Embeddings are not a good measure of cultural homogeneity

A critique of “AI Suggestions Homogenize Writing Toward Western Styles and Diminish Cultural Nuances”

175729 Psychology and Culture

Assignment 1

Vojtěch Formánek, 25021757
Masaryk University, Massey University
xforman@mail.muni.cz

18.8.2025

Introduction

"AI Suggestions Homogenize Writing Toward Western Styles and Diminish Cultural Nuances" (Argawal, Naaman & Vashistha, 2025) is an interdisciplinary article from Psychology and Natural Language Processing (NLP), studying the impact of bias in a Large Language Model-based (LLM) chatbot on its users. Specifically, it explores whether writing assistance from this chatbot, ChatGPT-4.5, in the form of auto-completions, homogenizes user's writing toward western styles. The article contains an inspiring exploratory analysis of data of 120 participants from the American and Indian culture across two cultures: American or Indian and two conditions: AI assistance or No AI.

As the title states, the main result of the article is that AI homogenizes writing toward Western styles, in this case American. A secondary result is that AI increases productivity more for Americans than Indians. Further qualitative and quantitative analyses were also performed, among them is a finding that using AI reduces the use of heroes or symbols from Indian culture in the writing. The writing consisted of four tasks, that roughly reflect relevant daily tasks. The task and participants' responses were in English; the participants were recruited using an online platform Prolific and screened for proficiency in English.

The authors reach their results by combining both NLP and Psychological approaches. Mixing typical statistical data with the use of relatively recent sentence embedding models, allowing for a quantitative analysis of otherwise qualitative text. While the combination of methodologies is necessary, and in many cases well done, it does sometimes draw invalid conclusions from statistical results and overlooks some cross-cultural aspects of embeddings, which undermine the claims about homogeneity.

I focused on methodology to keep the essay concise and because it is relevant to my own work. The essay is structured according to parts of the article; I'll first do a review of relevant theory before each section.

Does AI impact Indians and Americans differently?

The impact of p-values

P-values have been used as the gold standard in empirical research, to test scientific hypotheses and present results. But p-values alone aren't enough. There are many reasons justifying that statement, but the one relevant to my critique is sample sizes. The problem is simple, with sample that is large enough, the result of any statistical test will become statistically significant (Sullivan, 2012), even though the actual differences between populations might be negligible. P-value tells us whether the difference between populations is due to chance, but it is not enough to make a conclusion about the difference itself.

Suggestion Modification Behavior

The article examined whether there is a difference between American and Indian participants when accepting or rejecting modifications. The bootstrapping methodology used to test the hypothesis is unusual and not suitable to draw the conclusions. In general, bootstrapping is not the same as drawing the data naturally and here the comparison is between percentages from populations, instead of the populations themselves. The practical consequence is that a) a slight difference in the data might be overrepresented in the result and b) the result isn't related to the hypothesis directly.

	accepted	rejected
American	91.4	52.6
Indian	68.9	47.1

Fig. 1. – Estimation of values of the two groups and accept/reject modification condition. The estimates are not integers, to make the computation more accurate, since the real data are not accessible.

For a comparison of two populations with binary values (i.e. accepted x modified) a Chi-square test is more suitable. It is possible to reconstruct the contingency table from the values in the paper (Fig. 1.). Based on this contingency table and a Chi-square test, I tested a null hypothesis, whether ethnicity and suggestion acceptance on a task are independent on $\alpha=0.05$. The resulting difference is not statistically significant ($\chi^2(1)=0.3$, $p=0.58$), thus no conclusion should be drawn about the differences between modification acceptance of Americans and Indians.

Productivity Derived from AI Suggestions

Productivity is a misleading name for average writing speed (words per second). As the authors mention, the term productivity is multi-dimensional and whilst writing speed is a part of it, it is not the whole construct. Writing in non-native language is usually slower

(Schoonen et al., 2003), as can be seen in Figure 5.b of the article. Chatbots are also known to increase writing speed and productivity (e.g., finishing scientific papers; Chen, 2023) and the article results confirm previous findings.

The authors show that both American and Indian participants write faster when using AI. The improvements in both groups, per the results of the regression model, indicate that the relative improvements are similar. However, the authors ignore this result and present an alternative interpretation, based on a confusingly constructed score – per suggestion productivity (i.e. per suggestion writing speed). As I established earlier, Indian participants wrote slower, irrespective of whether AI was used, because English is their second language. The score doesn't reflect this. It only averages writing speed per suggestion, thus the slower writing speed heavily skews the results.

The authors state that the difference is significant, but statistical significance is not an indicator of effect size. Further, the sample is large, thus the critique the use of p-values applies here, so the estimate is unreliable. Instead, the effect size should be interpreted as the main score, in this case showing a minor difference.

Given that no precaution was taken to mitigate the effect of writing speed on productivity per second, and the fact that Indian participants wrote slower in English, it is likely that there is no difference between relative improvements of the groups. That hypothesis is also in part supported by the regression model results.

Does AI homogenize writing within cultures?

The Unreliability of Embeddings

NLP, a subset of AI that includes generative AI and LLMs, usually doesn't test statistical hypotheses, but instead evaluates on similar data that the models are trained on. LLMs are trained on huge datasets, the size of the world wide web, to work properly. This data is essential to the model and dictates what it outputs to the users. WEIRD data are usually overrepresented in the training and evaluation datasets (Mihalcea et al., 2025), which is why the article was made in the first place. However, the authors don't extend the same scrutiny to the sentence embedding model they used.

What is a sentence embedding? For any LLM to work, each word in its input (e.g., a user's prompt) must be separately translated into a string of numbers of the same length, called an embedding. This translation can be done using another model or a mathematical method. The sentence embedding model just translates the entire sentence into a single string of numbers, a sentence embedding. They are trained differently than LLMs, but also use WEIRD data, and the quality of the translation depends on how often the model sees similar data. Consequently, it can perform worse for unusual (non-WEIRD) sentence structures, symbols, etc. Making possible comparisons unreliable.

Cosine similarity is common measure used to show how similar two embeddings are. It is a score from 0 to 1; 0 indicating absolutely no similarity, whilst a 1 is given to identical inputs. Otherwise, there is no established interpretation of the score, such as with Cohen's d, and is task dependent. But various concerns have been raised about its reliability in measuring the similarity of embeddings (Steck, Ekanadham & Kallus, 2024).

Similarity analysis

Cosine similarity of embeddings might not be the best measure of homogeneity, because the sentence embedding model might not represent the participants' writing correctly (Schroder, Schulz & Hammer, 2024; Nikolaev, 2023). Indians might skip *"the"*, *"a"*, write incorrect grammatical or semantical structures. Although all aren't strictly mistakes, they might have outsized effect on the distance between the groups' writing. LanguageTool checks for grammatical mistakes, it doesn't detect and change sentence structure. Additionally, combining AI generated and original Indian text can create data that the sentence-embedding model is not trained to represent, negatively impacting the accuracy of their representation. I also believe turning off spelling is a big mistake, as it does impact the embeddings. It would've been insightful to know how many mistakes would've been corrected that way.

Experiments

Effect sizes can indicate important differences, nonetheless I believe it is important to show what concrete changes in text the cosine similarity scores represent. I want to highlight the non-trivial interaction of embeddings and cosine score. Because the model in the article isn’t open-source, I used all-MiniLM-L6-v2 (Reimers & Gurevych, 2020), but their behavior is similar. I'll use two examples from the article, both from the Indian participants, and compute the similarity scores between them.

A: *"My favorite food is the chicken biriyani. I like it because it tastes good and it is easy to prepare. I like to prepare it in the Malabar style. That recipe uses some exotic ingredients like the nutmeg. The chicken biriyani, along with the raita, the lemon pickle and the dates chutney, tastes divine. The chicken biriyani is supposed to be a dish that was brought to India by the Mughals.*"

B0: *"My favorite festival is Diwali. Diwali is usually celebrated in between October and November. On this occasion, I used to worship goddess Laxmi along with my family. Most fascinating thing is to pop crackers and eat sweets for minimum span of 4 days. I always wait for this festival. We lighten our houses with earthen lamps and worship cows and lords.*"

The cosine similarity of A and B0 is 0.42. Now, I'll show how slight changes impact the scores, by modifying the first two sentences of B0 and keeping the rest as is.

B1: *My favorite festival is Diwali, usually ...*
B2: *My favorite festival is usually ...*

The cosine similarity between B0 and B1, 2 is 0.99, 0.90 resp. The differences between A0 and B1 are minimal, and the cosine similarity (0.42 after rounding) is also very similar to A0 and B1. The similarity between A0 and B2 is 0.33 – a noticeable difference, larger than the average in the article. Last, I want to show how unusual structure impacts the embeddings. If the phrase *"chicken biriyani"* (a phrase repeated in A) is added in the middle of the B0 text, the similarity increases to 0.52. Further, if the entire sentence *"My favorite food is the chicken biriyani."* is added to the front of B0, the similarity increases to 0.70.

Discussion

I decided to focus on methodology, because I also had to bridge NLP and Psychological methodologies in my own research, and the essay allows me to write out the problems I see with them concisely, especially with embeddings. Nonetheless, there are many other issues with the article, such as the sample choice and size, or the homogenization that is defined as embedding similarity. However, the methodological gap that the authors are filling is a big one, and there isn't much theoretical research to draw on.

I highlighted that the authors misinterpret the statistical results, especially regarding the writing speed. The conclusion I've drawn, based on their results and my partial replication, is more optimistic – both American and Indian participants show similar improvements while using AI autocompletions, if the effect of writing speed in English is controlled. Nonetheless I can't test these hypotheses, as the authors can't release the data, due to ethical considerations. The reliance on embeddings might invalidate some of the quantitative results drawn, because the way the embeddings are computed is also biased and thus might not represent actual meaning accurately. The experiments highlight that removing a culture-based word such as Diwali, has a large impact on the score. Though this does differ for other cultural words, as the embeddings might represent them differently. The next part also highlights that adding context, can have an outsized effect on the cosine similarity, in an amount that might not reflect the semantic impact as the participant would see it.

Although modifying these semantic attributes can be considered homogenization, possibly even cultural, it does not imply the implicit change, as the authors are suggesting, since the embeddings might not be a good-enough tool to measure it. They also don't consider a third option, that the homogenization might not be directly toward the Western style, but sideways, into a space that is not necessarily Western. Instead, a homogenized version of all cultures, each having different level of influence, creating a melting pot. Consequently, diminishing not only Indian and other non-Western symbols and ideas, but also the Western. The qualitative analysis done at the end of the paper is very insightful into the possible workings of the homogenization. Omitting culture-based words from the generated text is a very plausible way that LLMs could be changing cultural expression, especially since these phrases are likely way less represented in their training data, than their more generic counterparts. Embeddings are sensitive to their removal and thus the difference would show when computing cosine similarity. The analysis is not exhaustive and should be followed upon in future research, as this finding represents a plausible, measurable and terrifying way that LLM-based autocompletion could be erasing deep cultural expression.

Conclusion

Within the essay I summarized relevant parts of the article, which compared the impact of AI assistance on textual expression of Indian and American participants. I highlighted some of the methodological issues, focusing on the misinterpretation of statistical results and overreliance on embeddings, with an extended discussion on their reliance on WEIRD data. I included a replication to the best of my ability and the data available, and extended experiments to show the unreliability of embeddings. I reached the conclusion that the impact of the autocompletions is similar across both groups and that the embedding-based homogenization measurement is unreliable, but the qualitative results of the article – erasure of specific cultural words – is a plausible cause behind the quantitative results and a dangerous way LLMs could impacting textual expression.

Bibliography

Agarwal, D., Naaman, M., & Vashistha, A. (2025). AI suggestions homogenize writing toward western styles and diminish cultural nuances. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, *1*, 1–21. https://doi.org/10.1145/3706598.3713564

Chen, T.-J. (2023). ChatGPT and other artificial intelligence applications speed up scientific writing. *Journal of the Chinese Medical Association: JCMA*. https://doi.org/10.1097/JCMA.0000000000000900

Mihalcea, R., Ignat, O., Bai, L., Borah, A., Chiruzzo, L., Jin, Z., ... & Solorio, T. (2025, April). Why AI Is WEIRD and Shouldn't Be This Way: Towards AI for Everyone, with Everyone, by Everyone. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39, No. 27, pp. 28657-28670).

Nikolaev, D., & Padó, S. (2023). Representation biases in sentence transformers. *arXiv preprint arXiv:2301.13039*.

Reimers, N., & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

Schoonen, R., van Gelderen, A., Glopper, K. de, Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2003). First language and second language writing: The role of linguistic knowledge, speed of processing, and metacognitive knowledge. *Language Learning*, *53*(1), 165–202. https://doi.org/10.1111/1467-9922.00213

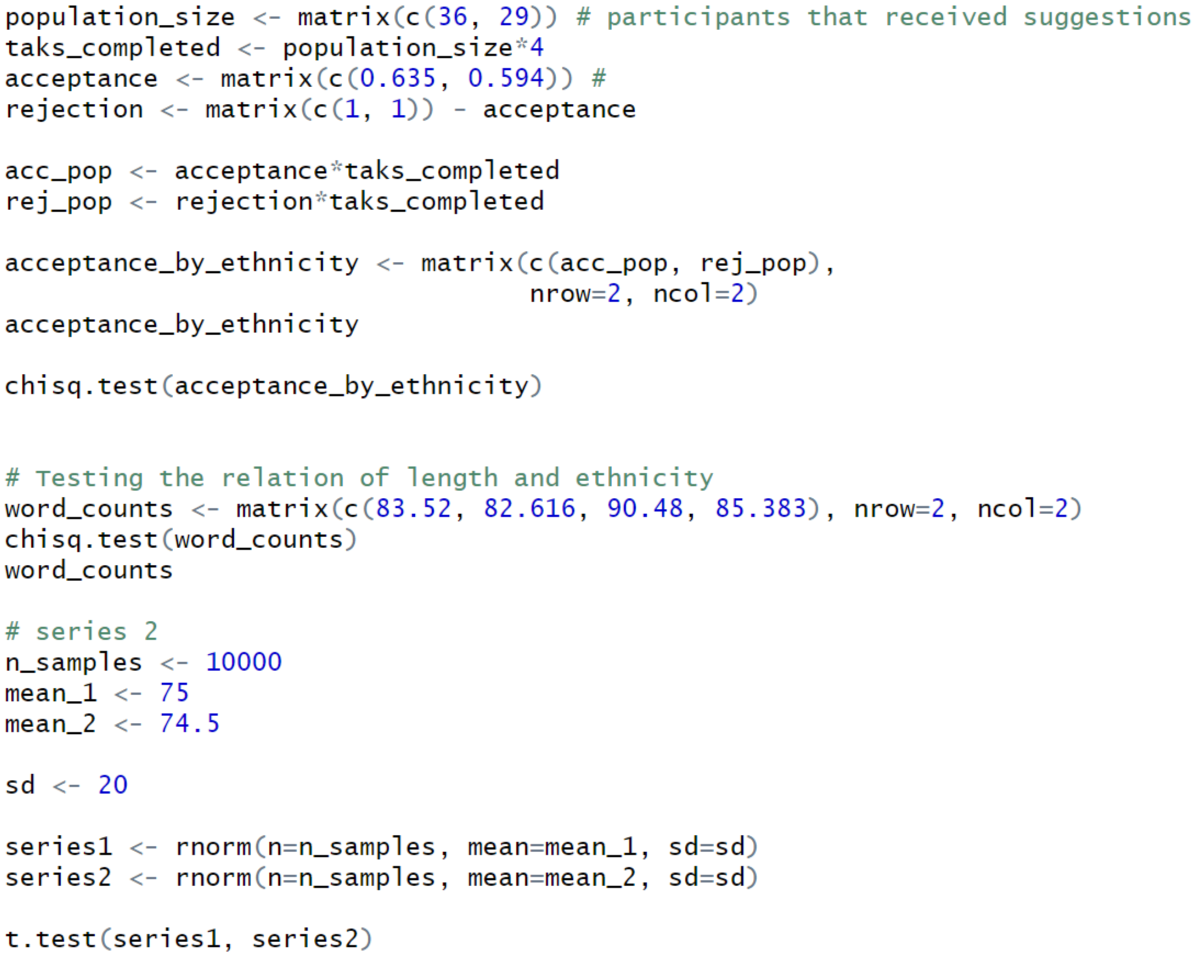
Schröder, S., Schulz, A., & Hammer, B. (2024, June). The SAME score: Improved cosine based measure for semantic bias. In *2024 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

Steck, H., Ekanadham, C., & Kallus, N. (2024, May). Is cosine-similarity of embeddings really about similarity?. In *Companion Proceedings of the ACM Web Conference 2024* (pp. 887-890).

Sullivan, G. M., & Feinn, R. (2012). Using Effect Size—or Why the P Value Is Not Enough. *Journal of Graduate Medical Education*, *4*(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

Appendix

R code:



```
population_size <- matrix(c(36, 29)) # participants that received suggestions
taks_completed <- population_size*4
acceptance <- matrix(c(0.635, 0.594)) #
rejection <- matrix(c(1, 1)) - acceptance

acc_pop <- acceptance*taks_completed
rej_pop <- rejection*taks_completed

acceptance_by_ethnicity <- matrix(c(acc_pop, rej_pop),
                                  nrow=2, ncol=2)
acceptance_by_ethnicity

chisq.test(acceptance_by_ethnicity)


# Testing the relation of length and ethnicity
word_counts <- matrix(c(83.52, 82.616, 90.48, 85.383), nrow=2, ncol=2)
chisq.test(word_counts)
word_counts

# series 2
n_samples <- 10000
mean_1 <- 75
mean_2 <- 74.5

sd <- 20

series1 <- rnorm(n=n_samples, mean=mean_1, sd=sd)
series2 <- rnorm(n=n_samples, mean=mean_2, sd=sd)

t.test(series1, series2)
```

Sentence-embedding code:

https://colab.research.google.com/drive/16fXURW9lybz-vwVhE-xsu5RkKlm8bKEK?usp=sharing