# Inclusivity in Large Language Models: Personality Traits and Gender Bias in Scientific Abstracts

Naseela Pervez[1] and Alexander J. Titus[1,2,3] ⋆

[1] Information Sciences Institute, University of Southern California
[2] Iovine and Young Academy, University of Southern California
[3] In Vivo Group

**Abstract.** Large language models (LLMs) are increasingly utilized to assist in scientific and academic writing, helping authors enhance the coherence of their articles. Previous studies have highlighted stereotypes and biases present in LLM outputs, emphasizing the need to evaluate these models for their alignment with human narrative styles and potential gender biases. In this study, we assess the alignment of three prominent LLMs—Claude 3 Opus, Mistral AI Large, and Gemini 1.5 Flash—by analyzing their performance on benchmark text-generation tasks for scientific abstracts. We employ the Linguistic Inquiry and Word Count (LIWC) framework to extract lexical, psychological, and social features from the generated texts. Our findings indicate that, while these models generally produce text closely resembling human-authored content, variations in stylistic features suggest significant gender biases. This research highlights the importance of developing LLMs that maintain a diversity of writing styles to promote inclusivity in academic discourse.

**Keywords:** Large Language Models (LLMs), Text Generation, Gender Bias, Linguistic Inquiry and Word Count (LIWC), Computational Linguistics

## 1 Introduction

Large Language Models (LLMs) have gained popularity in recent years for their performance on benchmark natural language processing (NLP) tasks including, but not limited to, text summarization, text generation, and question answering. LLMs have the ability to generate texts (stories) that are grammatically correct, and coherent and keep the readers engaged. It is likely that in the near future LLMs will be adopted as assistants for many writing tasks such as rewriting and improving the language of text (e.g. student essays, newspaper article, book writing). Scientific texts like abstracts, proposals, and journal publications use scientific jargon. However, due to the technical nature of scientific publications, to our knowledge, there are no full-text scientific articles that have been written by LLMs yet. In this paper we have proposed an evaluation framework that reflects the traits of LLM written text for scientific literature.

David C. Wyld et al. (Eds): NLAI, IBCOM, MLDS, GridCom, SCAI, ICCSEA, SPPR – 2024
pp. 01-16, 2024. - CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.142201

Although scientific papers are aimed at conveying a technique, research solution, or a research proposal, every author has a writing style. This writing style has an impact on their readership. In theoretical psychology, there are well-established frameworks that quantify a piece of writing by various features. These features are what can be called the "personality" of the text.

Like all texts, scientific texts have a personality and scientific authorship is male-dominated [1]. It has been established that male authors have a higher readership and citation count as well [2]. There are two aspects to this bias in research communities [3] - 1. the established stereotype that males are better researchers than females, and 2. the writing style of males (informative) is appreciated over the writing style of female researchers (descriptive) [4, 5].

Research communities have made efforts to reduce the gender gaps between males and females. This is reflected in the appointment of female faculty, promotion of male-female collaborations as well as high-prestige institutions hiring female researchers as project leads that have improved the representation of females in academia and improved the reach of female research [6]. It is therefore crucial to ensure that LLMs used for scientific writing do not create, or exacerbate, a gender bias based on the writing style.

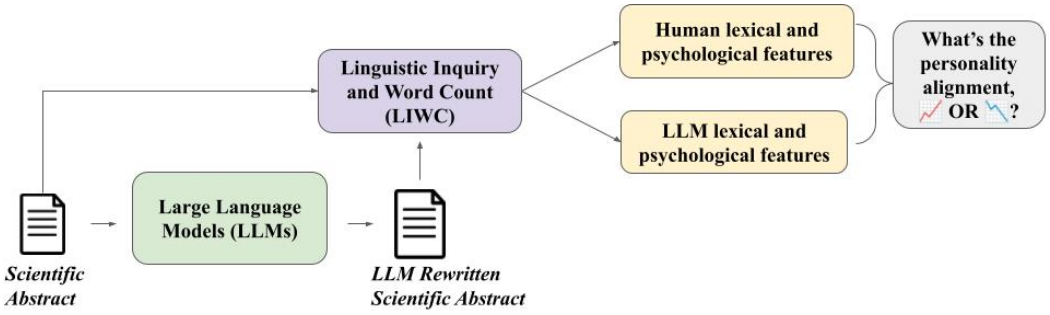


**Fig. 1.** Flowchart illustrating the comparison of LIWC features in scientific abstracts written by humans and rewritten by LLMs to assess personality alignment. This framework is adapted for male vs female comparison as well.

In this paper, we have focused on the following research questions:

1. When prompted to re-write a piece of scientific text, do LLMs maintain the narrative style of the text, implying that it maintains the author's personality?
2. Do LLMs alleviate or undermine the personality traits of a scientific text? Do they accentuate positive traits and diminish negative traits?

## 2 Related Work

The manner in which a narrative is constructed can significantly influence the impact and dissemination of scientific literature, with varying styles potentially affecting the learning process of readers [7, 8]. Gender differences in writing styles have been observed, with female authors often adopting a more intimate and engaged tone, while male authors tend to employ a more directive and informational approach [9]. These stylistic disparities can inadvertently contribute to gender and prestige biases within the research community, potentially impacting the visibility and reception of similar research findings [10,11]. However, it is crucial to prioritize the value and outcomes of research above stylistic preferences, fostering an inclusive appreciation for the diversity of writing styles in academic and scientific discourse.

In the realm of computational linguistics, the analysis of language goes beyond mere comprehension, delving into the intricate features that underpin it. This quantification of language has found diverse applications, including the identification of authors' genders, detection of underlying psychological issues, and recognition of hate speech, among others [12–14]. Established frameworks in linguistics and psychology [15–17] have been instrumental in text analysis, providing tools to examine the narrative style of a text from a lexical, psychological and social perspective. Research in this field posits that the these characteristics of a text serve as a mirror to the author's thought processes and personality traits [18, 19]. Consequently, a comprehensive understanding of an author's personality can be gleaned from a thorough analysis of their writings. This approach offers a panoramic view of the author's persona, further enriching the field of text analysis.

Previous studies have demonstrated that language styles can be associated with the author's gender [20,21]. Research has extensively examined the directive style is typically used by males and the involved style is often employed by females [22]. The writing style of the authors can help identify the authors of anonymized texts [23]. Studied have also demonstrated that not only can we identify the author from gthe text but we can also identify the psychological state of the author given his/her article [24]. Quantifying the narrative style of a text provides a framework for exploring gender differences in language use within STEM fields and the research community. Although research output should be valued regardless of writing style, directive writing tends to receive more citations due to its conciseness, which allows readers to grasp the content without extensive analysis [25].

Large language models (LLMs) are extensively utilized in text generation tasks, as discussed in Section 1. There have been instances where LLMs reflect stereotypes, leading to gender bias in various societal contexts [26]. However, studies have shown that with the use of efficient prompts, the personality traits of LLMs, such as extroversion and neuroticism, can be effectively tuned [27]. Understanding the linguistic markers of LLM-generated text which reflect the lexical and psychological personality traits is an active and ongoing area of research.

To understand the extent to which LLMs induce gender bias in generated text, we conducted a correlation analysis comparing lexical, psychological, and social features computed by LIWC [15] of scientific abstracts re-generated by LLMs with those of human-written abstracts. Our study further includes an examination of differences between human-written and LLM-generated scientific abstracts to identify and quantify the gender gaps in these features.

## 3 Methodology

In this section, we present the methodology and framework of our research (Figure 1). We provide a detailed discussion of the data used for analysis, the large language models (LLMs) employed for text regeneration, and the prompts provided to these LLMs. Additionally, we describe the framework utilized to compute the lexical and psychological traits of the text. Finally, we explain our approach to quantifying the alignment between human and LLM-generated text features.

**Table 1.** Distribution of Genders in Scientific Abstracts from the CORE Dataset

| Gender | Count |
|---|---|
| Female | 418 |
| Male | 946 |
| Mixed-Gender | 2026 |

### 3.1 Data

For the analysis presented in this paper, we focused exclusively on scientific abstracts rather than full-text articles. We selected a subset of 3,390 abstracts from the CORE dataset [28]. Although the dataset includes author details, it does not provide gender information. To address this, we used the Python library, *gender-extractor*, to assign genders to the authors[4]. Table 1 shows the distribution of publications among male-only, female-only, and mixed-gender authors.

Given that many publications have both male and female authors, for the analysis of male vs. female personality alignment, we considered only publications authored solely by males or solely by females (n=1,364). However, to evaluate the overall alignment between human and LLM-generated abstracts, we utilized the entire dataset.

### 3.2 Large Language Models

The use of large language models (LLMs) has become increasingly prevalent in various natural language processing tasks. In this study, we are using LLMs to

[4] https://pypi.org/project/gender-extractor/

rewrite scientific abstracts of authors. For regenerating the text of human-written scientific abstracts, we utilized three prominent LLMs: Claude 3 Opus[5], Mistral AI Large [29], and Gemini 1.5 Flash [30]. These models were selected due to their popularity and high performance on benchmark text-generation tasks including but not limited to question answering, mathematical reasoning, diagram understanding. Their performance on these benchmarks not only surpasses traditional machine learning algorithms but also stands on par with each other, showcasing their advanced capabilities in generating coherent and contextually accurate text.

In this study, we employ large language models (LLMs) to regenerate scientific abstracts. It is crucial to use LLMs that produce text that is consistent, concise, and factually accurate. Therefore, we selected LLMs that meet these criteria [31–33].

The following prompt was consistently employed across all three LLMs with default parameters for regeneration of scientific abstracts in zero-shot setting:

"**Given the scientific abstract, imagine yourself to be an author and researcher, and rewrite this abstract. The abstract is : *[content of the abstract]***"

### 3.3 Linguistic Inquiry and Word Count (LIWC)

Linguistic Inquiry and Word Count LIWC [15] is a text analysis framework comprising thousands of dictionaries. The software quantifies the lexical, psychological, and social features of a text based on these dictionaries. In this study, we utilized the LIWC-22 [34] dictionary for research and analysis. Given our focus on scientific writing, we employed LIWC features pertinent to this domain; for instance, the *curse* feature is not relevant for quantifying scientific abstracts.

This study analyzes the Linguistic Inquiry and Word Count (LIWC) features of scientific abstracts authored by humans and those rewritten by language models. We focus on five comprehensive LIWC-22 dictionaries: cognitive processes, affect (tone and emotion features), social processes, motives, and cultural features. Our analysis concentrates on features pertinent to the narration of scientific texts, emphasizing specific, narrow categories of LIWC features. The targeted features for this research include: Segment, WC, Analytic, Clout, Tone, affiliation, achieve, power, insight, cause, discrep, tentat, certitude, differ, tone_pos, tone_neg, emotion, emo_pos, emo_neg, emo_anx, emo_anger, emo_sad, prosocial, polite, conflict, moral, comm, politic, ethnicity, tech, reward, risk, curiosity, and allure.

### 3.4 Comparison Between Human-Written and LLM-Generated Scientific Abstracts

Our objective is to determine the alignment between human and LLM-generated texts by comparing various features. We conduct two primary statistical analyses:

[5] https://www.anthropic.com/news/claude-3-family

**Correlation Analysis:** We use the Pearson correlation coefficient to compare human-written abstracts with those generated by three different LLMs: Claude 3 Opus, Gemini 1.5 Flash, and Mistral AI Large. Specifically, we calculate the correlations for each LLM to see how closely their generated texts match human-authored texts. This helps us evaluate the similarities in word choice, psychological elements, and social features between human and LLM-generated abstracts. By doing this, we can determine how well these LLMs mimic human personality traits in their scientific abstracts.

**T-Test Analysis:** We use t-tests to compare the average features of the following scientific abstracts:

- Abstracts written by women vs. abstracts written by men (Human Female vs. Human Male).
- Abstracts generated by LLMs for female authors vs. those for male authors (AI Female vs. AI Male).

We perform these analyses for all three LLMs. The results are presented in the next section.

## 4 Results

In this section, we detail how the lexical, psychological, and social features of LLM-rewritten scientific abstracts, obtained from LIWC, differ from those written by humans. We explore these differences in two key aspects: *humans vs. LLMs*, which examines the alignment of LLM personality traits with human personality traits, and *males vs. females*, which investigates whether the narrative style of LLMs reflects any gender bias.

### 4.1 Humans vs LLMs: Correlation Analysis

To assess the alignment of LIWC features between human-generated and LLM-generated scientific abstracts, we computed Pearson correlation coefficients across all pairs of features. Our focus, however, lies specifically on the diagonal elements of this correlation matrix. In simpler terms, a positive coefficient of diagonal elements indicates that humans and LLMs express features in a similar way, reflecting similar personality traits. In contrast, a negative coefficient suggests that humans and LLMs express these traits differently, indicating dissimilar personality traits.

Table 2 presents Pearson correlation coefficients for LIWC features comparing human-generated texts with those produced by all LLMs. It is important to note that all correlations are statistically significant (p_value $< 0.05$)) demonstrated by heatmap in figure 3. The lexical feature *WC* (word count) shows a weaker positive

**Table 2.** Pearson correlation for humans vs LLMs

| LIWC Category | LIWC Abbrev. | Claude | Gemini | Mistral |
|---|---|---|---|---|
| Segment | Segment | NaN | NaN | NaN |
| Word Count | WC | 0.35 | 0.80 | 0.86 |
| Analytical thinking | Analytic | 0.33 | 0.49 | 0.37 |
| Clout | Clout | 0.73 | 0.80 | 0.77 |
| Tone | Tone | 0.75 | 0.75 | 0.72 |
| Affiliation | affiliation | 0.60 | 0.73 | 0.71 |
| Achievement | achieve | 0.65 | 0.65 | 0.66 |
| Power | power | 0.81 | 0.81 | 0.80 |
| Insight | insight | 0.81 | 0.81 | 0.82 |
| Causation | cause | 0.70 | 0.75 | 0.70 |
| Discrepancy | discrep | 0.22 | 0.22 | 0.22 |
| Tentative | tentat | 0.58 | 0.71 | 0.60 |
| Certitude | certitude | 0.51 | 0.43 | 0.48 |
| Differentiation | differ | 0.66 | 0.75 | 0.73 |
| Positive tone | tone_pos | 0.68 | 0.69 | 0.66 |
| Negative tone | tone_neg | 0.80 | 0.80 | 0.76 |
| Emotion | emotion | 0.51 | 0.55 | 0.55 |
| Positive emotion | emo_pos | 0.48 | 0.56 | 0.54 |
| Negative emotion | emo_neg | 0.72 | 0.73 | 0.63 |
| Anxiety | emo_anx | 0.86 | 0.88 | 0.88 |
| Anger | emo_anger | 0.75 | 0.76 | 0.72 |
| Sadness | emo_sad | 0.53 | 0.52 | 0.31 |
| Prosocial behavior | prosocial | 0.82 | 0.79 | 0.80 |
| Politeness | polite | 0.64 | 0.64 | 0.63 |
| Interpersonal conflict | conflict | 0.82 | 0.79 | 0.78 |
| Moralization | moral | 0.70 | 0.64 | 0.59 |
| Communication | comm | 0.65 | 0.61 | 0.66 |
| Politics | politic | 0.85 | 0.88 | 0.85 |
| Ethnicity | ethnicity | 0.86 | 0.92 | 0.91 |
| Technology | tech | 0.86 | 0.91 | 0.89 |
| Reward | reward | 0.66 | 0.66 | 0.66 |
| Risk | risk | 0.85 | 0.85 | 0.84 |
| Curiosity | curiosity | 0.74 | 0.79 | 0.80 |
| Allure | allure | 0.63 | 0.62 | 0.62 |

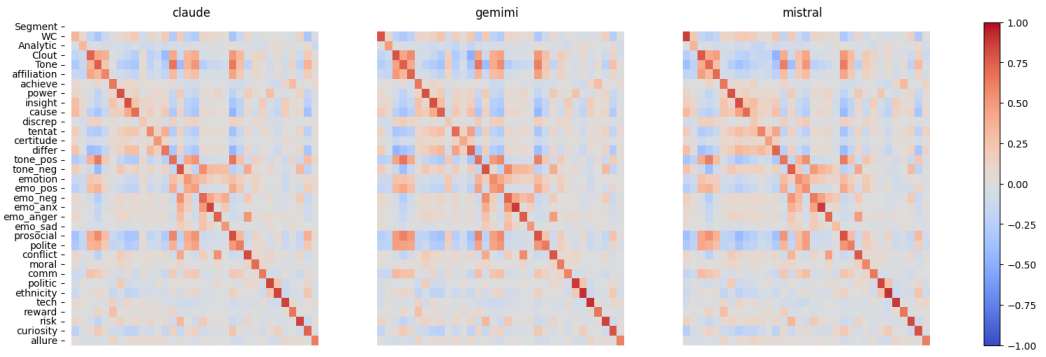


**Fig. 2.** Heatmap representing the pearson correlation coefficient of LIWC features between humans and LLMs - Claude, Gemini, Mistral (left to right)

correlation (0.35) for Claude Opus compared to Gemini and Mistral (refer to Table 2) We observed a significant positive correlation among the diagonal elements. However, we found minimal to no correlation between other pairs of features. Specifically, the diagonal elements prominently exhibit a strong positive correlation (see Figure 2).

**Cognitive Personality Traits** - Across cognitive features in Table 2, there is a generally higher positive correlation, indicating alignment between LLMs and humans. Notably, the feature "certitude," reflecting confidence in text, particularly lacks strength in Gemini, suggesting that scientific abstracts generated by LLMs may not convey certitude similarly to human-authored text.

**Tone and Emotion Related Traits** - *tone_pos* and *tone_neg* reflecting the tone of the text exhibit strong positive correlations between humans and all three LLMs. However, features related to emotions relevant to scientific writing (*emotion* and *emo_pos*) show positive correlations but not as pronounced.

**Social Processes Personality Traits** - Analysis of social process features such as 'prosocial', 'polite', 'conflict', 'moral', and 'comm' reveals consistently high positive correlation coefficients across all three LLMs. Claude Opus particularly stands out with the highest values, indicating its texts closely mirror human social processes.

**Motive Personality Traits** reflected by 'reward', 'risk', 'curiosity', and 'allure' demonstrate high positive correlations between LLM generated and human-written abstracts, suggesting precise capture and reflection of textual motives by LLMs.

**Culture Traits** reflected by LIWC features - 'politic', 'ethnicity', and 'tech'- exhibit higher correlations between humans and LLMs, indicating accurate reflection of document categories in regenerated texts.

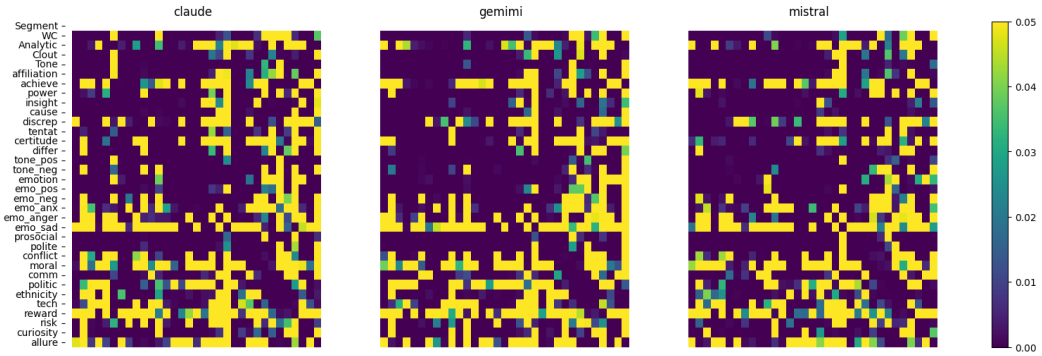


**Fig. 3.** Heatmap representing the significance (p_value) of pearson correlation coefficient of LIWC features between humans and LLMs - Claude, Gemini, Mistral (left to right)

Our findings demonstrate a strong positive correlation between LIWC features in human-written and LLM-generated scientific abstracts across all three LLM models. This suggests that LLMs effectively capture the lexical characteristics, psychological traits, and social dynamics observed in human-authored texts.

### 4.2 Gender Bias: Two Sample t-test

We conducted a two-sample t-test to compare the LIWC features between male and female authors. Among the 35 features analyzed, 15 features showed a statistically significant t-statistic value (p-value $< 0.05$). Table 3 presents the t-statistic values for scientific abstracts written by humans, as well as those generated by Claude, Gemini, and Mistral. In the table, the significant values are in bold and italic for clarity.

**Lexical Features** - Among the lexical features, *WC* (word count) and *tone* are statistically significant. On average, female authors use 5.86 times more words than male authors, a pattern that is also observed in the LLM-generated texts, as shown in table 3. Additionally, male authors tend to have a more positive tone in their abstracts compared to female authors, approximately 4 times more. Although LLMs reflect this same trend, it is noteworthy that the difference is less pronounced, especially for Claude. This highlights an important observation: while LLMs generally follow the authors' personality traits in lexical features, they can sometimes underestimate or overestimate these features.

**Table 3.** t-test statistics for Males vs Females

| LIWC Category | LIWC Abbrev. | Human | Claude | Gemini | Mistral |
|---|---|---|---|---|---|
| Segment | Segment | NaN | NaN | NaN | NaN |
| Word Count | WC | ***-5.86*** | ***-4.66*** | ***-6.13*** | ***-6.31*** |
| Analytical thinking | Analytic | -1.13 | -1.63 | -1.74 | -1.22 |
| Clout | Clout | -0.71 | -1.86 | -1.69 | -1.44 |
| Tone | Tone | ***3.93*** | 1.77 | ***2.37*** | ***2.43*** |
| Affiliation | affiliation | 0.19 | -0.71 | -0.57 | 1.57 |
| Achievement | achieve | -0.78 | ***-3.22*** | ***-2.15*** | ***-2.67*** |
| Power | power | -0.42 | 0.75 | -0.51 | -0.27 |
| Insight | insight | ***-6.74*** | ***-6.66*** | ***-7.64*** | ***-6.98*** |
| Causation | cause | ***-2.27*** | -0.51 | -1.56 | -1.38 |
| Discrepancy | discrep | -0.35 | -0.33 | ***-3.01*** | 0.73 |
| Tentative | tentat | -0.31 | 0.23 | 1.67 | 0.86 |
| Certitude | certitude | -0.01 | -0.54 | -1.66 | -1.90 |
| Differentiation | differ | ***-3.20*** | ***-2.17*** | ***-3.11*** | -1.59 |
| Positive tone | tone_pos | ***3.88*** | ***2.21*** | ***2.04*** | ***2.01*** |
| Negative tone | tone_neg | -1.25 | -1.12 | -0.77 | -0.7 |
| Emotion | emotion | ***3.33*** | ***5.8*** | ***3.73*** | ***5.05*** |
| Positive emotion | emo_pos | ***3.82*** | ***6.20*** | ***4.71*** | ***5.64*** |
| Negative emotion | emo_neg | 0.96 | 1.30 | 1.40 | 1.01 |
| Anxiety | emo_anx | -0.08 | -0.01 | 0.22 | 0.12 |
| Anger | emo_anger | -0.31 | 0.27 | 0.36 | -0.61 |
| Sadness | emo_sad | 1.07 | 0.27 | 0.67 | 0.79 |
| Prosocial behavior | prosocial | ***3.15*** | ***3.00*** | ***2.96*** | 1.57 |
| Politeness | polite | ***5.70*** | ***3.70*** | ***4.10*** | ***2.82*** |
| Interpersonal conflict | conflict | -1.95 | ***-2.53*** | ***-2.58*** | ***-2.00*** |
| Moralization | moral | -1.63 | -0.92 | 0.11 | -0.14 |
| Communication | comm | 1.08 | -1.42 | -0.84 | -1.52 |
| Politics | politic | -0.60 | -0.85 | -1.26 | -1.45 |
| Ethnicity | ethnicity | 0.37 | -0.51 | 0.28 | -0.13 |
| Technology | tech | 1.66 | 0.85 | 0.93 | 1.54 |
| Reward | reward | -1.84 | -1.73 | -.12 | -1.46 |
| Risk | risk | ***2.61*** | ***2.10*** | 1.94 | 1.93 |
| Curiosity | curiosity | 1.06 | ***3.75*** | ***2.60*** | ***2.66*** |
| Allure | allure | -1.77 | -0.48 | -0.9 | -0.3 |

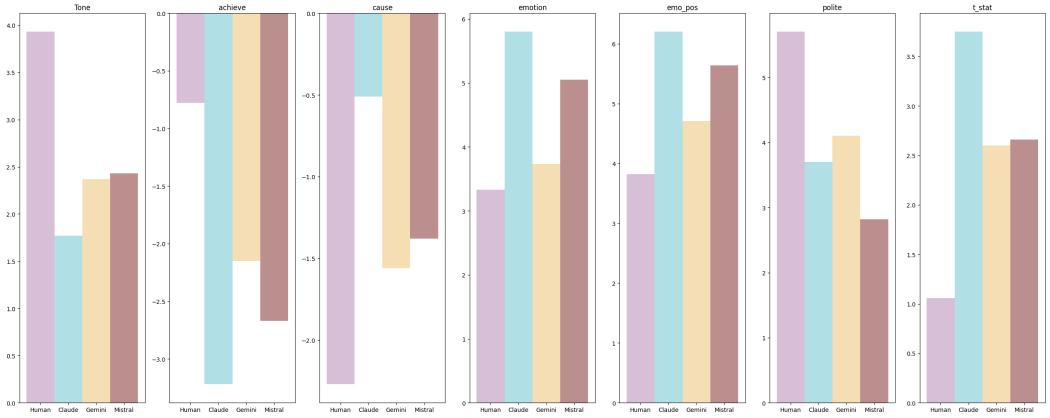


**Fig. 4.** t-statistic values of statistically significant features ('Tone', 'achieve', 'cause', 'emotion', 'emo_pos','polite','curiosity') which reflects gender gaps between human and LLM texts.

**Cognitive Personality Traits** - Statistically significant connitive processes are reflected by *insight*, *cause*, and *differ* LIWC features. The scientific abstracts show that female authors use approximately seven times more insightful words such as "know," "think," and "feel" compared to male authors. This pattern is also reflected in the LLM-generated texts, with similar values (table 3). For the feature *differ*, the difference in writing style between males and females is consistent across all LLMs except Mistral, where the t-statistic is not significant. Interestingly, for the feature *cause* (causation), the difference between male and female narration is significant, with females using more causation-centric terms. However, this trend is not followed by any of the LLMs. Although not a drastic difference, but LLMs do introduce bias in terms of using causation-centric words that is found approximately twice more in females than males.

**Tone and Emotion Related Traits** are significantly represented by three LIWC features: *tone_pos*, *emotion*, and *emo_pos* ($p < 0.05$). These features, out of the eight selected, were the only ones to show statistical significance. They reflect the positive tone and emotional content present in the text. Interestingly, both human authors and large language models (LLMs) consistently indicated that texts authored by males exhibited at least three times more positivity than those authored by females (Table 1). However, it is noteworthy that the LLMs, specifically Claude and Mistral, tended to overestimate the emotional content and positivity in texts authored by males. This overestimation could potentially amplify the perceived gender disparity between male and female authors in the field.

**Social Processes Personality Traits** - The categories *prosocial*, *polite*, and *conflict* are statistically significant. Notably, while the difference in narration style be-

tween males and females regarding conflict is non-significant, it becomes significant in the texts generated by large language models (LLMs). This suggests that female authors use a more conflicting style of writing compared to their male counterparts. Additionally, male authors tend to adopt a significantly more polite writing style, approximately five times more polite on average than females that is consistent with existing literature [25]. However, this distinction is underestimated in LLM-generated abstracts, contributing to a gender gap in the portrayal of politeness in writing styles (see table 3).

**Motive Personality Traits** include risk, curiosity and allure. We identified *risk* and *curiosity* as two significant features. Our analysis revealed that male authors used approximately twice as many risk-associated words compared to female authors. However, this difference was only statistically significant ($p < 0.05$) in texts regenerated by Claude, and not in those regenerated by Gemini or Mistral. In addition to this, the analysis shows a significant ($p\_value < 0.05$) difference between male and female re-written abstracts by LLMs for *achieve* category reflecting that females use approximately thrice more words reflecting achievement which is not reflected in the human authored articles. Additionally, our analysis found no significant difference ($p > 0.05$) between male and female authors in terms of reflecting curiosity in their narrative style. Nevertheless, LLMs indicated that male authors adopted a narrative style that reflected more curiosity than their female counterparts (Table 2). This finding highlights the potential for LLMs to further exacerbate the gender gap in scientific writing by perpetuating biased representations of author characteristics.

In cconclusion, we did not observe a gender gap in the lexical features and cognitive processes of scientific abstracts when comparing human authors to LLM-generated texts. However, for the psychological processes of affect and motive, LLMs tend to amplify the gender gap observed between males and females for certain features. Additionally, while not drastic, a gender gap is also present in the politeness feature within the social behavior category. The t-statistics for these features indicate a significant difference (see Figure 4).

## 5 Limitations and Future Works

While this study has provided valuable insights into the personality traits reflected by LLMs in the context of scientific writing, there are several limitations that need to be addressed.

Firstly, the study has not taken into account the imbalance between the number of female and male authors. It would be significant to observe if the average LIWC features would differ drastically when correction for class imbalance is considered. This could provide a more accurate representation of gender differences in scientific writing.

Secondly, the data for this study was randomly sampled. However, it is important to note that different academic fields have varying levels of gender representation. For instance, some areas such as 'arts' and 'psychology' may be female-dominated, while others like 'engineering' may be male-dominated. Therefore, it would be crucial to analyze the gender gap induced by LLMs for different academic circles to provide a more comprehensive understanding of the issue.

Lastly, the study has not considered longitudinal changes in the gender gap over the years. As the number of female authors in academic and scientific publications has grown over time, it is important to quantify how the gender gaps have evolved. This could provide valuable insights into the progress made towards gender equality in these fields.

In terms of future work, efforts should be made to refine LLM algorithms to minimize bias and enhance their ability to generate equitable representations across different genders. This could involve training LLMs on more diverse datasets and incorporating fairness metrics into their evaluation. Additionally, future research could explore the intersectionality of gender with other factors such as race, ethnicity, and age to provide a more nuanced understanding of bias in LLM-generated text.

## 6 Conclusion

The use of LLMs in scientific writing necessitates an examination of their ability to replicate the traits of human authors and researchers. In this paper, we have shown that LLMs can effectively mirror the social, psychological, and lexical traits exhibited by humans. However, our analysis also revealed that some psychological and social traits reflected by LLMs exhibit significant gender biases. We have highlighted the gender gaps present in three widely used LLMs. Our findings underscore the importance of addressing and mitigating biases in these models to ensure fair and equitable representation across different genders.

As LLMs continue to gain prominence in scientific writing, it is crucial to prioritize the development of unbiased and inclusive models. This can be achieved by incorporating diverse datasets, implementing fairness metrics, and conducting regular evaluations of LLM-generated text. By doing so, we can leverage the full potential of LLMs while promoting gender equality and minimizing the perpetuation of harmful stereotypes.

## Authors

**Naseela Pervez** received her M.S Computer Science from Viterbi School of Engineering, University of Southern California (USC). Currently, she is a pre-doctoral research staff at Management of Innovation, Entrepreneurial Research, and Venture

Analysis (MINERVA) and Information Sciences Institute (ISI) at the University of Southern California (USC). Her research interests include natural language processing, network science, social sciences, and fairness and bias in AI.

**Alexander J. Titus** received his Ph.D. in Quantitative Biomedical Sciences from Dartmouth College. Currently, he is a Principal Scientist at the Information Sciences Institute and Research Faculty at the Iovine and Young Academy at the University of Southern California (USC), and Founder and Principal Investigator at the In Vivo Group. His research interests include the applications of artificial intelligence to the life sciences and responsible AI development.

## References


1. N. Rinaldo, G. Piva, S. Ryder, A. Crepaldi, A. Pasini, L. Caruso, R. Manfredini, S. Straudi, F. Manfredini, and N. Lamberti, "The issue of gender bias represented in authorship in the fields of exercise and rehabilitation: A 5-year research in indexed journals," *J. Funct. Morphol. Kinesiol.*, vol. 8, p. 18, Jan. 2023.
2. P. van den Besselaar and U. Sandström, "Gender differences in research performance and its impact on careers: a longitudinal case study," *Scientometrics*, vol. 106, pp. 143–162, Nov. 2015.
3. A. H. Kerkhoven, P. Russo, A. M. Land-Zandstra, A. Saxena, and F. J. Rodenburg, "Gender stereotypes in science education resources: A visual content analysis," *PLoS One*, vol. 11, p. e0165037, Nov. 2016.
4. N. Arkin, C. Lai, L. M. Kiwakyou, G. M. Lochbaum, A. Shafer, S. K. Howard, E. R. Mariano, and M. Fassiotto, "What's in a Word? Qualitative and Quantitative Analysis of Leadership Language in Anesthesiology Resident Feedback," *Journal of Graduate Medical Education*, vol. 11, pp. 44–52, 02 2019.
5. `https://debuk.wordpress.com/2016/03/06/do-women-and-men-write-differently/`. [Accessed 21-06-2024].
6. S. J. Ceci, W. M. Williams, and S. M. Barnett, "Women's underrepresentation in science: sociocultural and biological considerations," *Psychol. Bull.*, vol. 135, pp. 218–261, Mar. 2009.
7. A. Hillier, R. P. Kelly, and T. Klinger, "Narrative style influences citation frequency in climate change science," *PLoS One*, vol. 11, p. e0167983, Dec. 2016.
8. G. H. Bower and M. C. Clark, "Narrative stories as mediators for serial learning," *Psychonomic Science*, vol. 14, pp. 181–182, Apr. 1969.
9. E. Levitskaya, K. Kedrick, and R. J. Funk, "Investigating writing style as a contributor to gender gaps in science and technology," 2022.
10. M. Helmer, M. Schottdorf, A. Neef, and D. Battaglia, "Research: Gender bias in scholarly peer review," *eLife*, vol. 6, p. e21718, mar 2017.
11. V. Larivière, C. Ni, Y. Gingras, B. Cronin, and C. R. Sugimoto, "Bibliometrics: Global gender disparities in science," *Nature*, vol. 504, pp. 211–213, Dec. 2013.
12. V. Pérez-Rosas and R. Mihalcea, "Gender differences in deceivers writing style," in *Human-Inspired Computing and Its Applications* (A. Gelbukh, F. C. Espinoza, and S. N. Galicia-Haro, eds.), (Cham), pp. 163–174, Springer International Publishing, 2014.
13. K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, pp. 407–422, 2016.
14. K. Englmeier, "The role of storylines in hate speech detection (short paper)," in *Machine Learning for Trend and Weak Signal Detection in Social Networks and Social Media*, 2020.

15. Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of Language and Social Psychology*, vol. 29, pp. 24 – 54, 2010.
16. S. A. Crossley, K. Kyle, and D. S. McNamara, "Sentiment analysis and social cognition engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis," *Behavior Research Methods*, vol. 49, pp. 803–821, June 2017.
17. E. Fast, B. Chen, and M. S. Bernstein, "Empath: Understanding topic signals in large-scale text," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, (New York, NY, USA), p. 4647–4657, Association for Computing Machinery, 2016.
18. K. Llerena, S. G. Park, S. M. Couture, and J. J. Blanchard, "Social anhedonia and affiliation: examining behavior and subjective reactions within a social interaction," *Psychiatry Res*, vol. 200, pp. 679–686, Aug. 2012.
19. K. Brauer, R. Sendatzki, and R. T. Proyer, "Testing associations between language use in descriptions of playfulness and age, gender, and self-reported playfulness in german-speaking adults," *Frontiers in Psychology*, vol. 13, 2022.
20. A. Haas, "Male and female spoken language differences: Stereotypes and evidence," *Psychological bulletin*, vol. 86, no. 3, pp. 616–626, 1979.
21. J. N. Martin and R. T. Craig, "Selected linguistic sex differences during initial social interactions of same-sex and mixed-sex student dyads," *Western Journal of Speech Communication*, vol. 47, no. 1, pp. 16–28, 1983.
22. X. Wen, P. M. McCarthy, and A. C. Strain, "A gramulator analysis of gendered language in cable news reportage," in *The Florida AI Research Society*, 2013.
23. N. E. Benzebouchi, N. Azizi, N. E. Hammami, D. Schwab, M. C. E. Khelaifia, and M. Aldwairi, "Authors' writing styles based authorship identification system using the text representation vector," in *2019 16th International Multi-Conference on Systems, Signals Devices (SSD)*, pp. 371–376, 2019.
24. X. Du and Y. Sun, "Linguistic features and psychological states: A machine-learning based approach," *Front Psychol*, vol. 13, p. 955850, July 2022.
25. Y. Ma, Y. Teng, Z. Deng, L. Liu, and Y. Zhang, "Does writing style affect gender differences in the research performance of articles?: An empirical study of BERT-based textual sentiment analysis," *Scientometrics*, vol. 128, pp. 2105–2143, Apr. 2023.
26. H. Kotek, R. Dockum, and D. Sun, "Gender bias and stereotypes in large language models," in *Proceedings of The ACM Collective Intelligence Conference*, CI '23, (New York, NY, USA), p. 12–24, Association for Computing Machinery, 2023.
27. S. Mao, X. Wang, M. Wang, Y. Jiang, P. Xie, F. Huang, and N. Zhang, "Editing personality for large language models," 2024.
28. P. Knoth, D. Herrmannova, M. Cancellieri, L. Anastasiou, N. Pontika, S. Pearce, B. Gyawali, and D. Pride, "Core: A global aggregation service for open access papers," *Nature Scientific Data*, vol. 10, p. 366, June 2023.
29. A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," 2023.
30. G. Team, P. Georgiev, V. I. Lei, R. Burnell, *et al.*, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," 2024.
31. P. Laban, W. Kryscinski, D. Agarwal, A. Fabbri, C. Xiong, S. Joty, and C.-S. Wu, "SummEdits: Measuring LLM ability at factual reasoning through the lens of summarization," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (H. Bouamor, J. Pino, and K. Bali, eds.), (Singapore), pp. 9662–9676, Association for Computational Linguistics, Dec. 2023.
32. D. Tam, A. Mascarenhas, S. Zhang, S. Kwan, M. Bansal, and C. Raffel, "Evaluating the factual consistency of large language models through news summarization," in *Findings of the Association for Computational Linguistics: ACL 2023* (A. Rogers, J. Boyd-Graber, and

N. Okazaki, eds.), (Toronto, Canada), pp. 5220–5255, Association for Computational Linguistics, July 2023.
33. R. Uceda-Sosa, K. N. Ramamurthy, M. Chang, and M. Singh, "Reasoning about concepts with llms: Inconsistencies abound," *arXiv preprint arXiv:2405.20163*, 2024.
34. R. Boyd, A. Ashokkumar, S. Seraj, and J. Pennebaker, "The development and psychometric properties of liwc-22," 02 2022.