to occur more frequently in AI-generated writing (e.g.,\ \citealt{matsui2024delving}). We developed a method for extracting lexical items of interest, described in Section~\ref{sec:corpusanalysis}. First we generated a list of non-content-related words that are overrepresented in recent scientific abstracts compared to abstracts from a few years ago, before the advent of ChatGPT, the most popular LLM \cite{Sarkar2023}. Then we compared these words to lexical items that LLMs use more frequently than humans when writing scientific abstracts. This produced a list of 21 ``focal words'':\ lexical items that (i) have recently spiked in Scientific English and (ii) seem to be (over)used by LLMs. Our method for generating such a list is rigorous, reproducible, and transferable to other data and models.

Prior research has largely focused on quantifying such focal words' increasing usage and leveraging their prevalence to estimate how much recent scientific writing has been produced with the assistance of LLMs (e.g.,\ \citealt{kobak2024delving,liang2024mapping}). By contrast, our research focused on the question of ``why'' rather than ``how much''. Our second goal was to explore the factors that might contribute to the overrepresentation of the focal words in scientific writing produced by LLMs. Why does ChatGPT use ``delve'' (and other focal words) so frequently when generating scientific text? We identified a set of possible factors, characterized in Section~\ref{sec:puzzle}, and then sought to rule several of them out. We did not find evidence that certain model choices or algorithmic decisions play a major role in the overrepresentation of focal words (Section 4). Large language models are trained in several stages. During pre-training, and through exposure to billions (or trillions) of documents, models learn to predict the next word in a sequence of text. At a later stage, models are often fine-tuned for specific purposes, which is done with additional data. We also did not find any evidence that the overrepresentation of focal words stems from the training or fine-tuning data (Section 5).

In a later phase of training, LLMs are given additional information about quality outputs from human evaluators. For instance, the system might be required to generate several responses to the same query, and a human evaluates which is the best. The model is then subject to reinforcement learning based on the human evaluations. We found preliminary evidence that reinforcement learning from human feedback (RLHF) contributes to the overrepresentation of certain lexical items in LLM-generated text. This evidence comes primarily from model testing on Meta's Llama LLM. An exploratory experiment described in Section 6 is also weakly suggestive, although our findings indicate that participants in the experiment became wary of the word ``delve'' in particular. This result raises important questions about the future of LLM-driven language change, discussed in Section 7.

2. Corpus Analysis: Systematic Identification of Overrepresented Lexical Items

To probe recent changes in language usage in scientific abstracts, we used PubMed's publicly available repository of scientific abstracts, which focuses on biomedical literature \cite{PubMed}. We downloaded the database through the PubMed API using a script written in Python \cite{Python3} and extracted and analyzed all abstracts available at the time of download, on May 4, 2024 (all code can be found on our GitHub). Our analysis includes more than 5.2~billion tokens from the abstracts of 26.7 million papers. To track changes in word usage over time, we measured occurrences per million (opm) of a given token in each year. Figure~\ref{fig:baselineopms} illustrates the usage trajectories of some baseline items over time. We focus on the period from 1975 to May 2024 as data prior to 1975 are less extensive.

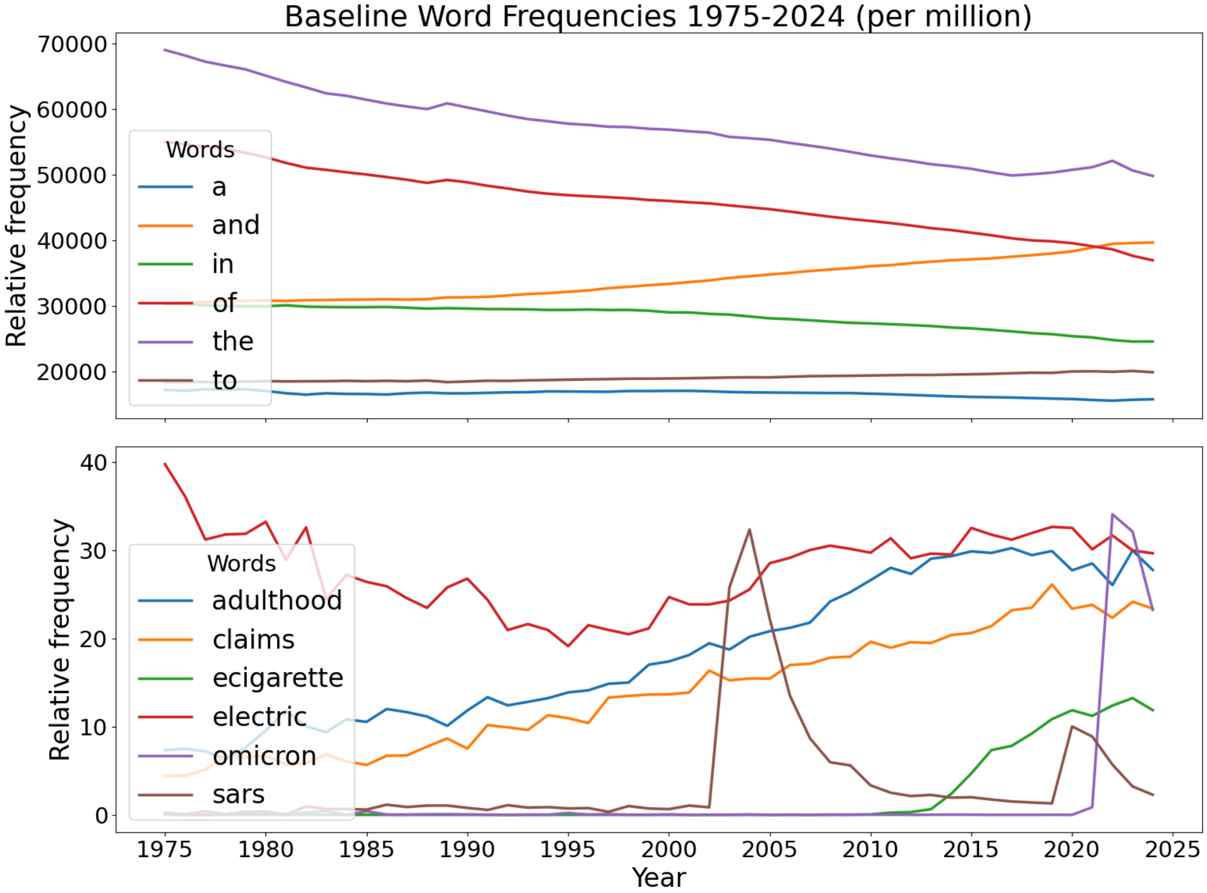


FIGURE 1a (top) and 1b (bottom): \caption{Occurrences per million words in PubMed for selected lexical entries.}

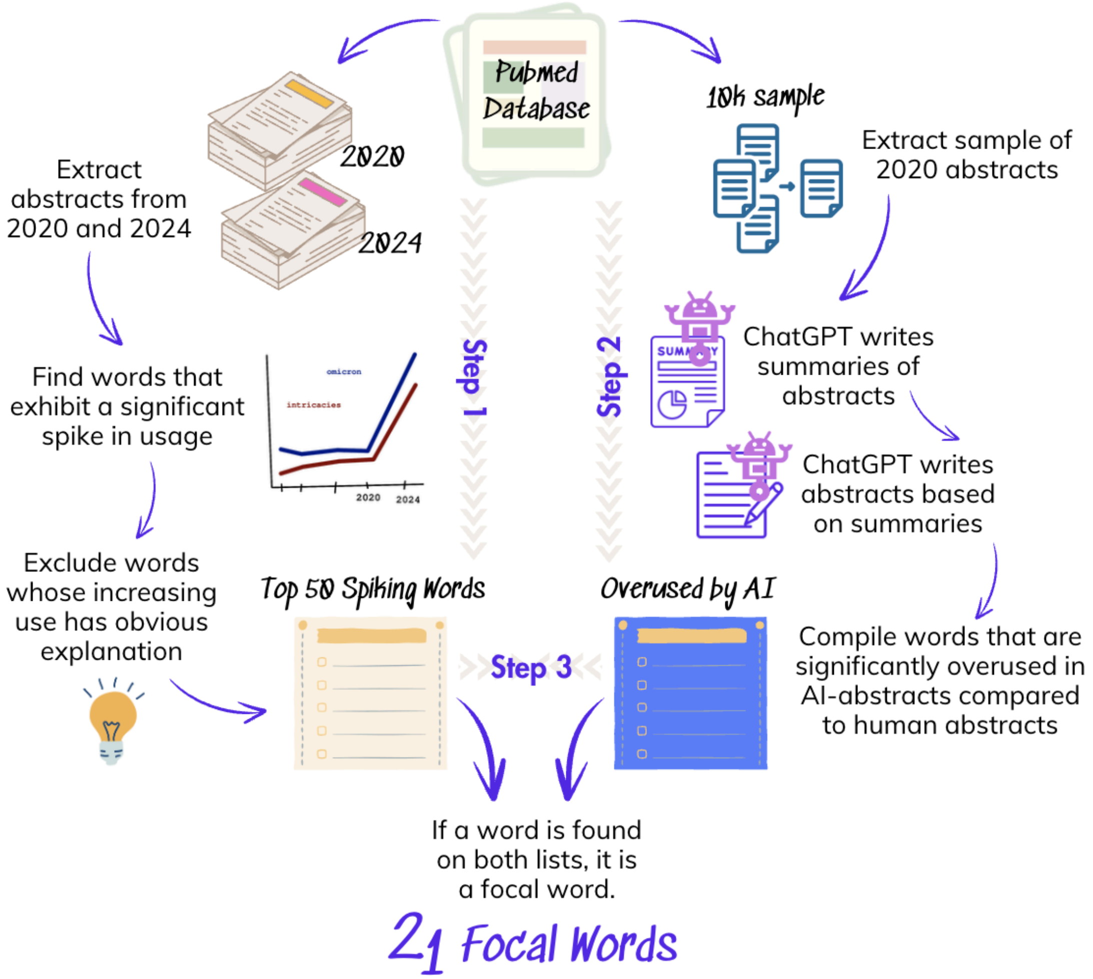
The goal of our corpus analysis was to develop a more rigorous approach to identifying words in recent scientific writing whose overuse is likely to be the result of LLM deployment. Our approach involved three steps. First, we sought to determine which words were more prevalent in scientific abstracts from 2024 compared to 2020. (Since LLMs were not in widespread use in

2020, we were confident that most or all 2020 abstracts were human-produced.) We calculated the percentage increase in opm for each token in the database between 2020 and 2024. Unsurprisingly, for many of the words which spiked in usage during that time, there was a straightforward explanation. For example, ``omicron'' and ``metaverse'' were two of the words that showed the largest percentage increase (see Figure~\ref{fig:baselineopms}). We only considered increases deemed significant by chi-square tests. (There were about 7300 tokens that exhibited a significant spike.)

For the purpose of this study, we were interested in isolating the words whose spike in usage was unexplained. The authors functioned as annotators and independently reviewed the list of words that had the highest percentage change in opm to exclude irrelevant tokens (like year numbers) and words whose increasing usage had a straightforward explanation in terms of scientific advancements or world events. In cases of disagreement, we included the word on our list. This process yielded a list of 50 words whose usage spiked without any obvious explanation (see name.tsv in our GitHub repo). The resulting list contained several of the words that had been the focus of online conversation, including ``delve'' and ``intricate''.

However, a spike without an obvious real-world explanation is not necessarily an LLM-induced spike. For example, the usage of 'mash' increased tenfold, but it is not a word that ChatGPT is known to overuse. Thus, the second step called for identifying words that are overrepresented in AI-generated scientific abstracts compared to human-generated abstracts. Producing AI-generated abstracts for this comparison was a delicate process; as frequent users of LLMs can attest, the outputs of these models are highly sensitive to the formulation of the prompt \cite{wei2022chain,zhou2022least}.Our aim was to imitate as closely as possible the process by which researchers might have deployed an LLM in 2022-early 2024. After exploring several different methods of abstract generation, we ended up with a two-stage process:\ (1) We randomly sampled 10,000 abstracts from papers published in 2020 from the PubMed database., We then used a Python script to have ChatGPT-3.5 summarize the associated paper (Prompt:\ ``The following is an abstract of an article. Summarize it in a couple of sentences.'' Code available on GitHub.) (2) The ChatGPT-generated summary was input via the OpenAI API, and a Python script that used the ChatGPT-3.5 API produced a corresponding scientific abstract. (Prompt:\ ``Please write an abstract for a scientific paper, about 200 words in length, based on the following notes.'') We suspect that the most common way of using an LLM to generate an abstract for a scientific paper, at least during the period in which ChatGPT could not accept paper-length inputs, involved inputting important fragments of a paper. The two-step procedure for abstract generation was intended to roughly imitate such a process, albeit imperfectly.\footnote{Researchers probably use ChatGPT in its intended dialogue mode, rather than with one-and-done prompting. We expect that scientists engage in a back-and-forth with ChatGPT, tweaking the prompting in various ways depending on its outputs. Such a procedure cannot be easily imitated for a study like ours that required generation of a significant number of abstracts.} We used ChatGPT-3.5 for the entirety of our project because if scientific abstracts in our dataset contain AI-generated language, it is most likely from ChatGPT-3 or ChatGPT-3.5.

In total, from the 10,000 human abstracts, we used ChatGPT-3.5 to generate 9,953 scientific abstracts. (For a small number of abstracts, ChatGPT would not provide a response, presumably due to the sensitivity of the topics.) We then compared the word usage in these AI-generated abstracts with word usage in the original scientific abstracts. Again, we only considered words for which a chi-square test indicated a significant difference in opm between the human- and AI-produced text. This gave us a list of items overused by ChatGPT. The procedure is illustrated in Figure REF.


Pubmed Database
2020
2024
Extract abstracts from 2020 and 2024
10k sample
Extract sample of 2020 abstracts
Find words that exhibit a significant spike in usage
omicron
intricacies
2020
2024
Step 1
Step 2
SUMMARY
ChatGPT writes summaries of abstracts
ChatGPT writes abstracts based on summaries
Exclude words whose increasing use has obvious explanation
Top 50 Spiking Words
Overused by AI
Step 3
Compile words that are significantly overused in AI-abstracts compared to human abstracts
If a word is found on both lists, it is a focal word.
21 Focal Words

\caption{Our procedure of identifying focal words.}

In the third step of our analysis, we returned to our list of 50 spiking words to ask:\ Is the word also on the ChatGPT-overuse list? If so, then it became a lexical item of interest – what we call a ``focal word'' (see Figure~\ref{fig:focalwordprocedure} for an illustration of this procedure). This gave us a list of 21 focal words (see Figure~\ref{fig:focalwordsopms} and Appendix A-REF). Each focal word (a) showed a significant spike in human usage opm between 2020 and 2024, (b) its spike lacks an obvious real-world explanation, and (c) ChatGPT tends to overuse it significantly more than humans when prompted to write a scientific abstract. Thus a plausible explanation for the increasing prevalence of each focal word is the use of AI for scientific writing.

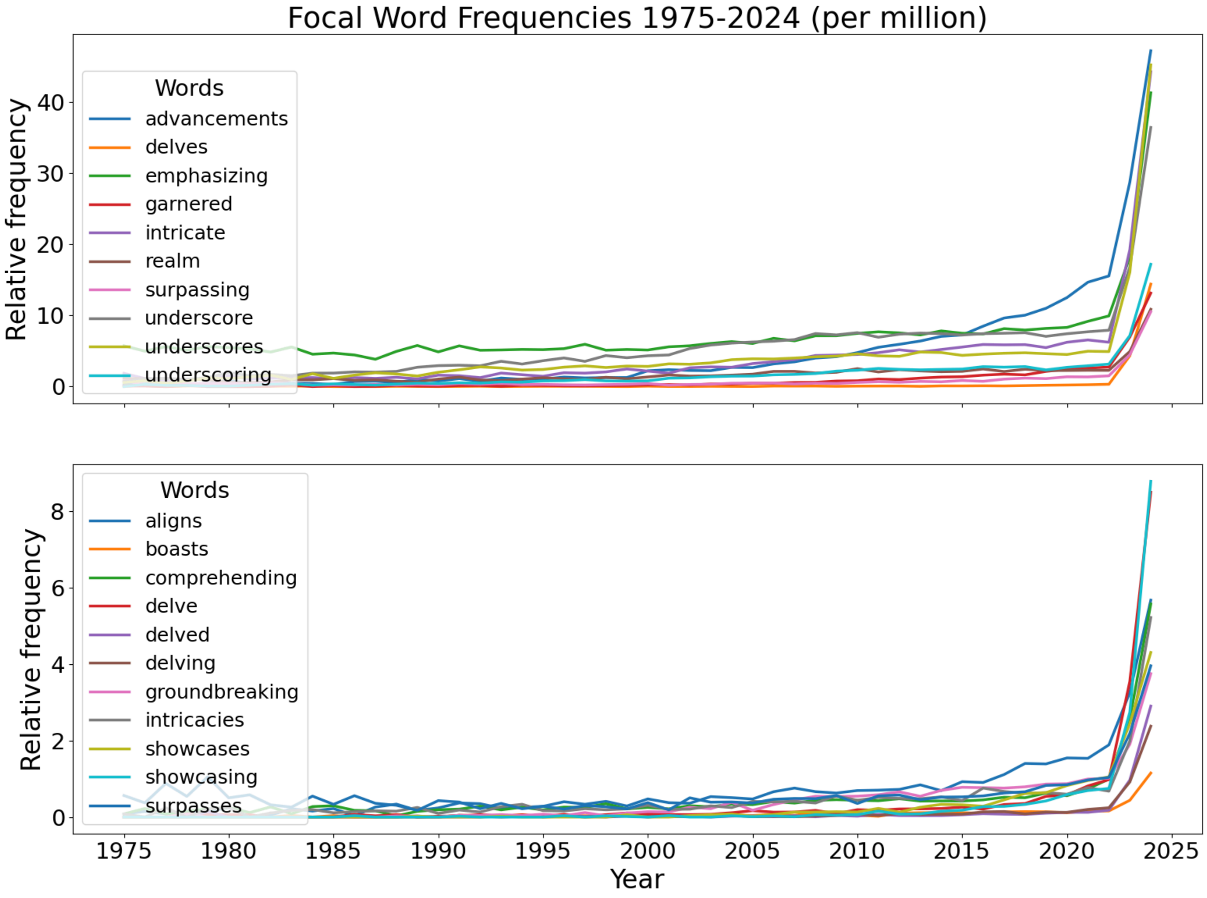


FIGURE 3a (top) and 3b (bottom): \caption{Occurrences per million words in PubMed for our 21 focal words.}

Since the part-of-speech category is not always clear for a given token, the focal word list contains inflected forms instead of lemmata. Several interesting trends can be seen in Figs. 3a and 3b. 18 of the 21 focal words showed a positive trajectory between 2010 and 2020 (see also Matsui 2024), a point we will return to in Section 7. Note also how inorganic the spikes look when compared with the more content-driven spikes shown as baselines in Figure~\ref{fig:baselineopms}.

This systematic, three-step method for constructing a list of focal words is novel. It represents an improvement over more informal ways of identifying AI-associated words for further analysis, and it can be applied to other corpora and other LLMs. Because we only used ChatGPT-3.5 to generate comparison abstracts (in the second step of the method), our list of focal words is specific to ChatGPT-3.5. We focused on ChatGPT because it is the most popular LLM \cite{Sarkar2023}, and it was probably especially dominant during the time period under study, as the range of available LLMs has expanded over time. In Appendix REF, we validate our results with ChatGPT-4.0(-mini). Future research should apply the method to other LLMs. This would shed light on whether the same words are overrepresented in the outputs of different models – or indeed, whether there are any LLMs that do not exhibit significant lexical overrepresentation at all.

3. The Puzzle of Lexical Overrepresentation

An obvious and pressing question now presents itself:\ Why are certain words used so often in AI-generated scientific writing? We call this ``the puzzle of lexical overrepresentation.'' There are a number of different factors that might be responsible for the overrepresentation of focal words in scientific abstracts generated by ChatGPT. These potential explanations have different implications about how ``baked in'' LLMs' tendency to overuse certain words is. It is important to emphasize that the explanations are not mutually exclusive:\ multiple factors may (and probably do) contribute to the behavior of interest, and some of the factors listed here can only operate in combination with others.

Initial Training Data: Although the focal words we have identified are overrepresented in ChatGPT-generated scientific abstracts relative to human-written scientific abstracts, it is possible that they are not overrepresented relative to the data on which ChatGPT was trained to do next-word prediction. One potential explanation of the prevalence of focal words in AI-generated abstracts is therefore that these words are actually being used by LLMs with the same frequency as in their training data. This would suggest that the focal words are more prevalent in the training text than in scientific abstracts.

Fine-Tuning Training Data: As mentioned above, after LLMs have been trained on a simple next-word prediction task, they are fine-tuned in various ways. For instance, an LLM might be fine-tuned with a corpus from a particular domain in which it will be used. Chatbots are often presented with sample dialogues to familiarize them with the structure of a conversation/Q\&A. Even if the focal words are not overrepresented with respect to ChatGPT's initial training data, it is possible that something about its fine-tuning data leads it to prefer the identified words. For example, if the focal words are overrepresented in the human-produced dialogues that are used in fine-tuning, it would be unsurprising that their prevalence would increase in ChatGPT's outputs.

Architecture: One possibility is that there is something about the architecture of LLMs, or perhaps ChatGPT in particular, that causes them to overproduce certain words. One of LLMs' characteristic features is their transformer architecture. Perhaps transformers tend to privilege some lexical items over others in an as-yet-unrecognized way. Note that even if this explanation proves correct – that is, if the use of our focal words can be shown to depend on the presence of a particular architectural feature – the question remains why this particular set of words tends to be overrepresented.

Choice of Algorithm: Many different algorithms are required in the construction of LLMs. Tokenization algorithms, for example, divide up an input string into discrete lexical items. Optimization algorithms determine how a model maximizes its reward function. It is possible that the choice of one of these algorithms over another will be shown to cause lexical overrepresentation of the focal words. Why the algorithm does so, and why particular words that are overrepresented are selected, would then be further important questions.

Context Priming: When we produced AI-generated text to construct our list of focal words, we asked ChatGPT to write an abstract for a scientific paper. A well-known strength of LLMs is their sensitivity to writing genre (REF). As many users have noticed, the output of LLMs is highly dependent on the domain and style requested by the prompt – a fact that prompt engineers have taken advantage of (REF). Perhaps there is something about being asked to engage in scientific writing that prompts LLMs to engage in excessive usage of the focal words. That is, maybe ChatGPT associates scientific writing in particular with words like ``delve'' and ``intricate.'' Note that this explanation, even if true, raises a further question:\ why exactly does ChatGPT associate the identified focal words with scientific writing? Answering that further question requires appeal to other factors on (or off) this list.

Reinforcement Learning from Human Feedback (RLHF): As mentioned in Section~\ref{sec:introduction}, human feedback is used in the later stages of training to give LLMs additional information about what a quality output looks like. A human evaluator might rate several potential responses, for example, with reinforcement learning then used to train the model to produce responses similar to highly-rated exemplars. It is possible that the human feedback used for fine-tuning encodes a preference for certain words. If responses featuring the words ``delve'' and ``intricate'' are rated more highly by human evaluators, for example, it would explain why there is overrepresentation of these words in the outputs of ChatGPT.

Other factors: This list of potential explanations is not intended to be exhaustive. Many other choices in model development – e.g., parameter settings, including temperature, Top-K, Top-P – could be found to influence the degree of lexical overrepresentation in LLM outputs. We leave it to the reader to extend this list.

Apportioning responsibility for lexical overrepresentation to each of these factors, individually or in combination, is not straightforward. The puzzle of overrepresentation arises in part because, as is widely recognized, LLMs are to a large extent ``black boxes'' whose inner workings are mysterious even to the developers building them (Knight 2017). Until further advances are made in LLM explainability or interpretability (e.g., Templeton et al. 2024), we will not understand many aspects of their behavior. There are, however, additional obstacles to solving the puzzle of lexical overrepresentation arising from the secrecy surrounding LLMs, including ChatGPT. Plenty of information that would be useful for discriminating between the potential explanations listed above is not public, even for open source models like Llama. For instance, we do not know exactly what LLMs' training data are (relevant to \#1 above), what fine-tuning steps are used (\#2), which sample dialogues are used in chatbot fine-tuning (\#2), what genres the models are exposed to during training (\#5), and who the human evaluators are (\#6). Since many aspects of LLM construction are closely-guarded secrets, in the remaining sections we develop several indirect ways of probing potential explanations of the puzzle of lexical overrepresentation.

4. Searching for Overrepresentation in (Potential) Training Data

The analysis in Section 2 indicates that the focal words are overrepresented in the outputs we elicited from ChatGPT compared to the pre-2023 PubMed data. Other research indicates that such words also appear less frequently in related datasets (Cheng et al. 2024, Liang et al. 2024a, Liang et al. 2024b, Gray 2024 [these papers use other datasets; DOUBLE CHECK that they do indeed fail to find overrepresentation). These results cast some doubt on the hypothesis that ChatGPT is using words like “delve” and “surpass” frequently because such words occur frequently in its training data. It is important to remember, however, that we do not know for certain what data any particular LLM has been trained on.

To further demonstrate that the focal words are probably not overrepresented in either the initial or fine-tuning training data, we analyzed several additional datasets. First, we checked the prevalence of the focal words in Arxiv abstracts (accessed 4 Aug 2024; contains data from 1986 onwards, averaged over all years), the Leipzig Corpus Collective (REFs in Overleaf; the English LCC contains mostly news texts and transcriptions, data from 2005 onwards; preprocessed snapshot from a previous project), and Wikipedia articles and discussions (accessed 4 Aug 2024). The results are presented in Table REF in Appendix B-REF. The opm of the focal words in our ChatGPT-3.5-generated abstracts far exceeds their opm in any of the four datasets examined. For example, “underscore” appears about 18.1 times per million words in our ChatGPT-3.5-generated text, compared to 5.2 (Arxiv), 1.5 (LCC), 7.9 (PubMed), and 0.7 (Wikipedia) times per million words in the other datasets.

Second, we conducted a similar analysis for various varieties of English from across the world using the International Corpus of English (ICE; kirk greenbaum in Overleaf). The results can be found in Figure REF in Appendix C-REF. Although ICE is relatively small compared to the other datasets (the subcorpora for most varieties contain about one million words), the results do not indicate that the focal words are especially prevalent in any particular variety of English. This suggests that the overrepresentation of focal words in ChatGPT’s output is probably not due to an overrepresentation of a certain variety of English in its training data. It has been hypothesized that LLMs might frequently use words like “delve” because they are more common in varieties of English spoken by the human evaluators hired to provide fine-tuning data, such as Nigerian English (Guardian REF). Although more research is needed, our initial analysis of ICE does not support this hypothesis.

5. Model Choices: Architecture and Algorithms

Having found no evidence for the overuse of focal words in likely sources of training data, we turn next to the hypothesis that choices about model architecture or algorithms are responsible for the puzzle of lexical overrepresentation. Ideally, to probe the effects of these choices, we would build a ChatGPT-like LLM ourselves and test the impact of each potential factor on the prevalence of our focal words. This requires vast resources, however, and is beyond our capabilities. A more feasible alternative would be to investigate a model that has several released variants – e.g., different versions of the same model using different optimization algorithms. Such a model must also be queryable with respect to information-theoretic measures like entropy (Shannon REF). To our knowledge, no LLM offers such fine-grained releases.

The closest we could find is the comparison between Llama 2-Base (Llama-2-7b-hf) and Llama 2-Chat (Llama-2-7b-chat-hf; \citealt{touvron2023llama}). We used the Llama 2 models in part because they are more similar to ChatGPT-3.5 than Llama 3. The main difference between these two versions of Llama is that Llama 2-Chat includes fine-tuning and RHLF, whereas Llama 2-Base does not. The Llama 2 models can also be queried for per-word entropy, which we normalized for length (see Formula 1; REF Jurafsky Martin).

\[
H_{\text{norm}} = -\frac{1}{L} \sum_{i=1}^{n} p(x_i) \log p(x_i)
\] REF get formula from J&M

By comparing the two models' per-word entropy for human- and AI-generated abstracts, we could assess whether each was more "surprised" by abstracts with an overrepresentation of focal words. If there is a difference between the models, that provides evidence about the source of lexical overrepresentation. We provided our sample of 10,000 human-written abstracts to the two versions of Llama 2, followed by the abstracts rewritten by ChatGPT-3.5 (see method description in Section 2). The results are presented in Table REF.

	Llama 2-Base	Llama 2-Chat
human	1.664	1.104
ai	1.844	1.045

\caption{Your caption here}

Table REF: Normalised entropy for human abstracts compared to ChatGPT-generated abstracts. Higher values of entropy mean that the model is more "surprised." Lower values indicate that it finds language more predictable.

We observe that Llama 2-Base is less "surprised" by human-written text, while Llama 2-Chat is less "surprised" by AI-generated abstracts, in which the focal words are overrepresented. This suggests the overuse of focal words is driven by some factor that differs between the models, rather than something they have in common. Given that the model architecture and algorithms are held constant across Llama 2-Base and Llama 2-Chat, our results suggest that these factors are not the primary causes of lexical overrepresentation. Instead, they indicate that fine-tuning and RHLF – which differ between the models – might be important contributors.

These results are necessarily limited. They show that Llama 2-Base finds human-generated abstracts more predictable than AI-generated abstracts, but we cannot claim definitively that this difference is driven by the prevalence of focal words rather than some other feature of AI-generated text. Moreover, most of our paper is concerned with ChatGPT rather than Llama. The difficulty is that there are no models of ChatGPT (v.3 or above) that can be queried in the described fashion. We think Llama is a useful approximation, even if it is not the ideal testing ground.

6 RLHF: An Experimental Approach

Our model testing with Llama highlighted a potential role for RHLF in causing lexical overrepresentation. This hypothesis has some intuitive plausibility: when human raters evaluate alternative answers to a query, perhaps they are exhibiting a preference for answers containing certain words. Since the LLM is trained to align its answers with human preferences, it would learn to use those words more frequently (REF , christiano2017deep, ziegler). To further investigate this potential explanation, we conducted an exploratory online study in which participants indicated whether they preferred scientific abstracts that contained some of our 21 focal words.

Materials. We randomly sampled shorter PubMed abstracts (70-100 words) from the year 2020 and, with Python and using the OpenAI API, asked ChatGPT-3.5 to rewrite them with and without focal words. For the focal-word abstracts, the prompt for the rewrite included four randomly selected words from our 21 focal words. An example prompt is: “Please write a 100-word abstract for the following scientific paper, using words such as 'delves,' 'underscores,' 'surpasses,' and 'emphasizing': [SUMMARY].” (The summary was generated via the procedure described in Section 2.) The script instructed ChatGPT to generate and revise an abstract until it contained at least three of the focal words. For the no-focal-word abstracts, we used a similar prompt: “Please write a 100-word abstract for the following scientific paper, ensuring that none of the following words are used: [list of blockwords].” The blockwords included the 21 focal words plus another 21 words identified using the methodology described in Section 2. The script prompted ChatGPT to generate and revise an abstract until it contained none of the blockwords.

We created 200 items, each consisting of one abstract with focal words and one without (for the same paper). We manually filtered out a handful of potentially problematic items (e.g., one abstract opened with REFYZ). Considerably more than half of the abstracts with focal words included “delve” in the first sentence; we call items containing these abstracts “delve-initial” items. To compile a bank of 30 critical items, we selected the 15 delve-initial items and the 15 other items with the smallest difference in length between the abstracts with and without focal words. (We capped delve-initial items at 50\% to prevent participants from detecting the study’s purpose.) We also constructed 30 pairs of distractor items in the same manner as the critical items, except both abstracts in each distractor item were generated using the non-focal-word prompt. A full list of experimental items can be found on our Github, and two examples are in Appendix C-REF.

Participants. We used Prolific (prolific.com) to recruit participants. As noted above, public information about the human evaluators employed to provide feedback in RLHF is limited. One report indicates that, at one point, about half of the evaluators employed by OpenAI were from the Philippines and Bangladesh (Ouyang et al., 2022); another suggests that at least some of OpenAI’s evaluators were from Kenya (Perrigo, 2023). We recruited 201 participants from India

(140 male, 61 female). Average age was 31.3 years (stdev: 10.6). We also collected data on self-assessed English proficiency and first languages (see our GitHub for the data). Participants were compensated at an average rate of \$15 per hour.

Task and Exclusions. The study began with IRB information, followed by task instructions, and then the items. An image of the interface can be found in Appendix D-REF. Participants evaluated 20 items in total, indicating which abstract they preferred out of the two presented. The first item was a calibration item, followed by (in random order) five critical items, ten distractor items, two items checking language abilities, and two attention checks ("This is not a real item, please click on the left button" inserted in the middle of the text). Thus, the proportion of critical items was 25\%. Each time an item was displayed, it was randomly determined which abstract was displayed on the left vs. right. If a participant failed one of the attention checks, their data was disregarded. Participants were warned if they were proceeding unrealistically fast (0.25 * (225 ms + 25ms * character length of an item); following self-citation Jana), and items with excessively fast rating times were excluded from our analysis. We also excluded data from participants who completed less than 10 out of the 20 items. After exclusions, we analyzed a total of 1822 ratings, with 1215 ratings for distractor items and 607 ratings for critical items, resulting in each critical item receiving an average of 20.2 ratings (stdev: 3.4).

Analysis. Our original plan was to test all 30 critical items together in a chi-square analysis against the distractor items as an approximation of random choices. These results are still reported below. However, during the creation of the items, we noticed the mentioned excess of text-initial delves and split the critical items into delve-initial items and other items. A lower N per condition and a higher-than-expected exclusion rate left us considerably below the originally estimated sample size from a pre-study power analysis, thus, we added an explorative mixed-effects logistic regression model, with rating as the dependent variable and condition as the independent variable, including items as a random effect (rating ~ condition + (1 | item_id)). Distractor items served as the intercept condition. For delve-initial items and other items, a focal word preference was encoded with 0, and the non-focal word preference with 1. For the distractors, there are two non-focal word conditions, encoded as 0 and 1.

Results. Contrary to our expectations, when all critical items are analyzed together, there is a slight preference for the non-focal word condition. The overall difference, as indicated by the chi-square test, between all critical items and distractor items is not significant (p = 0.174). This dispreference is driven by the delve-initial items, as Figure REF illustrates. As to the logistic regression model, we observe that the coefficient for the distractor items, represented by the intercept condition, is 0.500. This is close to random, indicating that participants did not exhibit a significant preference between the abstracts in the distractor items, validating our methodology (also see results in Appendix E-REF). The analysis also shows that delve-initial items differed significantly from the distractors (p = 0.023), with a coefficient of 0.082, indicating that for the delve-initial items, participants preferred the abstracts without focal words. While participants exhibited a slight preference for abstracts with focal words for the other critical items (coefficient = -0.017), the difference from the distractor items was non-significant (p = 0.651). The group variance was small (0.003), indicating that most of the variability in the ratings was due to the fixed effects. The model converged successfully (log-likelihood = -1324.9522, mean group size = 30.4). A Wald test to determine whether delve-initial items and the other items differed from each other was statistically significant (p = 0.03, Wald Test Statistic: 4.77).

We provide descriptives for individual items, and in the following, we consider a preference to be robust if a random outcome falls outside the margin of error, and marginal if it falls within that margin.[1] Figure REF in Appendix REF illustrates this logic for the distractor items. Out of the 15 delve-initial items, we observe a robust preference for the focal-word abstract in only one item. In six items, there was a robust preference for the non-focal-word abstract. This pattern shifts for the other critical items, which did not have 'delve' in the first sentence. For five of these other items, there is a robust preference for the focal-word abstract. For three items, there is a robust preference for the non-focal-word abstract (Figure REF).

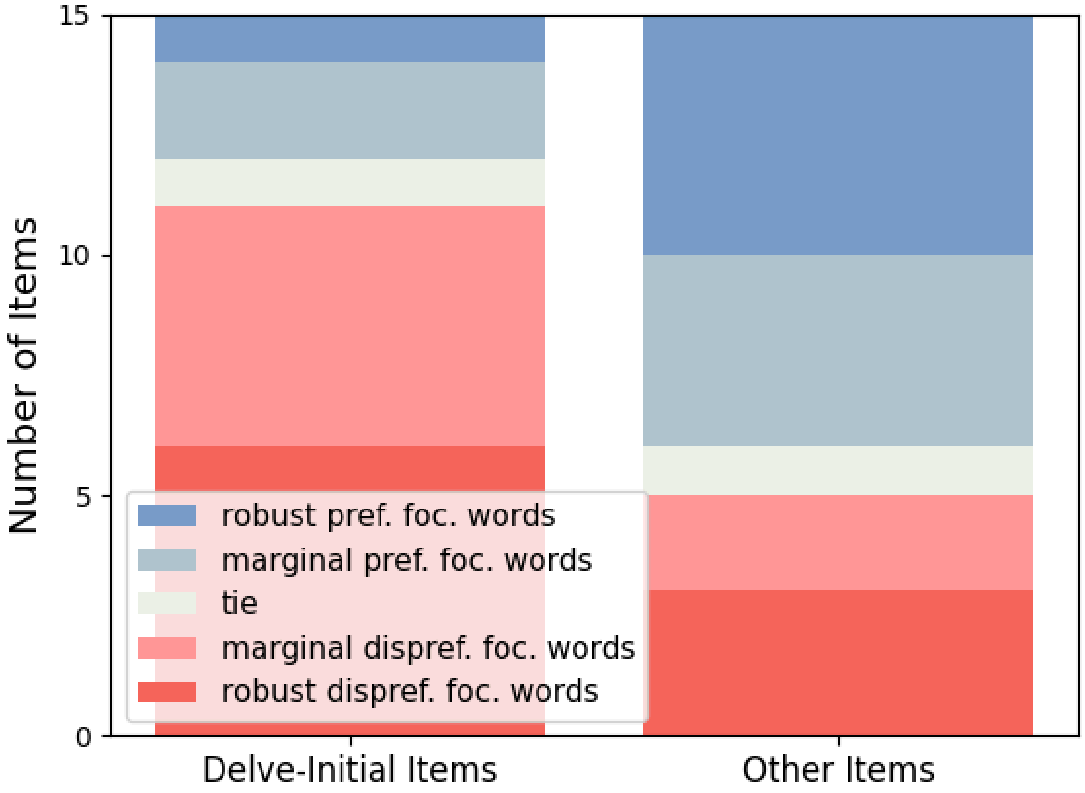


Figure REF:\caption{The results of our experiment.}

[1] Testing for significance for individual items is hindered by a relatively low N per item (an average of 20.2) and the lack of sensitive tests suitable for this scenario. In a chi-square test, 17 out of 20 participants would have to prefer one abstract over the other to yield a statistically significant result; for a binomial test, it would have to be 15 out of 20.

What explains the difference between delve-initial and the other critical items? Although we took steps to limit the number of delve-initial items each participant saw, we suspect that some of them became or were already sensitive to the occurrence of “delve.” Our participants were probably disproportionately young people with an affinity for technology, and so more likely to be familiar with the discourse surrounding AI language use. Wariness about the word “delve” might explain why participants preferred the abstracts without focal words in the delve-initial items, though we would want to see these results confirmed with a larger sample before drawing definitive conclusions.

The analysis of the remaining items is weakly consistent with the idea that RLHF is a source of lexical overrepresentation. Even a slight preference for responses with focal words among human evaluators could explain why ChatGPT tends to overuse those words in its outputs. There is, however, an inherent obstacle to research on the role of RLHF. Companies building LLMs often solicit human feedback from workers who are underpaid, stressed, and under time pressure (toxtli2021quantifying roberts2022precarious novick2023dirty). It is difficult to simulate these conditions ethically in a research environment. For example, many online recruitment platforms, including Prolific, rightly require decent compensation. Although it complicates further study, this economic reality lends plausibility to RLHF as a source of lexical overrepresentation. Rushed human evaluators might well base their evaluations on form over content, as the former is often easier and quicker to evaluate than the latter. If certain words are treated as a proxy for quality, that could explain their overrepresentation in LLM outputs. Again, much more research is needed to follow up on our exploratory findings and assess this hypothesized explanation of the puzzle of lexical overrepresentation.

7. Discussion and Concluding Remarks

It has been observed that LLMs overuse certain lexical items, a fact even acknowledged by OpenAI (REF openai tweet). Our work formalized this finding and identified 21 focal words whose usage has spiked in scientific abstracts and that are overused by ChatGPT-3.5. These results provide additional evidence that recent changes to Scientific English are partly driven by AI. Unlike previous research, our work also explored possible explanations of the puzzle of lexical overrepresentation. We failed to find evidence that training data, model architecture, or algorithm choices play a role. However, our model testing with Llama and our experimental results are consistent with the hypothesis that RLHF contributes to overuse of particular words by ChatGPT.

Future research should further probe the impact of each factor canvassed in Section 3 on lexical overrepresentation. (This includes model choices and training data; despite our negative results, we suspect that these factors do influence the lexical choices of LLMs.) We would also like to see further confirmation of the role of RHLF. Unfortunately, the lack of procedural and data transparency surrounding LLM development is an obstacle to the investigation of potential sources of lexical overrepresentation. This issue becomes even more pressing when considering the impact that LLMs have on language usage.

It would also be interesting to apply the present methods to LLMs besides ChatGPT. Our qualitative impression, partially substantiated by the results in Section 5, is that ChatGPT and Llama overuse many of the same words, but a systematic investigation of the degree of overlap is needed. This research could also be extended to domains beyond Scientific English. We suspect that Scientific English played a minor role in LLMs' RLHF. It seems more likely that human evaluators rated academic writing in general, with their preferences then shaping LLMs' scientific writing through overspill. Finally, we wish to follow up on our loose impression that poorer quality inputs tend to lead to greater overrepresentation of focal words in model output.

We believe that more attention should be paid to how LLMs are changing language. Almost all of our 21 focal words were already increasing in usage in the years leading up to the release of ChatGPT, suggesting that LLMs may accelerate language change. With the increasing prevalence of AI-generated text in many areas of life, LLMs are arguably influencing the language usage even of people who don't themselves interact with these models. Our findings show that lexical overrepresentation remains a feature of current iterations of ChatGPT (REF appendix), indicating that the phenomenon is here to stay.

Still, it is difficult to predict just how AI will shape language in the future. Discussions on social media and in academic discourse, plus our exploratory findings for items with "delve," indicate that there is some public awareness of LLMs' overuse of particular words. This awareness could influence future rounds of RLHF, leading to a realignment of AI and human preferences. At the same time, the language of today – lexical representations and all – will become the training data for the models of tomorrow, raising concerns about model degradation over time (REFs \cite{alemohammad2023self,briesch2023large,hataya2023will,shumailov2023curse)}.

A major consequence of the advent of LLMs seems to be a decoupling of form and content (REFs). Many of us use the heuristic of taking the fluency or style in which something is written as a proxy for the quality of its content (mcnamara2010linguistic , and in an L2 context kim2018modeling). Because LLMs are masterful at generating fluid text in just about any style, this heuristic is radically undermined by the increasing ubiquity of LLM-generated text (other REFs). The irony is that, if our hypothesis about RLHF proves correct, this heuristic has shaped model training as well. Human evaluators are using the presence of certain words as a proxy for response quality, causing the models to overuse those words in their outputs. LLMs are thus undermining the very same heuristic that has shaped their own lexical preferences.

One thing is certain: through LLMs, tech companies are having a global impact on language usage. We believe this strengthens the case for broader societal debate about the power and responsibilities these companies have. Moreover, our speculations about how the feedback of rushed and underpaid workers might contribute to lexical overrepresentation compound ethical worries about the poor working conditions of tech companies' employees in the Global South (REFS). There are thus both moral and non-moral reasons to apply greater scrutiny to how human feedback is collected and used in the training of LLMs.

Appendix A: List Of Focal Words

word	opm-2020	opm-2024	increase-\%
delves	0.21	14.38	6697.14
delved	0.12	2.9	2240.47
delving	0.12	2.38	1816.83
showcasing	0.59	8.79	1396.03
delve	0.58	8.5	1374.92
boasts	0.11	1.15	918.18
underscores	4.5	45.19	903.61
comprehending	0.56	5.58	898.95
intricacies	0.6	5.22	772.85
surpassing	1.37	10.5	667.48
intricate	6.22	44.22	611.24
underscoring	2.7	17.17	536.94
garnered	2.44	13.13	437.19
showcases	0.82	4.31	422.45
emphasizing	8.3	41.27	397.12
underscore	7.42	36.4	390.65
realm	2.25	10.85	381.1
surpasses	0.85	3.96	367.55
groundbreaking	0.87	3.75	330.42
advancements	12.49	47.17	277.59
aligns	1.55	5.68	266.97

\caption{Your caption here}

Appendix B: Analysis Of Further Corpora and GPT-4o

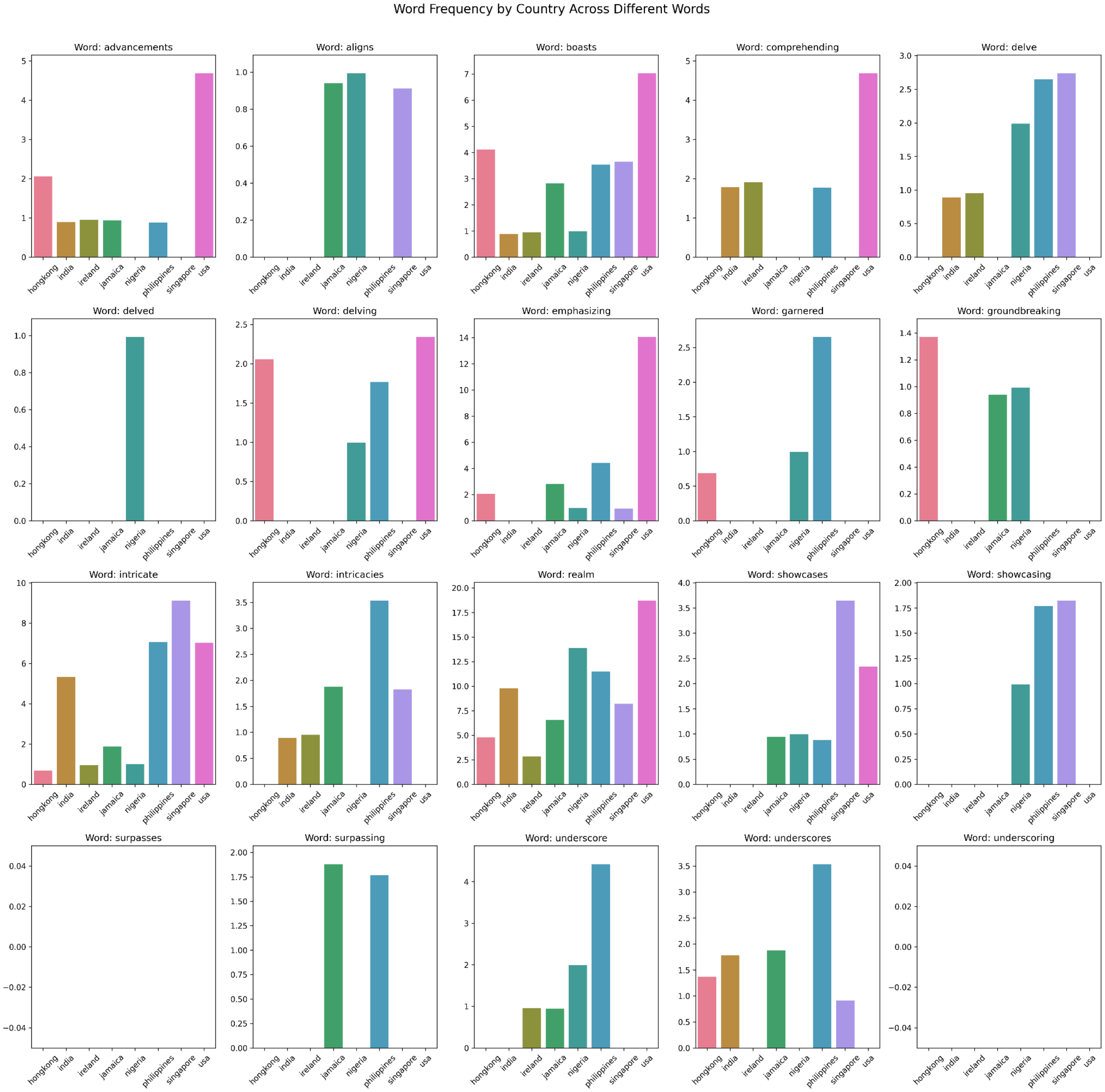
Word	ChatGPT-3.5	ChatGPT-4o-mini	Arxiv	LCC	Pubmed	Wiki
of	45624.843	42622.645	42842.716	27363.47	38634.994	23116.178
and	38889.239	32537.785	26395.275	28488.532	39469.962	21149.631
the	63174.047	55111.229	72009.628	59324.621	52139.049	53379.319
data	978.907	1075.589	2484.2	418.292	1734.747	142.809
results	4074.641	3307.316	2352.127	244.522	1722.072	95.365
i	32.205	61.168	414.027	4715.417	214.824	8041.613
year	78.5	61.774	37.583	1076.291	217.246	397.608
patients	4416.822	3936.56	48.974	131.475	4775.725	23.036
advancements	319.369	407.585	22.536	2.561	15.534	1.112
aligns	6.709	19.986	6.684	1.323	1.889	0.901
boasts	5.368	0.606	0.432	14.112	0.16	1.483
comprehending	6.709	7.267	1.765	0.371	0.993	0.312
delve	19.457	18.169	4.072	2.229	0.98	1.214
delves	183.168	23.014	3.196	0.792	0.316	0.525

delved	6.709	0.606	0.3	0.615	0.179	0.384
delving	8.722	0.606	0.724	0.764	0.243	0.606
emphasizing	138.214	367.614	10.206	2.815	9.92	2.638
garnered	20.799	173.209	4.094	4.344	2.738	4.613
groundbreaking	38.915	17.563	2.467	5.91	1.015	2.262
intricate	163.039	316.136	17.873	4.795	6.217	2.132
intricacies	15.432	27.253	1.978	1.236	0.677	0.676
realm	10.735	54.506	11.531	9.217	2.272	8.461
showcases	28.851	4.239	3.193	4.65	1.05	1.462
showcasing	30.192	58.14	5.885	5.417	0.75	1.651
surpasses	4.026	4.239	11.16	1.137	1.037	0.401
surpassing	5.368	17.563	7.608	1.663	1.506	1.421
underscore	18.115	1365.078	5.172	1.527	7.909	0.715
underscores	60.385	1048.942	4.95	1.896	4.912	0.896
underscoring	10.064	313.714	2.573	0.661	3.153	0.198

\caption{Occurrences per million for selected baseline words and our 21 focal words. Results are averaged across all given years of the corpus.}

We used the same summaries from the sample of 10,000 abstracts and used a Python script to generate abstracts using GPT-4o-mini, which were then analyzed as per Section REF. The system's role was set as 'You are a world-leading scientist.' and the prompt was the same as that used with GPT-3.5. The results are very similar, with a few exceptions: 'boasts' is no longer overused; 'delve' is still overused, but to a lesser extent; and the usage of 'underscore' has increased significantly. These differences could be artifacts of the methodology (the GPT-4-generated abstracts are based on the same GPT-3.5 summaries used in Section-REF), the consequence of active intervention, RLHF workers responding to overuse, and/or other factors. We also conducted a spot-check with a prompt specifying the role 'You are a helpful assistant.' using GPT-4o. For each role, we generated 500 abstracts and analyzed them. There was no noticeable difference with GPT-4o-mini.

Appendix C1: Analysis of the International Corpus of English



\caption{Your caption here}

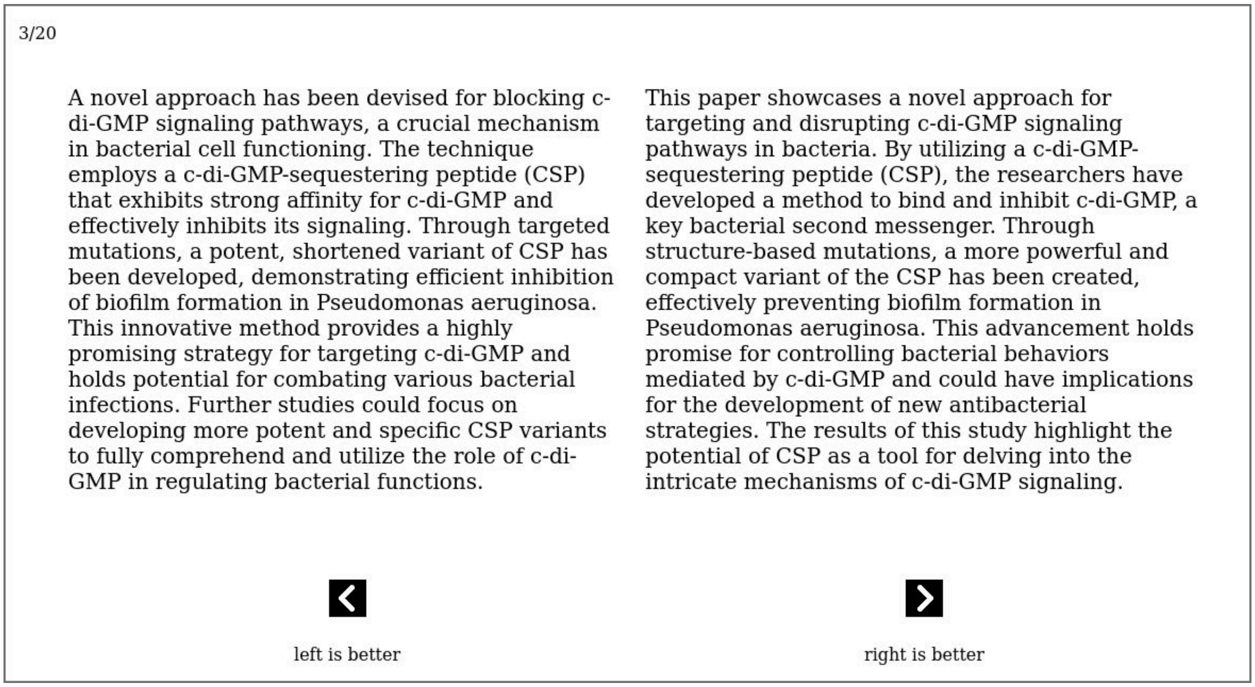
Appendix C2: Examples Of Experimental Items

A delve-initial item.
Focal word condition: This study delves into the impacts of maintaining mean arterial blood pressure (MABP) at a specific level during the prehospital and initial hospital phases of treatment for patients with traumatic spinal cord injury (TSCI). The results show a strong correlation between maintaining MABP at 80 mm Hg and improved long-term neurological outcomes. Specifically, prehospital and operating room levels had the greatest effect, while maintaining MABP at 80 mm Hg during the first 2 days in the neurointensive care unit (NICU)

was also beneficial. These findings surpass previous knowledge and highlight the advancements in comprehending the relationship between MABP and TSCI recovery.
Non-focal word condition: This study explores the impact of maintaining mean arterial blood pressure (MABP) at 80mm Hg during prehospital and initial hospital treatment on long-term neurological outcomes in patients with TSCI. Results showed a significant correlation between higher MABP levels and improved outcomes, with the strongest impact observed in the prehospital and operating room phases. The benefits of maintaining MABP at 80mm Hg were also observed in the first 2 days in the NICU. These findings highlight the crucial role of MABP management in minimizing neurogenic shock-induced damage and emphasize the importance of maintaining adequate blood pressure in TSCI patients.

A non delve-initial item.
Focal word condition: This paper showcases a novel approach for targeting and disrupting c-di-GMP signaling pathways in bacteria. By utilizing a c-di-GMP-sequestering peptide (CSP), the researchers have developed a method to bind and inhibit c-di-GMP, a key bacterial second messenger. Through structure-based mutations, a more powerful and compact variant of the CSP has been created, effectively preventing biofilm formation in Pseudomonas aeruginosa. This advancement holds promise for controlling bacterial behaviors mediated by c-di-GMP and could have implications for the development of new antibacterial strategies. The results of this study highlight the potential of CSP as a tool for delving into the intricate mechanisms of c-di-GMP signaling.
Non-focal word condition: A novel approach has been devised for blocking c-di-GMP signaling pathways, a crucial mechanism in bacterial cell functioning. The technique employs a c-di-GMP-sequestering peptide (CSP) that exhibits strong affinity for c-di-GMP and effectively inhibits its signaling. Through targeted mutations, a potent, shortened variant of CSP has been developed, demonstrating efficient inhibition of biofilm formation in Pseudomonas aeruginosa. This innovative method provides a highly promising strategy for targeting c-di-GMP and holds potential for combating various bacterial infections. Further studies could focus on developing more potent and specific CSP variants to fully comprehend and utilize the role of c-di-GMP in regulating bacterial functions.

Appendix D: The Rating Interface

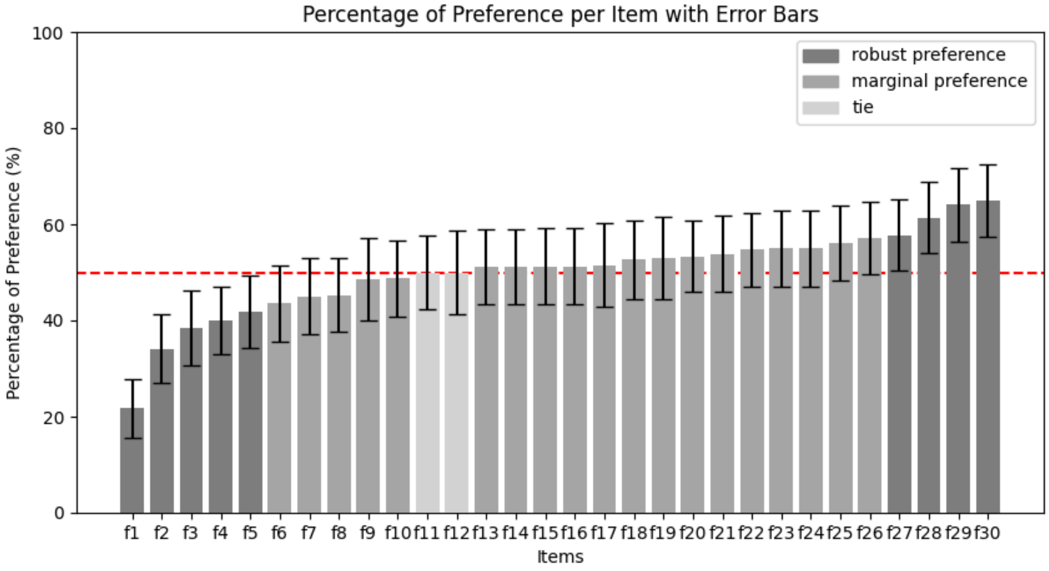


A novel approach has been devised for blocking c-di-GMP signaling pathways, a crucial mechanism in bacterial cell functioning. The technique employs a c-di-GMP-sequestering peptide (CSP) that exhibits strong affinity for c-di-GMP and effectively inhibits its signaling. Through targeted mutations, a potent, shortened variant of CSP has been developed, demonstrating efficient inhibition of biofilm formation in Pseudomonas aeruginosa. This innovative method provides a highly promising strategy for targeting c-di-GMP and holds potential for combating various bacterial infections. Further studies could focus on developing more potent and specific CSP variants to fully comprehend and utilize the role of c-di-GMP in regulating bacterial functions.

This paper showcases a novel approach for targeting and disrupting c-di-GMP signaling pathways in bacteria. By utilizing a c-di-GMP-sequestering peptide (CSP), the researchers have developed a method to bind and inhibit c-di-GMP, a key bacterial second messenger. Through structure-based mutations, a more powerful and compact variant of the CSP has been created, effectively preventing biofilm formation in Pseudomonas aeruginosa. This advancement holds promise for controlling bacterial behaviors mediated by c-di-GMP and could have implications for the development of new antibacterial strategies. The results of this study highlight the potential of CSP as a tool for delving into the intricate mechanisms of c-di-GMP signaling.

\caption{Your caption here}

Appendix E: Ratings For The Distractor Items



\caption{Your caption here}