How Learning from Human Feedback Influences the Lexical Choices of Large Language Models

Anonymous ACL submission

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. 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 than ever before by experimentally emulating the LHF procedure and demonstrating that participants systematically prefer text variants that include certain words. To address the overuse of such words, developers now have a clear starting point: LHF datasets. This lexical overuse may 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. Possible causes of these divergences include demographic differences and/or features of the feedback solicitation task. Our work challenges the view of artificial neural networks as impenetrable black boxes and emphasizes the critical importance of both data and procedural transparency in alignment research.

1 Introduction

Following the arrival of Large Language Models (LLMs), observers were quick to note their tendency to overproduce certain lexical entries (Koppenburg, 2024; Nguyen, 2024; Shapira, 2024; Gray, 2024; Kobak et al., 2024; Liang et al., 2024; Liu and Bu, 2024; Matsui, 2024; Juzek and Ward, 2025). Much of the discourse centered on Scientific and academic English, focusing on words such as “delve”, “intricate”, and “realm”. Moreover, while changes in Scientific English over decades and centuries are well-documented (Degaetano-Ortlieb and Teich, 2018; Degaetano-Ortlieb et al., 2018; Bizzoni et al., 2020; Menzel, 2022), 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, *that* certain lexical biases exist in LLMs has been established, with evidence demonstrating their influence on human language usage. However, the question of *why* this lexical overrepresentation arises remains open. While some have pointed to Learning from Human Feedback (LHF) as a significant contributor to these lexical choices (Hern, 2024; Sheikh, 2024), conclusive evidence to substantiate this claim 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; Christiano et al. 2017; Ziegler et al. 2019), though a more recent and increasingly popular form of LHF is Direct Preference Optimization (DPO) (Rafailov et al., 2024). 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” (He et al., 2024), is a major challenge in AI (Bender et al., 2021; Santurkar et al., 2023; Durmus et al., 2023). 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 (Ouyang et al., 2022). However, researching the effects of LHF is difficult due to lack of transparency surrounding the procedures and datasets used in model development (Bommasani et al., 2021), including for many popular open models.

Very broadly, our research aims to investigate how technology and language interact (Erdocia et al., 2024). The present study addresses the po-

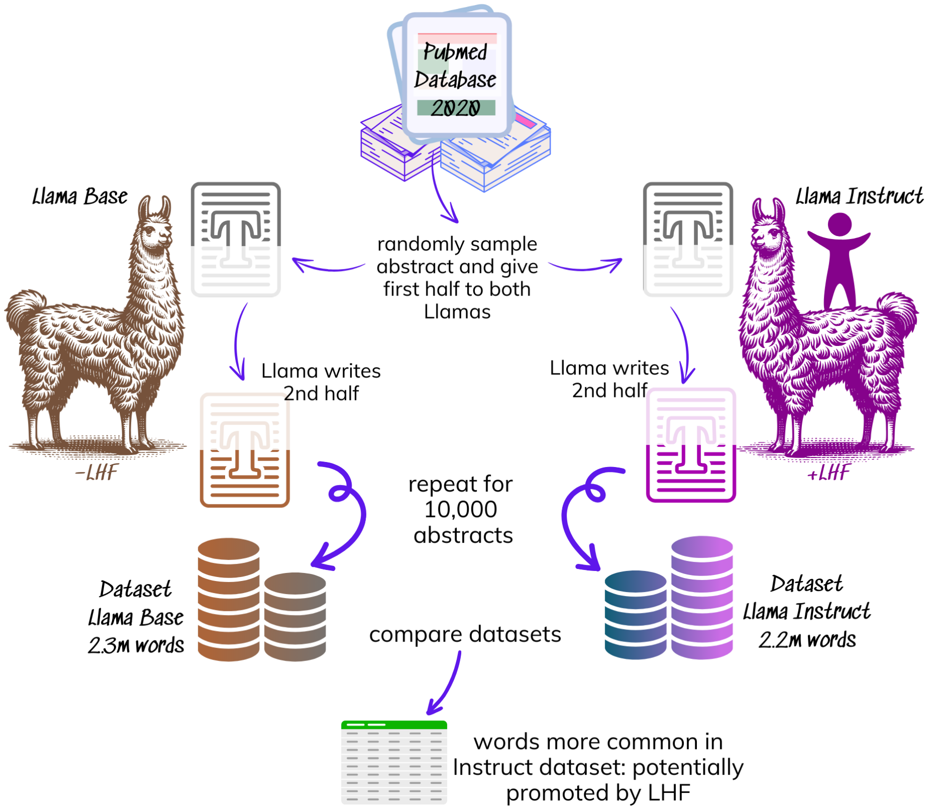

Pubmed Database 2020
Llama Base
-LHF
Llama Instruct
+LHF
randomly sample abstract and give first half to both Llamas
Llama writes 2nd half
Llama writes 2nd half
repeat for 10,000 abstracts
Dataset Llama Base 2.3M words
Dataset Llama Instruct 2.2M words
compare datasets
words more common in Instruct dataset: potentially promoted by LHF

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

tential 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 has possible applications in industry, as it can aid efforts to mitigate the most extreme cases of lexical overrepresentation and align models more closely with general language usage (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 the most rigorous test to date of the hypothesis that LHF significantly shapes LLMs' lexical choices (Section 3). Finally, we highlight the fact that LLMs are not impenetrable black boxes: meaningful insights into their behavior can indeed be gained. We also explore whether the lexical preferences we have identified are inherently problematic. The answer depends in part on the specific mechanisms through which these preferences arise, underscoring the importance of research on the sources of lexical overuse (Section 4).

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 likely 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 resulting generations. Here, we use Llama 3.2-3B Base and Llama 3.2-3B Instruct (Dubey et al., 2024) (via the Hugging Face Transformers library, Wolf et al. 2020), as the Llama family is, to our knowledge, the closest approximation to pure +/-LHF models available. Llama 3 makes use of Direct Preference Optimization. (Notably, OpenAI has not offered access to base models for years now.) While there are other differences between Llama Base and Llama Instruct (Dubey et al., 2024), the use of LHF to train Llama Instruct remains the main one (other differences include optimizing Instruct for tooling purposes and safety mitigations). This makes Llama well-suited for our purposes. All technical implementations described in this paper were carried out in Python (Python Software Foundation, 2024).

Since most of the academic discourse on LLMs has focused on Scientific English, we chose this domain for our study, though the procedure we present is transferable to other domains. Here the procedure is applied to abstracts from PubMed from 2020 (National Library of Medicine, 2024), 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 (Achiam et al., 2023; OpenAI, 2025) (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 (Montani et al., 2023), 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, Google 2024). 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 in Appendix A for further entries and our anonymous GitHub for the full list (github.com/arizus/delve3).

This is a straightforward procedure for identifying lexical items that are likely preferred by an LLM (in this case, the Llama Instruct model) as a result of training with LHF. Many of the identified words have been discussed in the literature on the distinctive lexical choices of LLMs (see references in Section 1). However, the procedure also identifies lexical entries that are not known to be overused by LLMs and so are more difficult to interpret. For instance, the Instruct model uses the item "radar_NOUN" considerably more often than the Base model (an increase of 2590%). A qualitative examination 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.

Our procedure serves as a proof of concept that it is possible to automate the search for potentially LHF-induced lexical preferences. Our application of the procedure is limited to the domain of Scientific English and to corpora of about two million words each. Thus, scaling it could improve the results. It is important to keep in mind that 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. Virtually all 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 (813 out of 814). Thus, when it comes to the lexical items that distinguish LLM-generated text from human-generated text, the identification procedure in its current form is effective in picking out many of the most extreme cases.

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. Our insights could also inform the discourse on AI-generated text detection (Lavergne et al., 2008; Chakraborty et al., 2023; Mitchell et al., 2023; Huang et al., 2025), as such methods often rely on identifying atypical lexical items and distributions. Although the simplicity of the procedure might raise questions about its value, the degree of such bias observed in LLM outputs suggests that either no robust identification mechanisms were previously applied, or existing mechanisms have proven too weak. There is therefore a need for even basic procedures like the one presented here.

We believe the above results are consistent with the hypothesis that LHF is a major source of the lexical bias discussed in the literature. However, more evidence is needed to more conclusively support this hypothesis. 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 for each abstract ("The following text is an abstract from a scientific paper:\n\n{input_text}\n\nSummarize the abstract in keywords, separate keywords by commas."; an example output is provided in Appendix B). 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. 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 either below 90 words or above 110 words. There has been a widespread recognition that "delve" is an LLM-associated word (see references in Section 1) and a corresponding backlash against it (Juzek and Ward, 2025). Thus, we removed any variants containing any of the 21 most overused 'AI words' as discussed in (Juzek and Ward, 2025), including words like "realm" and "groundbreaking". After applying these filters, we retained a final set of 8710 variants.

For these items, which were also part-of-speech tagged, we calculated a score to measure a word's potential to have been favored by LHF ("LHF-Potential-Preference-Score", or simply "LP-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 a significant increase of +1160%, is probably more indicative of the influence of LHF than using a term like "of_ADP", which saw a much smaller increase of only 2%.

The LP-Score for a sequence is the sum of LP-Scores for each token (w). The LP-Score for a given token is its increase in percent between Llama Base (B) and Llama Instruct (I), divided by one thousand; "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.

$$\text{LP-Score}(S) = \sum_{i=1}^{n} \text{LP-Score}(w_i)$$

$$\textit{where}$$

$$\text{LP-Score}(w) =$$

$$\frac{1}{1000} \cdot \left(\frac{\text{opm}_I(w) - \text{opm}_B(w)}{\text{opm}_B(w)} \times 100 \right)$$

$$\textit{where}$$

$$\text{opm}(w) = \frac{\text{count}(w)}{\text{N}} \times 10^6$$

An LP-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 LP-Score and the one with the highest LP-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 LP-Scores were calculated. The LP-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 in Appendix C.

(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 LP-Score for the variants with many of the lexical items identified in Section 2 is 7.2 (average length: 105 words), and the average LP-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 anonymous GitHub repository. As discussed in Section 2, some of the words identified by the procedure above do not seem likely to have been promoted by LHF, such as "radar". This introduces noise into the experiment. For instance, one variant of an abstract might include "radar", resulting in a higher LP-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 (Haslwanter, 2016). 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). It has been claimed that tech companies often recruit LHF workers from the Global South (Kwet, 2019; Perrigo, 2023; Gray, 2024; Rohde et al., 2024). 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 D 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.

The Task. The task began with IRB information, followed by an introduction to the task, including an example to familiarize participants with the process (for general best practices of experimental design, we followed Cowart 1997 and Berinsky et al. 2014). An illustration of the interface can be found in Appendix E. Each participants 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. Berinsky et al. 2014; Maniaci and Rogge 2014), 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 (Friedman et al., 1994; Chyung et al., 2018). 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.

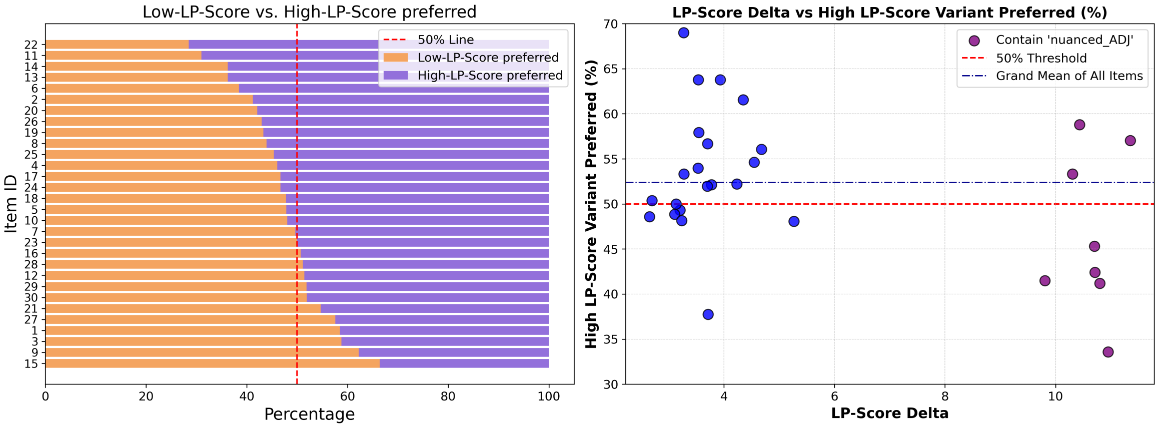


Figure 2: (a) Experimental results: Preferences between low LP-Score variant vs. high LP-Score variant, for the 30 items. (b) Participant preferences for pairs with different LP-Score Deltas. Each dot represents the mean preference for one of 30 abstract pairs. High LP-Score Delta pairs contained "nuanced_ADJ."

Exclusions. To ensure high-quality data, which is crucial for statistical power (Mahowald et al., 2016), 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). Häussler and Juzek (2017) report that (225 ms + 25ms * character length of an item) is a good approximation of the minimum time physically required to read text. 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 applying these 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). Our exclusion rate of 46.8% of the participants is in line with the percentages reported in the literature (Downs et al., 2010; Zhu and Carterette, 2010; Kazai et al., 2011; Thomas and Clifford, 2017; Daniel et al., 2018).

3.2 Analyses

The null hypothesis is that participants' choices between the high and low LP-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 LP-score. For categorical, binary preference data like ours, where observations are tested against an expected baseline, a chi-square test is an excellent choice of statistical test (Haslwanter, 2016). 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.

3.3 Results

Overall, participants exhibited a significant preference for variants with a high LP-Score over variants with a low LP-Score, with a highly significant 52.4% to 47.6% split ($\chi^2 = 9.4, p < 0.01$). This trend is consistent across items and is not driven by a small subset of items, as confirmed by the regression model and the low variance observed across items (also see Figure 2). 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 LP-Score Delta (Figure 2 (b)). Further, the average preference for the high LP-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” (Juzek and Ward, 2025). Additional data is needed to substantiate this interpretation, however.

4 Discussion and Conclusion

There is little doubt *that* Large Language Models exhibit lexical overuse – that is, that they output certain words more frequently than a human baseline (see references in Section 1). Our research advances the discourse by addressing the *why*, providing stronger evidence than ever before that Learning from Human Feedback could be a major 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 (Gray, 2024; Kobak et al., 2024; Liang et al., 2024; Liu and Bu, 2024; Matsui, 2024; Juzek and Ward, 2025). This resulted in a list of 32 lexical entries (see Appendix F). 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.

4.1 Broader impacts and concluding remarks

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 (Matsui, 2024). Taken together, these facts suggest that lexical overuse in LLMs might be a form of normal intergenerational language change (Labov, 2011), 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 (Young et al., 2024).

LHF workers are also typically located in the Global South, whereas criticism of the increased usage of words like “delve” has predominantly originated from 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 (Hern, 2024; Sheikh, 2024), though follow-up work has not yet substantiated this conjecture (Juzek and Ward, 2025). It is also possible that it is the nature of the LHF task, rather than demographic factors, that is responsible. Perhaps human evaluators, skimming quickly through unfamiliar text, rely on the presence of certain words as a proxy for quality. Wu and Aji (2025) showed that human evaluators tend to prioritize style over content, 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.

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 (and comparison with a human baseline) is still needed, the procedure effectively identifies many of the most extreme instances of potential overuse. Importantly, our findings also highlight where interventions should 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 (Sheikh, 2024), 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, the direction of influence has predominantly flowed in the opposite direction (Kwet, 2019; hMensa, 2024).

Finally, our research challenges the idea that artificial neural networks (ANNs) are impenetrable black boxes (Knight, 2017; Sculley et al., 2015). Through systematic investigation, meaningful insights into their workings can indeed be gained (see also discussion in Templeton 2024). However, a key difficulty for such research is the lack of transparency surrounding LLM development (Bommasani et al., 2021). 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 (Toxtli et al., 2021; Roberts, 2022; Novick, 2023). Lack of data transparency remains an issue as well, with 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.

5 Limitations

As noted in Section 2, the application of the procedure proposed here to identify potentially LHF-favored words is limited in both domain and size. The procedure should be scaled to domains other than Scientific English and well beyond the few millions words that we have analyzed. It is also important to keep in mind that 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 ("It's not about [X], it's about [Y]", the AI Whisperer 2024), 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.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Adam J Berinsky, Michele F Margolis, and Michael W Sances. 2014. Separating the shirkers from the workers? making sure respondents pay attention on self-administered surveys. *American journal of political science*, 58(3):739–753.

Yuri Bizzoni, Stefania Degaetano-Ortlieb, Peter Fankhauser, and Elke Teich. 2020. Linguistic variation and change in 250 years of english scientific writing: A data-driven approach. *Frontiers in Artificial Intelligence*, 3:73.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Souradip Chakraborty, Amrit Singh Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. 2023. On the possibilities of ai-generated text detection. *arXiv preprint arXiv:2304.04736*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Seung Youn Chyung, Megan Kennedy, and Ingrid Campbell. 2018. Evidence-based survey design: The use of ascending or descending order of likert-type response options. *Performance Improvement*, 57(9):9–16.

Wayne Cowart. 1997. *Experimental syntax*. Sage.

Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)*, 51(1):1–40.

Stefania Degaetano-Ortlieb, Hannah Kermes, Ashraf Khamis, and Elke Teich. 2018. An information-theoretic approach to modeling diachronic change in scientific english. In *From data to evidence in English language research*, pages 258–281. Brill.

Stefania Degaetano-Ortlieb and Elke Teich. 2018. Using relative entropy for detection and analysis of periods of diachronic linguistic change. In *Proceedings of the second joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature*, pages 22–33.

Julie S Downs, Mandy B Holbrook, Steve Sheng, and Lorrie Faith Cranor. 2010. Are your participants gaming the system? screening mechanical turk workers. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2399–2402.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Iker Erdocia, Bettina Migge, and Britta Schneider. 2024. Language is not a data set—why overcoming ideologies of dataism is more important than ever in the age of ai. *Journal of Sociolinguistics*.

Hershey H Friedman, Paul J Herskovitz, and Simcha Pollack. 1994. The biasing effects of scale-checking styles on response to a likert scale. In *Proceedings of the American statistical association annual conference: survey research methods*, volume 792, pages 792–795.

Google. 2024. Google books ngram viewer. Accessed: 2025-01-02.

Andrew Gray. 2024. Chatgpt" contamination": estimating the prevalence of llms in the scholarly literature. *arXiv preprint arXiv:2403.16887*.

Thomas Haslwanter. 2016. An introduction to statistics with python. *With applications in the life sciences. Switzerland: Springer International Publishing*.

Jana Häussler and Tom Juzek. 2017. Hot topics surrounding acceptability judgement tasks. In S. Featherston, R. Hörnig, R. Steinberg, B. Umbreit, and J. Wallis, editors, *Proceedings of Linguistic Evidence 2016: Empirical, Theoretical, and Computational Perspectives*. University of Tübingen, Tübingen.

Zihao He, Siyi Guo, Ashwin Rao, and Kristina Lerman. 2024. Whose emotions and moral sentiments do language models reflect? *arXiv preprint arXiv:2402.11114*.

Alex Hern. 2024. TechScape: How cheap, outsourced labour in Africa is shaping AI English. Accessed: 2024-08-12.

Patience Afrakoma hMensa. 2024. Artificial intelligence and the future of sociolinguistic research: An african contextual review. *Journal of Sociolinguistics*.

Yifei Huang, Jiuxin Cao, Hanyu Luo, Xin Guan, and Bo Liu. 2025. Magret: Machine-generated text detection with rewritten texts. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8336–8346.

Tom S Juzek and Zina B Ward. 2025. 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 (COLING 2025)*.

Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2011. Worker types and personality traits in crowdsourcing relevance labels. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1941–1944.

Will Knight. 2017. The dark secret at the heart of ai.

Dmitry Kobak, Rita González Márquez, Emőke-Ágnes Horvát, and Jan Lause. 2024. Delving into chatgpt usage in academic writing through excess vocabulary. *arXiv preprint arXiv:2406.07016*.

Patrick Koppenburg. 2024. Tweet on 01 april 2024. https://x.com/PKoppenburg/status/1774757167045788010. Accessed: 2024-08-12.

Michael Kwet. 2019. Digital colonialism: Us empire and the new imperialism in the global south. *Race & Class*, 60(4):3–26.

William Labov. 2011. *Principles of linguistic change, volume 3: Cognitive and cultural factors*, volume 3. John Wiley & Sons.

Thomas Lavergne, Tanguy Urvoy, and François Yvon. 2008. Detecting fake content with relative entropy scoring. *Pan*, 8(27-31):4.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, et al. 2024. Mapping the increasing use of llms in scientific papers. *arXiv preprint arXiv:2404.01268*.

Jialin Liu and Yi Bu. 2024. Towards the relationship between aigc in manuscript writing and author profiles: evidence from preprints in llms. *arXiv preprint arXiv:2404.15799*.

Kyle Mahowald, Peter Graff, Jeremy Hartman, and Edward Gibson. 2016. Snap judgments: A small n acceptability paradigm (snap) for linguistic acceptability judgments. *Language*, 92(3):619–635.

Michael R Maniaci and Ronald D Rogge. 2014. Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, 48:61–83.

Kentaro Matsui. 2024. Delving into pubmed records: Some terms in medical writing have drastically changed after the arrival of chatgpt. *medRxiv*, pages 2024–05.

Katrin Menzel. 2022. 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*, pages 79–104. John Benjamins.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pages 24950–24962. PMLR.

Ines Montani, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2023. explosion/spacy: v3.7.2: Fixes for apis and requirements. Version v3.7.2, Zenodo.

National Library of Medicine. 2024. PubMed Database. https://pubmed.ncbi.nlm.nih.gov/. Accessed: 2024-11-24.

Jeremy Nguyen. 2024. Tweet on 30 march 2024. https://x.com/JeremyNguyenPhD/status/1774021645709295840. Accessed: 2024-08-12.

Michael Novick. 2023. A.i.'s dirty secret: It's powered by digital sweatshops. Blog post.

OpenAI. 2025. *OpenAI Python API*. Version 1.57.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Billy Perrigo. 2023. Exclusive: Openai used kenyan workers on less than $2 per hour to make chatgpt less toxic. *Time Magazine*, 18:2023.

Python Software Foundation. 2024. Python 3.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Jennafer Roberts. 2022. The precarious human work behind ai. Blog post.

Friederike Rohde, Josephin Wagner, Andreas Meyer, Philipp Reinhard, Marcus Voss, Ulrich Petschow, and Anne Mollen. 2024. Broadening the perspective for sustainable artificial intelligence: sustainability criteria and indicators for artificial intelligence systems. *Current Opinion in Environmental Sustainability*, 66:101411.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. 2015. Hidden technical debt in machine learning systems. *Advances in neural information processing systems*, 28.

Philip Shapira. 2024. Delving into "delve". Accessed: 2024-09-21.

Hesam Sheikh. 2024. Why does chatgpt use "delve" so much? mystery solved. Accessed: 2025-01-14.

Adly Templeton. 2024. *Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet*. Anthropic.

Jim the AI Whisperer. 2024. How one sentence pattern can expose ai writing. Medium article.

Kyle A Thomas and Scott Clifford. 2017. Validity and mechanical turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior*, 77:184–197.

Carlos Toxtli, Siddharth Suri, and Saiph Savage. 2021. Quantifying the invisible labor in crowd work. *Proceedings of the ACM on human-computer interaction*, 5(CSCW2):1–26.

Thomas Wolf et al. 2020. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Minghao Wu and Alham Fikri Aji. 2025. Style over substance: Evaluation biases for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 297–312, Abu Dhabi, UAE. Association for Computational Linguistics.

Jordyn Young, Laala M Jawara, Diep N Nguyen, Brian Daly, Jina Huh-Yoo, and Afsaneh Razi. 2024. The role of ai in peer support for young people: A study of preferences for human-and ai-generated responses. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Dongqing Zhu and Ben Carterette. 2010. An analysis of assessor behavior in crowdsourced preference judgments. In *SIGIR 2010 workshop on crowdsourcing for search evaluation*, pages 17–20.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Top Potentially LHF-Favored Words

Table 1 gives the Top 20 words used more frequently by Llama Instruct than Llama Base, identified using the procedure presented in Section 2. The full list can be found on our GitHub.

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 for Llama Base (Ll-B) vs. Llama Instruct (Ll-I).

B Example of Abstract and AI-Generated Keywords for Summary

Example of original PubMed abstract: “Using a life course theory perspective, this qualitative descriptive study explored how Hispanic adolescent fathers view fatherhood, and how their perception of parenthood is shaped by critical life events. Hispanics are one of the largest ethnic groups, as well as one of the populations that is overrepresented in adolescent births in the United States. Despite this, Hispanic adolescent fathers are understudied and underrepresented in research. Participants were recruited from a community-based fatherhood program. Semi-structured interviews were conducted with Hispanic fathers, ages 16 years to 23 years. Participants conveyed their grief over fragmented family relationships and limited interactions with their own father. Some lived in hostile environments where they frequently experienced racism, discrimination, and neighborhood violence. The cumulative impact of these events resulted in substance use and emotional distress. Becoming a father was a sentinel event that helped resolve negative perceptions about fatherhood. Fatherhood also motivated participants towards a more productive, meaningful life.”

AI-generated keywords: “Hispanic, adolescent fathers, fatherhood, life course theory, qualitative descriptive study, critical life events, underrepresented, community-based program, semi-structured interviews, grief, family relationships, racism, discrimination, neighborhood violence, substance use, emotional distress, sentinel event, positive perceptions, meaningful life.”

C A Full Example: High- and Low-LP-Score Variants

For readability, words with an LP-score of >0.1 are highlighted in boldface, but part-of-speech is omitted. All items in both forms, with and without part-of-speech, can be found on our GitHub.

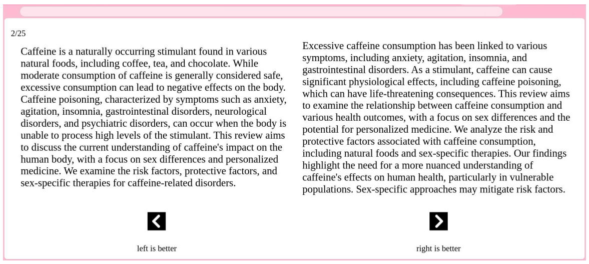
An example with a high LP-score: “In a transgenic mouse model of melanoma, we **investigated** the effects of glutamine supplementation on tumour growth and survival under conditions of nutrient deprivation. Glutamine supplementation enhanced tumour growth, but when combined with a BRAF inhibitor, reduced tumour growth and increased survival. Metabolomic analysis revealed increased αKG levels, **leading** to hypomethylation and H3K4me3 demethylation, promoting oncogenic pathways. Dietary intervention and **targeted** therapy **strategies targeting** these **epigenetic** modifications hold **promise** for melanoma treatment. **Furthermore**, **our** results **suggest** that glutamine supplementation may promote tumour growth, **potentially** through its role in αKG synthesis, **highlighting** the **need** for **nuanced** nutritional approaches in cancer treatment.” (100 words, LP-score: 12.6)

The following is the counterpart with a low LP-score: “This study **employed** a transgenic mouse model of melanoma to **investigate** the effects of glutamine supplementation on tumour growth and survival under conditions of nutrient deprivation. The model was treated with a BRAF inhibitor, a common **targeted** therapy for melanoma. Metabolomic analysis revealed increased αKG levels, **indicative** of glutamine metabolism, and associated with tumour growth and survival. Transcriptome analysis showed alterations in **epigenetic** marks, including hypomethylation and H3K4me3 modifications, in response to glutamine supplementation. These changes were correlated with activation of oncogenic pathways and improved tumour growth. Dietary intervention with glutamine also demonstrated enhanced tumour growth and survival in the model.” (101 words, LP-score: 2.1)

D Full List of Permitted Countries

Bangladesh, Belize, Botswana, Cameroon, Ethiopia, Fiji, Gambia, Ghana, Guyana, Indonesia, Kenya, Liberia, Malawi, Malaysia, Mauritius, Micronesia, Montserrat, Namibia, Nigeria, Pakistan, Papua New Guinea, Philippines, South Africa, Sri Lanka, Swaziland, Tanzania, Uganda, Zambia, Zimbabwe.

E Rating Interface



Caffeine is a naturally occurring stimulant found in various natural foods, including coffee, tea, and chocolate. While moderate consumption of caffeine is generally considered safe, excessive consumption can lead to negative effects on the body. Caffeine poisoning, characterized by symptoms such as anxiety, agitation, insomnia, gastrointestinal disorders, neurological disorders, and psychiatric disorders, can occur when the body is unable to process high levels of the stimulant. This review aims to discuss the current understanding of caffeine's impact on the human body, with a focus on sex differences and personalized medicine. We examine the risk factors, protective factors, and sex-specific therapies for caffeine-related disorders.

Excessive caffeine consumption has been linked to various symptoms, including anxiety, agitation, insomnia, and gastrointestinal disorders. As a stimulant, caffeine can cause significant physiological effects, including caffeine poisoning, which can have life-threatening consequences. This review aims to examine the relationship between caffeine consumption and various health outcomes, with a focus on sex differences and the potential for personalized medicine. We analyze the risk and protective factors associated with caffeine consumption, including natural foods and sex-specific therapies. Our findings highlight the need for a more nuanced understanding of caffeine's effects on human health, particularly in vulnerable populations. Sex-specific approaches may mitigate risk factors.

Figure 3: The rating interface for our experiment.

F List of words discussed in the literature on lexical overrepresentation in LLMs

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.