Word Overuse and Alignment in Large Language Models: The Influence of Learning from Human Feedback

Tom S. Juzek[1 [0000−0002−3204−3879]] and Zina B. Ward[1 [0000−0003−0160−6656]⋆]

Florida State University, Tallahassee FL 32306, USA

Abstract. Large Language Models (LLMs) are known to overuse certain terms like "delve" and "intricate." The exact reasons for these lexical choices, however, have been unclear. Using Meta's Llama model, this study investigates the contribution of Learning from Human Feedback (LHF), under which we subsume Reinforcement Learning from Human Feedback and Direct Preference Optimization. We present a straightforward procedure for detecting the lexical preferences of LLMs that are potentially LHF-induced. Next, we more conclusively link LHF to lexical overuse by experimentally emulating the LHF procedure and demonstrating that participants systematically prefer text variants that include certain words. This lexical overuse can be seen as a sort of misalignment, though our study highlights the potential divergence between the lexical expectations of different populations – namely LHF workers versus LLM users. Our work contributes to the growing body of research on explainable artificial intelligence and emphasizes the importance of both data and procedural transparency in alignment research.

Keywords: Computational linguistics · Large Language Models · Alignment · Preference Learning · Lexical Overuse.

1 Introduction

Following the arrival of Large Language Models (LLMs), observers were quick to note their tendency to overproduce certain lexical entries [1,2,3,4,5,6,7,8,9]. Much of the discourse centered on Scientific and academic English, focusing on words such as "delve", "intricate", and "realm." For this reason, we also concentrate on Scientific English here. While changes in Scientific English over decades and centuries are well-documented [10,11,12,13], the language shifts following the introduction of LLMs have been unprecedented, with certain words (like "delve") seeing a sudden and dramatic increase in usage.

Thus, it has been established *that* certain lexical biases exist in LLMs, with evidence demonstrating their influence on written language. However, the question of *why* this lexical overrepresentation arises remains open. While some have

⋆ Conceptualization: TSJ, ZBW (eq.). Code, Methodology: TSJ. Write-up: TSJ, ZBW (eq.). GitHub repository: github.com/tjuzek/lhf. Computational setup: 2024 Thelio Custom machine, GeForce RTX 3090.

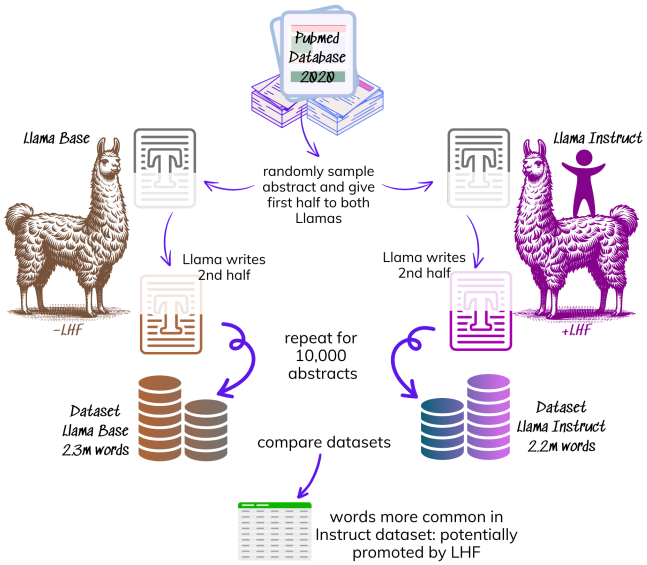

Fig. 1. An illustration of the procedure used to identify lexical preferences that are potentially induced by Learning from Human Feedback (LHF); created with Canva.

pointed to Learning from Human Feedback (LHF) as a significant contributor to these lexical choices [14,15], conclusive evidence is still missing.

Learning from Human Feedback is a procedure applied after initial model training during which human evaluators indicate preferences through A/B testing or ranking. It was first introduced in the form of Reinforcement Learning from Human Feedback (RLHF; [16,17]), though a more recent and increasingly popular form of LHF is Direct Preference Optimization (DPO), which aligns models by directly optimizing for human preferences without relying on reinforcement learning [18]. LHF was introduced to align models more closely with human preferences. Alignment, which reflects "how closely the model's opinions or stances mirror those of different social groups" [19], is a major challenge in AI [20,21,22]. A model is *misaligned* for a target group when its output does not align with the group's opinions, values, and/or expectations. LHF is recognized as a key factor contributing to the success of models like ChatGPT [23]. However, researching the effects of LHF is difficult due to lack of transparency surrounding the procedures and datasets used in model development [24].

The present study addresses the potential link between LHF and the lexical choices of LLMs through a two-step process. First, we introduce a method for identifying lexical preferences in LLMs that are potentially induced by LHF. This procedure can aid efforts to mitigate the most extreme cases of lexical overrepresentation (Section 2). Second, we conduct an experiment that emulates the LHF procedure in order to test whether humans indeed prefer texts containing the words identified by our initial procedure. This represents an empirical test of the hypothesis that LHF plays a role in shaping LLMs' lexical choices (Section 3),

based on one of Meta's popular Llama models. While our findings provide evidence for an LHF effect, other contributing factors remain to be systematically investigated. Finally, we discuss the implications of our study (Section 4) and its limitations (Section 5).

Related Work

Many studies explore the linguistic behavior of LLMs and their effects on (written) human language [1,2,3,4,5,6,7,8], with a few investigating spoken language [25,26]. Most of this work is situated at the word level, though there is also research on syntactic behavior [27]. Procedures for identifying LHF-induced overuse have been proposed [5,9], but these involve a manual component [5,9] and/or have a different focus [5]. Similar concerns have been raised for other behaviors exhibited by LLMs (see the xAI literature; [28,29,30]). Overlap between human linguistic preferences and model behavior has been shown, though with a small sample [9]. For non-lexical form (such as boldface or emoji use), it has been found that even subtle differences in preferences during human preference training can result in substantial differences in model behavior [31].

2 Procedure to Identify Potentially LHF-Induced Lexical Preferences

As a first step, we develop a low-cost procedure to identify lexical preferences in LLMs that may originate from LHF training. Our approach involves generating language outputs from both a pre-LHF model and a post-LHF model and then comparing word usage in the outputs. Here, we use Llama 3.2-3B Base and Llama 3.2-3B Instruct [33] (via the Hugging Face Transformers library [34]). The Llama family is, to our knowledge, the closest available approximation between models trained with and without LHF, which for Llama 3 involves Direct Preference Optimization. At the time of our research, Llama 3.2 was the most recent version of the Llama model family. While larger variants (11B and 90B) were available, they primarily added multimodal capabilities; improvements in textual reasoning abilities were minor [35]. At the time of research, a broader model comparison would have been difficult: of all the major LLM developers, only Meta had released both base and instruction-tuned models. Since then, models like OLMo [36] and Falcon [37] have gained popularity and would now be strong options.

There are other differences between Llama Base and Llama Instruct [33], most notably instruction tuning, optimization for tooling, and safety mitigation. However, none of these, including instruction tuning [38], are known to contribute to lexical overrepresentation. LHF remains the most plausible contributor to shifts in language output. This makes the Llama models well-suited for our purposes. All technical implementations described in this paper were carried out in Python 3 ([39]; v3.12.3).

Although our study focuses on Scientific English, the procedure we present is transferable to other domains. Here the procedure is applied to abstracts

from PubMed from 2020 [40], as this predates the mainstream availability of LLMs. We randomly sampled 10 000 abstracts and filtered out those with fewer than 40 words, which resulted in 9 853 abstracts. Each abstract was split in half by word count (rounding down), and each of the Llama models, Base and Instruct, were prompted to continue writing based on the initial half of the abstract (Prompt: 'Continue the following academic article: \"{first_half} '). Models were, if needed, cut off after twice the input length. The generated continuations were cleaned in order to remove issues such as generation loops (e.g., repetitive sentences) and meta-comments (e.g., "Certainly, here is ..."), using GPT-4o [41,42] (Prompt: 'The following text is meant to be a continuation of a scientific abstract. In some of the continuations, however, the AI finishes the abstract and continues with commentary. Please detect potential switches, and remove any commentary: \n\n"{input_text}"\n\n Output only the cleaned abstract. If the entire text is commentary, output an empty string.').

This process resulted in two corpora of PubMed abstract continuations: one generated by Llama Base (totaling 2.3m words) and the other by Llama Instruct (2.2m words). Both corpora were tagged for part-of-speech using spaCy ([43]; v3.8.3, en_core_web_sm v3.8.0, tagging of all data took about 140hrs), enabling the disambiguation of identical surface forms across word categories (e.g., "to_PART run_VERB" vs. "a_DET run_NOUN") and the grouping of conceptually related forms under a common lemma ("delve" and "delves"). Relative frequency usage was compared between the two corpora (similar to what one sees in the Google Ngram Viewer [44]). Here and in Section 3, we focus on statistically significant differences between Base and Instruct lexical usage, determined through a chi-square test. The top five items showing an increase in usage in the Instruct model compared to the Base model are as follows: "nuanced_ADJ (+8342%)", "nuance_VERB (+6301%)", "firstly_ADV (+4794%)", "reliance_NOUN (+3193%)", "generalizability_NOUN (+3124%)"; also see Table 1 for further entries and our GitHub for the full list.

Our procedure serves as a proof of concept: the identification of lexical items potentially favored by LHF can be automated. The procedure is validated in part by the observation that many of the identified words have been discussed in the literature on the distinctive lexical choices of LLMs [3,4,5,6,7,8,9]. However, the procedure does not necessarily identify words that are overused by Llama Instruct relative to human-generated text; the operative comparison is with Llama Base. Nevertheless, there seems to be considerable overlap between the words overused by Instruct relative to Base, and the words overused by Instruct relative to a human baseline. We compared the Llama Instruct outputs to a human baseline, the actual second halves of the randomly sampled PubMed abstracts. Almost all (813 out of 814) of the words used significantly more by Llama Instruct than Llama Base (Table 1) were also used significantly more by Instruct than in the human baseline. Thus, when it comes to the lexical items that distinguish LLM-generated text from human-generated text, the procedure in its current form effectively identifies many of the most extreme cases.

Lemma_POS	opm Ll-B	opm Ll-I	Incr. %
nuanced_ADJ	0.6	51.4	8342.8
nuance_VERB	0.6	39	6301.7
firstly_ADV	2.4	119.2	4794
reliance_NOUN	1.2	40.1	3193.6
generalizability_N	2.4	78.5	3124
underscore_VERB	4.3	124.9	2829.1
radar_NOUN	0.6	16.4	2590.6
staffing_NOUN	0.6	13	2033.9
socioemotional_ADJ	0.6	13	2033.9
multifacete_VERB	0.6	11.9	1848.3
flake_NOUN	0.6	10.7	1662.8
interoceptive_ADJ	0.6	10.7	1662.8
vocabulary_ADJ	0.6	10.7	1662.8
theanine_NOUN	0.6	10.7	1662.8
secondly_ADV	6.1	103.4	1597.8
finish_NOUN	0.6	10.2	1570
daa_NOUN	0.6	10.2	1570
necessitate_VERB	0.6	9.6	1477.2
behavioral_NOUN	0.6	9.6	1477.2

Table 1. Lemmata and part-of-speech for the Top 20 words identified using the procedure described in Section 2. Compared are occurrences-per-million (opm) for Llama Base (Ll-B) vs. Llama Instruct (Ll-I).

Assuming such divergences from human-generated text are undesirable and hence a form of bias (a point to which we will return in Section 4), the procedure is a method for uncovering lexical biases in LLMs. The degree of such bias observed in LLM outputs suggests that either no robust identification mechanisms were applied during model development, or existing mechanisms have proven too weak, which motivates the need for a procedure like ours. Our insights could also inform the discourse on AI-generated text detection [45,46,47,48], as such methods often rely on identifying atypical lexical items and distributions.

The above results are consistent with the hypothesis that LHF is a primary source of the lexical bias discussed in the literature. However, more conclusive evidence is needed; and specifically, experimental validation is required to confirm that the lexical items whose usage by LLMs we pinpointed as potentially LHF-induced are indeed preferred by human evaluators, thereby strengthening the causal link between LHF and LLMs' lexical choices.

3 Experimental Validation

At the core of the hypothesized link between LHF and LLMs' lexical choices is the idea that evaluators exhibit a subtle preference for certain lexical items, a preference that is in fact so slight that it has obscured this very link. However, when scaled up, these minor preferences for specific lexical items become

entrenched and ultimately manifested in the output generations of LLMs. To test this hypothesis, we created experimental items consisting of pairs of text variants. In each pair, one variant exhibits fewer words previously identified as potentially favored by LHF, while the other exhibits more such words, with all other factors held as equal as possible, including length and content. This design aims to isolate the effect of the presence of the lexical items identified above on evaluator judgments.

3.1 Experimental Setup

Creation of Experimental Items. The ideal test of the hypothesis would involve creating two random variants of a given abstract, repeating this for tens of thousands of pairs, collecting human evaluations for all these pairs, and then analyzing the ratings. The problem, however, is that detecting the hypothesized subtle effect experimentally under this approach would require an extraordinarily high number of ratings to achieve statistical significance. Thus, we opted for a procedure that increases the lexical differences between items, while at the same time maintaining comparable validity and being less resource-intensive.

For 50 randomly selected PubMed abstracts from 2020, we prompted GPT-4o to write summary notes ("The following text is an abstract from a scientific paper:\n\n{input_text}\n\nSummarize the abstract in keywords, separate keywords by commas."; see example on our GitHub). Using these summary notes as input, we then had Llama Instruct generate 500 abstracts (variants) for each item (Prompt: 'Based on the following keywords, write a 100-word abstract for a scientific journal article: "{line_of_keywords}." Reply with the abstract only.'), resulting in a total of 25 000 variants (50 random abstracts * 500 variants). We used GPT-4o to clean the abstracts (Prompt: 'The following text contains a scientific abstract, but sometimes further text:\n\n" {input_text}"\n\nPlease remove any irrelevant text, which can include titles, incomplete sentences, even a comment that an abstract is to follow (\"Abstract: \"). Output only the cleaned abstract.'). We controlled for length by filtering out candidates that were below 90 or above 110 words. It has been widely recognized that "delve" is an LLM-associated word [3,5,7,8,9] and a corresponding backlash against it [9]. Thus, we removed any variants containing any of the 21 most overused 'AI words' as discussed in [9], including words like "realm" and "groundbreaking". After applying these filters, our final set contained 8710 variants.

For these items (also part-of-speech tagged), we calculated a score to measure a word's potential to have been favored by LHF ("LHF-Score"). Using the lexical items identified in Section 2 as potentially promoted by LHF, we assigned a score to each variant by summing occurrences of these items, weighted by their relative rate of increase. This weighting reflects the idea that a single usage of a term like "revolutionize_VERB", which experienced an increase of +1160%, is probably more indicative of the influence of LHF than using a term like "of_ADP", which saw an increase of only 2%. As such, the score focuses on relative changes: A 100% shift from 1 to 2 occurrences of a given word should be treated the same as a shift from 1000 to 2000 occurrences in that same token span.

The LHF-Score for a sequence is the sum of LHF-Scores for each token (w). The LHF-Score for a given token is its increase in percent between Llama Base (B) and Llama Instruct (I), divided by one thousand (for ease of interpretability); "opm" stands for occurrences per million and is just the frequency of a token divided by the total number of tokens (N), multiplied by one million.

$$\begin{aligned}
\text{LHF-Score}(S) &= \sum_{i=1}^{n} \text{LHF-Score}(w_i) \\
&\textit{where} \\
\text{LHF-Score}(w) &= \frac{1}{1000} \cdot \left(\frac{\text{opm}_I(w) - \text{opm}_B(w)}{\text{opm}_B(w)} \times 100 \right) \\
&\textit{where} \\
\text{opm}(w) &= \frac{\text{count}(w)}{N} \times 10^6
\end{aligned} \tag{1}$$

An LHF-Score was calculated for all 8710 variants generated for the 50 summarized abstracts. For each of the 50 abstracts, we calculated the difference between the variant with the lowest LHF-Score and the one with the highest LHF-Score. We then selected the Top 30 abstract pairs with the largest Deltas while ensuring that the pair of variants were length-matched (in two cases, a length match was difficult, and we took the runners-up). The following hypothetical example between Sequence 1 and Sequence 2 illustrates how the LHF-Scores were calculated. The LHF-Score Delta is 0.31 (the score is calculated on lemmata and part-of-speech, which are omitted below for simplicity). A real example can be found on our GitHub.

(1)	*This*	*is*	*an*	*intricate*	*example*	*full*	*of*	*complex*	*words*	*(SUM)*
	0.03	0	0	0.36	0.03	0	0	0.2	0	(=0.44)

(2)	*This*	*is*	*a*	*baseline*	*example*	*free*	*from*	*these*	*words*	*(SUM)*
	0.03	0	0	0	0.03	0	0	0.07	0	(=0.13)

For the 30 selected items, the average LHF-Score for the variants with many of the lexical items identified in Section 2 is 7.2 (average length: 105 words), and the average LHF-Score for the variants with the fewest such items is 1.7 (average length: 104 words). The complete set of experimental item pairs is available on our GitHub repository. A small number of the words identified by the procedure above do not seem likely to have been promoted by LHF, such as "radar" (see Section 5). This introduces noise into the experiment. For instance, one variant of an abstract might include "radar", resulting in a higher LHF-Score, even though the in- or exclusion of such a word is unlikely to affect human preference between the two variants. Such cases weaken the statistical power of the experiment and increase the risk of a false negative outcome (the beta rate), thereby favoring the null hypothesis [49]. We anticipate this effect to be minor, however, given that the majority of lexical items previously identified do seem plausibly the sort that are potentially promoted by LHF.

Participants. We recruited 400 participants (231 female, 169 male; average age: 30.1 years, standard deviation: 9.8) through Prolific (www.prolific.com). Tech companies often recruit LHF workers from the Global South [50,51,4,52]. To more closely emulate the process by which LLMs are trained, we recruited participants from countries in the Global South where English is an official or widely used language (see Appendix A for a full list of countries). 90% of our participants were from Africa and 10% were from Southeast Asia. Participants were compensated at a rate equivalent to an average of $15 per hour.

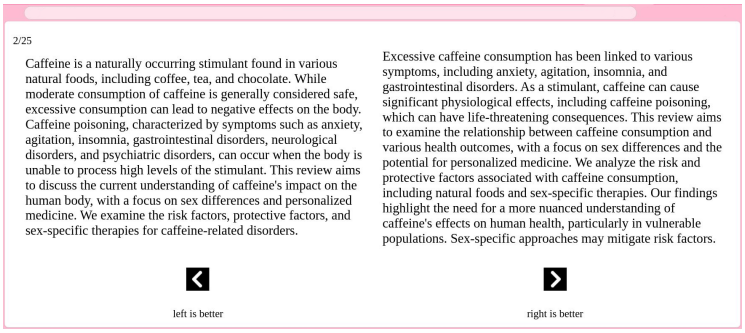

Fig. 2. The rating interface for our experiment.

The Task. The task began with IRB information (full instructions can be found on our GitHub), followed by an introduction to the task ("In the following, you will read a series of research summaries, with two alternatives next to each other. Please express which alternative you overall prefer. Some of the items are hard, do the best you can!", with an example as per Figure 2), including an example to familiarize participants with the process (for general best practices of experimental design, we followed [53] and [54]). Each participant rated 25 pairs of text variants, consisting of 20 critical item pairs (in random order), one calibration item at the beginning of the survey (where one variant was deliberately poor), two randomly interspersed "gotcha" items (which contained mid-sequence, "This is not a real item, please click on the left button"; cf. [54,55]), and two randomly interspersed items to assess language proficiency, similar to the calibration item. For each item, the left-right positioning of the abstracts was randomly flipped to avoid positional bias [56,57]. We did not include fillers, as the differences between the variants were subtle, and we were not concerned that participants would guess the purpose of the study.

Exclusions. To ensure high-quality data, which is crucial for statistical power [58], we applied exclusions. Only participants who completed 10 or more of the 25 items were included in the analysis (11 participants excluded). Participants who failed to correctly answer both "gotcha" items were also excluded from

the analysis (158 participants excluded). The literature reports that (225 ms + 25ms * character length of an item) is a good approximation of the minimum time physically required to read text [59]. To account for skimming or decisions made on the basis of reading only part of each abstract, we used a less strict threshold, excluding only ratings completed in less than 40% of this minimum time. Participants were warned if they responded more quickly than this. If a participant fell below this threshold on 5 or more items, all of their ratings were excluded from the analysis (18 additional participants excluded; many of the participants who failed the "gotcha" items would also have been excluded by this speed criterion). After exclusions, we retained 4039 ratings (out of a maximum of 8000 ratings: 400 participants * 20 ratings each), averaging about 135 ratings per item pair (minimum: 125 ratings). An exclusion rate of 46.8% is in line with previous work [60,61,62,63,64].

3.2 Analyses

The null hypothesis is that participants' choices between the high and low LHF-Score abstracts do not diverge from what one would expect when flipping a fair coin. The relevant alternative hypothesis is that participants show a preference for variants containing more of the words identified previously as potentially promoted by LHF – i.e., variants with a high LHF-score. For categorical, binary preference data like ours, where observations are tested against an expected baseline, a chi-square test is an excellent choice [49]. This is our main analysis. Additionally, we provide descriptives for the 30 item pairs, and we perform a mixed linear regression analysis to account for random effects. Our model includes the intercept as a fixed effect and participant and item as random effects.

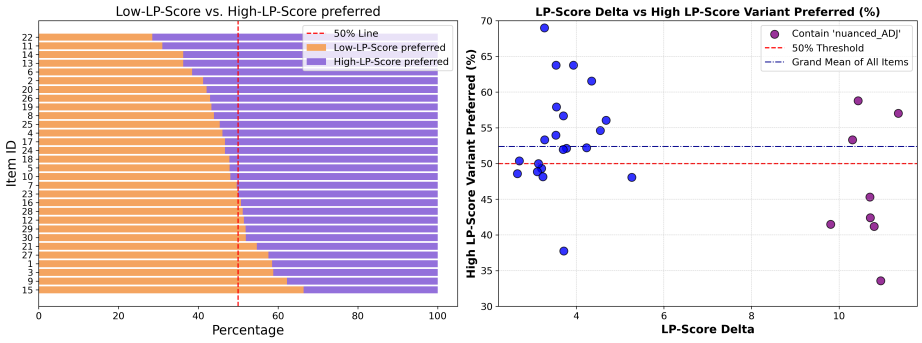

Fig. 3. (a) Experimental results: Preferences between low LHF-Score variant vs. high LHF-Score variant, for the 30 items. (b) Participant preferences for pairs with different LHF-Score Deltas. Each dot represents the mean preference for one of 30 abstract pairs. High LHF-Score Delta pairs contained "nuanced_ADJ."

3.3 Results

Overall, participants exhibited a highly significant preference for variants with a high LHF-Score over variants with a low LHF-Score (52.4% to 47.6%; $\chi^2 = 9.4, p < 0.01$). This trend is consistent across items, as confirmed by the regression model and the low variance observed across items (also see Figure 3). The mixed-effects model (REML, $N = 4038$, log-likelihood $= -2903.53$) revealed a significant intercept ($\beta = 0.524, z = 33.20, p < 0.001$), with low variance across items ($\sigma^2_{\text{item}} = 0.006$) and low to moderate variance across users ($\sigma^2_{\text{user}} = 0.104$). Based on these findings, we reject the null hypothesis and accept the alternative hypothesis: participants systematically and significantly prefer variants containing more of the items identified in Section 2 as words whose use by LLMs was likely promoted by LHF.

Although we did not initially intend to analyze abstracts containing any particular word, we noticed that sentence pairs in which the high RP-Score abstract contains the adjective "nuanced" had a substantially higher LHF-Score Delta (Figure 3 (b)). Further, the average preference for the high LHF-Score variant is markedly lower for items containing "nuanced" (46.6%) compared to sentence pairs without it (54.5%). It could be that items containing "nuanced" stuck out to participants, leading them to disprefer those items, similar to what has been observed with text that includes "delve" [9]. Additional data is needed to substantiate this interpretation, however.

4 Discussion

It has been well established *that* Large Language Models output certain words more frequently than a human baseline [3,4,5,6,7,8,9]. Our research advances the discourse by addressing the *why*, providing evidence that Learning from Human Feedback could be a primary source of this lexical overuse. We have identified lexical entries that models trained on LHF use considerably more than models without LHF training and then shown that texts containing many of these words are preferred to texts with fewer of them.

Furthermore, there is reason to think that the words used more by Llama Instruct than by Llama Base are also the sorts of words overused by LLMs compared to humans. To probe this connection to human language use, we extracted the lexical entries discussed in the academic literature on lexical overrepresentation [4,5,6,7,8,9]. This resulted in a list of 32 lexical entries (see Appendix A). We observe that 28 of these are also present in our Llama Base vs. Llama Instruct list. Thus, almost all of the words that researchers have identified as overrepresented in LLM-generated text compared to human-generated text appear more in the outputs of Llama Instruct than Llama Base. And as we have shown experimentally, these words are also favored by human evaluators, lending credibility to the hypothesis that the overuse of certain words by LLMs (relative to human usage) is at least partly the product of LHF. Our work therefore substantiates the previously speculative link between lexical overrepresentation and LHF.

It remains to be seen whether it is the demographics of the human evaluators or something about the feedback task they are engaged in that explains why they favor the sorts of words under discussion here. One notable observation is that LHF workers tend to be young, and almost all of the words overrepresented in LLM-generated text relative to human-generated text were already increasing in usage before the advent of LLMs [8]. Taken together, these facts suggest that lexical overuse in LLMs might be a form of normal intergenerational language change [65], albeit an accelerated one, wherein the preferences of younger generations are propagated in LLMs. This aligns with observations that young people tend to prefer AI-generated output over human-produced output [66].

LHF workers are also typically located in the Global South [50,51], whereas criticism of the increased usage of words like "delve" has predominantly originated from the Global North. Most of the academic research on the topic, such as [4,5,6,8,9], has been conducted at institutions based in the Global North. Some have speculated that the words overrepresented in LLM outputs might be more common in the dialects of English spoken by these LHF workers [14,15], though follow-up work has not yet substantiated this conjecture [9].

It is also possible that it is the nature of the LHF task that is responsible instead. Perhaps human evaluators, skimming quickly through unfamiliar text, rely on the presence of certain words as a proxy for quality. It was shown that human evaluators tend to prioritize style over content [67], which may explain why evaluators treat certain words as indicative of good outputs. In that case, the lexical preferences baked into LLMs through LHF might simply be task-driven. Discriminating between these explanations – that is, determining whether age, geographic location, dialect, or task features lead LHF workers to favor particular words – requires future research.

5 Limitations

This work has several limitations. First, our analysis is restricted to Meta's Llama. Broader validation would require access to base and instruction-tuned model variants from other LLM developers (such as OLMo or Falcon). Our analysis also focuses on English. Expanding this work to other languages would be valuable. Furthermore, while our dataset contains approximately 2m tokens per model, future work could scale this up. A likely artifact of the corpus size is the occasional identification of lexical items that are not commonly cited as overused by LLMs. For instance, the Instruct model uses the item "radar_NOUN" considerably more often than the Base model (+2590%). A qualitative analysis of the dataset, however, helps to make sense of this result: several PubMed abstracts in our sample discuss "radar_NOUN", and the Instruct model incorporates this into its continuations, whereas the Base model does not. Thus, scaling our procedure could improve the results.

Potential language confounds in the experimental items might have impacted our results. While we controlled for abstract length, other distinctive linguistic features of LLM-generated text, such as specific syntactic structures or stylistic

elements (e.g., "It's not about [X], it's about [Y]" [73]), might correlate with the presence of the words that we have identified, unknowingly contributing to higher preference ratings. A qualitative inspection of the item pairs did not reveal any clear patterns of such confounding features, but the possibility cannot be entirely ruled out. Furthermore, although our experimental procedure aimed to emulate the task situation of LHF workers, it did so imperfectly, as we cannot perfectly simulate their working conditions for both ethical and practical reasons. Lastly, while our experimental results clearly bear on the existing discourse about lexical biases, the connection to human language use remains somewhat preliminary. Further strengthening this connection would yield still further support for the hypothesis that LHF is at least partly responsible for lexical overuse in LLM outputs compared to human-generated text.

6 Conclusion

LHF is known to be a useful tool for aligning the outputs of LLMs more closely with human expectations. Our results, however, suggest that an accidental byproduct of such alignment efforts is lexical overuse. Does the overuse of particular words by LLMs constitute a failure of alignment? And should developers intervene to reduce the prevalence of these words? The answers to both questions depend on whose lexical preferences LLMs ought to reflect. Our research suggests that these models are making lexical choices that align with the preferences and expectations of LHF workers; but these same lexical choices may not satisfy consumers unhappy with LLMs' overuse of words like "delve."

If intervention is desired, our procedure offers a straightforward way of identifying potential cases of lexical overuse. While some manual verification is still needed, the procedure effectively identifies many of the most extreme instances of potential overuse. Importantly, our findings also highlight one place where interventions could be targeted: LHF datasets. Different strategies could be employed. For instance, developers and data scientists could diversify the workforce of human evaluators providing feedback for LHF [15], or datasets could be adjusted post-collection to ensure greater balance.

While we leave open the question of whether intervention is necessary, we note a shift in the dynamics of language change: Workers from the Global South are now influencing the language of language technologies, which are subsequently deployed globally. In the past, changes have predominantly flowed in the opposite direction [50,68]. However, those who wield this linguistic influence are in positions of economic precarity rather than positions of power.

Finally, our research contributes to the growing body of work on explainable AI [28,29,30]: Through systematic investigation, meaningful insights into the workings of artificial neural networks can be gained (see also discussion in [69]). However, a key difficulty for such research is the lack of transparency surrounding LLM development [24]. This includes lack of process transparency, as all major tech companies obscure the details of their LHF procedures, arguably in part to avoid scrutiny of poor working conditions for human evaluators, who

are frequently underpaid and stressed [70,71,72]. Lack of data transparency remains an issue as well, with many LHF datasets not being publicly available. These failures of transparency are worrisome in light of the significant impact that language technology has on global language usage. By facilitating insights like those presented here, publicizing information about model training can aid efforts to align LLMs more closely with human expectations.

References

1. Koppenburg, P.: Tweet on 01 April 2024. https://x.com/PKoppenburg/status/1774757167045788010, last accessed 2024/08/12
2. Nguyen, J.: Tweet on 30 March 2024. https://x.com/JeremyNguyenPhD/status/1774021645709295840, last accessed 2024/08/12
3. Shapira, P.: Delving into "delve". https://pshapira.net/2024/03/31/delving-into-delve/, last accessed 2024/09/21
4. Gray, A.: ChatGPT "contamination": Estimating the prevalence of LLMs in the scholarly literature. arXiv preprint arXiv:2403.16887 (2024)
5. Kobak, D., González Márquez, R., Horvát, E.-Á., Lause, J.: Delving into ChatGPT usage in academic writing through excess vocabulary. arXiv preprint arXiv:2406.07016 (2024)
6. Liang, W. et al.: Mapping the increasing use of LLMs in scientific papers. arXiv preprint arXiv:2404.01268 (2024)
7. Liu, J., Bu, Y.: Towards the relationship between AIGC in manuscript writing and author profiles: Evidence from preprints in LLMs. arXiv:2404.15799 (2024)
8. Matsui, K.: Delving into PubMed Records: Some Terms in Medical Writing Have Drastically Changed after the Arrival of ChatGPT. medRxiv (2024)
9. Juzek, T.S., Ward, Z.B.: Why Does ChatGPT "Delve" So Much? Exploring the Sources of Lexical Overrepresentation in Large Language Models. In Proceedings of the 31st International Conference on Computational Linguistics (pp. 6397-6411). https://doi.org/10.48550/arXiv.2412.11385 (2025)
10. Degaetano-Ortlieb, S., Teich, E.: Using relative entropy for detection and analysis of periods of diachronic linguistic change. In: Proc. 2nd Joint SIGHUM Workshop, pp. 22–33 (2018)
11. Degaetano-Ortlieb, S., Kermes, H., Khamis, A., Teich, E.: An information-theoretic approach to modeling diachronic change in scientific English. In: From Data to Evidence in English Language Research, pp. 258–281. Brill, Leiden (2018)
12. Bizzoni, Y., Degaetano-Ortlieb, S., Fankhauser, P., Teich, E.: Linguistic variation and change in 250 years of English scientific writing: A data-driven approach. Front. Artif Intell. 3(73) (2020)
13. Menzel, K.: Medical discourse in Late Modern English: Insights from a multidisciplinary corpus of scientific journal articles. In: Corpus Pragmatic Studies on the History of Medical Discourse, pp. 79–104. John Benjamins, Amsterdam (2022)
14. Hern, A.: TechScape: How cheap, outsourced labour in Africa is shaping AI English. https://www.theguardian.com/technology/2024/apr/16/techscape-ai-gadgest-humane-ai-pin-chatgpt, last accessed 2024/08/12
15. Sheikh, H.: Why does ChatGPT use “Delve” so much? Mystery Solved. https://hesamsheikh.substack.com/p/why-does-chatgpt-use-delve-so-much, last accessed 2025/01/14

16. Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. In: Adv. Neural Inf. Process. Syst. (30) (2017)
17. Ziegler, D.M., Stiennon, N., Wu, J., Brown, T.B., Radford, A., Amodei, D., Christiano, P., Irving, G.: Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593 (2019)
18. Rafailov, R. et al.: Direct preference optimization: Your language model is secretly a reward model. In: Adv. Neural Inf. Process. Syst. (36) (2024)
19. He, Z., Guo, S., Rao, A., Lerman, K.: Whose Emotions and Moral Sentiments Do Language Models Reflect? arXiv preprint arXiv:2402.11114 (2024)
20. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? In: Proc. 2021 ACM Conf. on Fairness, Accountability, and Transparency, pp. 610–623 (2021)
21. Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., Hashimoto, T.: Whose opinions do language models reflect? In: Int. Conf. on Machine Learning (ICML), pp. 29971–30004 (2023)
22. Durmus, E. et al.: Towards measuring the representation of subjective global opinions in language models. arXiv preprint arXiv:2306.16388 (2023)
23. Ouyang, L. et al.: Training language models to follow instructions with human feedback. In: Adv. Neural Inf. Process. Syst. (35), 27730–27744 (2022)
24. Bommasani, R. et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)
25. Geng, M., Chen, C., Wu, Y., Chen, D., Wan, Y., Zhou, P.: The impact of large language models in academia: from writing to speaking. arXiv preprint arXiv:2409.13686 (2024)
26. Yakura, H., Lopez-Lopez, E., Brinkmann, L., Serna, I., Gupta, P., Rahwan, I.: Empirical evidence of large language model's influence on human spoken communication. arXiv preprint arXiv:2409.01754 (2024)
27. Zamaraeva, O., Flickinger, D., Bond, F., Gómez-Rodríguez, C.: Comparing LLM-generated and human-authored news text using formal syntactic theory. arXiv preprint arXiv:2506.01407 (2025)
28. Sculley, D. et al.: Hidden technical debt in machine learning systems. In: Adv. Neural Inf. Process. Syst. (28) (2015)
29. Zhao, H. et al.: Explainability for large language models: A survey. ACM Transactions on Intelligent Systems and Technology 15(2), 1–38 (2024)
30. Cambria, E. et al.: XAI meets LLMs: A survey of the relation between explainable AI and large language models. arXiv preprint arXiv:2407.15248 (2024)
31. Zhang, X., Xiong, W., Chen, L., Zhou, T., Huang, H., Zhang, T.: From lists to emojis: How format bias affects model alignment. arXiv:2409.11704 (2024)
32. Erdocia, I., Migge, B., Schneider, B.: Language is not a data set—Why overcoming ideologies of dataism is more important than ever in the age of AI. J. Sociol. (2024)
33. Dubey, A. et al.: The LLaMA 3 herd of models. arXiv:2407.21783 (2024)
34. Wolf, T. et al.: Transformers: State-of-the-Art Natural Language Processing. arXiv preprint arXiv:1910.03771 (2020)
35. Hugging Face Team: Open LLM Leaderboard. (2024). https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard
36. Allen Institute for AI. (2024). OLMo 2. https://allenai.org/olmo
37. Technology Innov. Inst. (2024). Falcon 3. https://falconllm.tii.ae/falcon3/index.html
38. Juzek, T.S., Ward, Z.B.: Supplementary materials for "Word overuse and alignment in large language models: The influence of learning from human feedback". OSF, https://osf.io/4nvjk https://doi.org/10.17605/OSF.IO/4NVJK (2025)

39. Python Software Foundation: Python 3. https://www.python.org/, accessed 2024
40. National Library of Medicine: PubMed Database. https://pubmed.ncbi.nlm.nih.gov/, last accessed 2024/11/24
41. Achiam, J. et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
42. OpenAI: OpenAI Python API. Version 1.57. https://platform.openai.com/docs/, last accessed 2025/01/18
43. Montani, I., Honnibal, M., Boyd, A., Van Landeghem, S., Peters, H.: explosion/spaCy: v3.7.2: Fixes for APIs and requirements. Zenodo. https://doi.org/10.5281/zenodo.10009823 (2023)
44. Google: Google Books Ngram Viewer. https://books.google.com/ngrams/, last accessed 2025/01/02
45. Lavergne, T., Urvoy, T., Yvon, F.: Detecting fake content with relative entropy scoring. Pan 8(4), pp. 27–31 (2008)
46. Chakraborty, S. et al.: On the possibilities of AI-generated text detection. arXiv preprint arXiv:2304.04736 (2023)
47. Mitchell, E., Lee, Y., Khazatsky, A., Manning, C.D., Finn, C.: DetectGPT: Zero-shot machine-generated text detection using probability curvature. In: Int. Conf. on Machine Learning (ICML), pp. 24950–24962 (2023)
48. Huang, Y. et al.: MAGRET: Machine-generated Text Detection with Rewritten Texts. In: Proc. COLING 2025, pp. 8336–8346 (2025)
49. Haslwanter, T.: An Introduction to Statistics with Python. Springer, CH (2016)
50. Kwet, M.: Digital colonialism: US empire and the new imperialism in the Global South. Race Class 60(4), 3–26 (2019)
51. Perrigo, B.: Exclusive: OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic. Time Magazine (18) (2023)
52. Rohde, F. et al.: Broadening the perspective for sustainable artificial intelligence: Sustainability criteria and indicators for Artificial Intelligence systems. Curr. Opin. Environ. Sustain. (66), 101411 (2024)
53. Cowart, W.: Experimental Syntax. Sage, Thousand Oaks (1997)
54. Berinsky, A.J., Margolis, M.F., Sances, M.W.: Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. Am. J. Polit. Sci. 58(3), 739–753 (2014)
55. Maniaci, M.R., Rogge, R.D.: Caring about carelessness: Participant inattention and its effects on research. J. Res. Pers. (48), 61–83 (2014)
56. Friedman, H.H., Herskovitz, P.J., Pollack, S.: The biasing effects of scale-checking styles on response to a Likert scale. In: Proc. Amer. Stat. Assoc. Conf. on Survey Research Methods (792), pp. 792–795 (1994)
57. Chyung, S.Y., Kennedy, M., Campbell, I.: Evidence-based survey design: The use of ascending or descending order of Likert-type response options. Perform. Improv. 57(9), 9–16 (2018)
58. Mahowald, K., Graff, P., Hartman, J., Gibson, E.: SNAP judgments: A small N acceptability paradigm (SNAP) for linguistic acceptability judgments. Language 92(3), 619–635 (2016)
59. Häussler, J., Juzek, T.: Hot topics surrounding acceptability judgement tasks. In: Featherston, S., Hörnig, R., Steinberg, R., Umbreit, B., Wallis, J. (eds.) Linguistic Evidence 2016: Empirical, Theoretical, and Computational Perspectives. University of Tübingen, Tübingen. https://doi.org/10.15496/publikation-19039 (2017)
60. Downs, J.S., Holbrook, M.B., Sheng, S., Cranor, L.F.: Are your participants gaming the system? Screening Mechanical Turk workers. In: Proc. SIGCHI Conf. on Human Factors in Computing Systems, pp. 2399–2402 (2010)

61. Zhu, D., Carterette, B.: An analysis of assessor behavior in crowdsourced preference judgments. In: SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation, pp. 17–20 (2010)
62. Kazai, G., Kamps, J., Milic-Frayling, N.: Worker types and personality traits in crowdsourcing relevance labels. In: Proc. 20th ACM Int. Conf. on Information and Knowledge Management, pp. 1941–1944 (2011)
63. Thomas, K.A., Clifford, S.: Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. Comp. Hum. Behav. (77), 184–197 (2017)
64. Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B., Allahbakhsh, M.: Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. ACM Comput. Surv. 51(1), 1–40 (2018)
65. Labov, W.: Principles of Linguistic Change, vol. 3: Cognitive and Cultural Factors. Wiley, Hoboken (2011)
66. Young, J. et al.: The Role of AI in Peer Support for Young People: A Study of Preferences for Human-and AI-Generated Responses. In: Proc. CHI Conf. on Human Factors in Computing Systems, pp. 1–18 (2024)
67. Wu, M., Aji, A.F.: Style Over Substance: Evaluation Biases for Large Language Models. In Proc. COLING 2025, pp. 297–312. Association for Computational Linguistics, Abu Dhabi, UAE. https://aclanthology.org/2025.coling-main.21/ (2025)
68. hMensa, P.A.: Artificial intelligence and the future of sociolinguistic research: An African contextual review. J. Socioling. (2024)
69. Templeton, A.: Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Anthropic (2024)
70. Toxtli, C., Suri, S., Savage, S.: Quantifying the invisible labor in crowd work. Proc. ACM Hum.-Comput. Interact. 5 (CSCW2), 1–26 (2021)
71. Roberts, J.: The Precarious Human Work Behind AI. https://www.accel.ai/anthology/2023/5/22/jyzu7sbpzyxufu5l1ekidxj0g7jafh, last accessed 2023
72. Novick, M.: A.I.'s Dirty Secret: It's Powered by Digital Sweatshops. https://change-links.org/a-i-s-dirty-secret-its-powered-by-digital-sweatshops/, last accessed 2023
73. Jim the AI Whisperer: How One Sentence Pattern Can Expose AI Writing. Medium. https://generativeai.pub/how-to-spot-ai-writing-with-one-sentence-pattern-8aa5b3ec5a63, accessed 2024/12

A Appendix

Permitted Countries: Bangladesh, Belize, Botswana, Cameroon, Ethiopia, Fiji, Gambia, Ghana, Guyana, Indonesia, Kenya, Liberia, Malawi, Malaysia, Mauritius, Micronesia, Montserrat, Namibia, Nigeria, Pakistan, P. N. G., Philippines, S. Africa, Sri Lanka, Swaziland, Tanzania, Uganda, Zambia, Zimbabwe.

Words from overuse literature: advancements, aligns, boasts, commendable, comprehending, crucial, delve, delved, delves, delving, emphasizing, garnered, groundbreaking, intricacies, intricate, invaluable, meticulous, meticulously, notable, noteworthy, pivotal, potential, realm, showcases, showcasing, significant, strategically, surpasses, surpassing, underscore, underscores, underscoring.