Pre-Registration for The Effect of Group Status on the Variability of Group Representations in LLM-generated Text

July 25, 2023

Template adapted from AsPredicted.org

1) **Data collection.** Has any data been collected for this study already?

We have not yet collected data for this study.

2) **Hypothesis.** What’s the main question being asked or hypothesis being tested in this study?

Question 1: Does ChatGPT depict U.S. racial/ethnic minority groups (African, Asian, and Hispanic Americans) as more homogeneous compared to the U.S. racial/ethnic majority group (White Americans)?

Question 2: Does ChatGPT depict the gender minority group (woman) as more homogeneous compared to the gender majority group (man)?

Question 3: Is the effect of gender on the homogeneity of text generated by ChatGPT consistent across racial / ethnic groups?

3) **Dependent variable(s).** Describe the key dependent variable(s) specifying how they will be measured.

HOMOGENEITY OF CHATGPT-GENERATED TEXT: We will compute the pairwise cosine similarity between sentence embeddings of the Chat-GPT generated text. Sentence embeddings represent the meanings of sentences as vectors in a high-dimensional semantic space. Sentence embeddings can be derived using different types of models that capture context and meanings of words in sentences. Here, we will use the BERT-base-uncased model ([Devlin et al., 2018](https://huggingface.co/bert-base-uncased)). After performing minimal preprocessing steps (lower-casing, remove non-alphanumeric characters, and trim extra whitespaces) on the collected texts, we will use the text package in R ([Kjell et al., In progress](https://r-text.org/authors.html" \l "citation)) to input all the sentences into the BERT model and extract the second to last hidden layer (the default layer in the text package) to use as our sentence embeddings.

Then, we will compute the pairwise cosine similarity between the sentence embeddings of all texts generated for each intersectional group. Cosine similarity is calculated by taking the dot product of two vectors and dividing it by the product of their magnitudes. The value can range from -1 and 1, where 1 indicates that the two vectors are perfectly identical and where -1 indicates that the two vectors are completely dissimilar.

4) **Conditions.** How many and which conditions will participants be assigned to?

We did not assign participants to conditions, but we asked ChatGPT to generate 500 text completions each for 104 different vignettes (writing prompts). The vignettes involved three independent variables, each with multiple levels, so that we could investigate their effects on the homogeneity of ChatGPT-generated text. The following are the levels of the independent variables we used in our vignettes:

* Race / ethnicity: African, Asian, Hispanic, and White Americans
* Gender: man and woman
* Text format: story, character description, biography, introduction, social media profile, synopsis, narrative, self-introduction, tragic story, funny story, romantic story, horror story, dramatic story

5) **Analyses.** Specify exactly which analyses you will conduct to examine the main question/hypothesis.

To understand whether race / ethnicity and gender changed our key outcome, we will conduct mixed models using the following equation:

Note that the key outcome variable (cosine similarity) will be standardized in the main text to aid interpretability. The results with the raw measurements will be provided in the Supplemental Materials.

6) **Outliers and Exclusions.** Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.

In an earlier pilot study, we noticed that ChatGPT would sometimes avoid answering some of the vignettes and generate responses like, “As an AI language model, I cannot answer this prompt…” or “I’m sorry, but I cannot create content that may be offensive or perpetuate stereotypes.” These responses constituted a small minority of completions, at approximately 1.13% of the pilot study data. We excluded such observations and replaced them with another text completion.

7) **Sample Size.** How many observations will be collected or what will determine sample size? No need to justify the decision but be precise about exactly how the number will be determined.

We will generate 500 text completions per vignette (after exclusions). As there will be four racial / ethnic groups, two gender groups, and 13 different text formats, we will generate 52,000 text completions in total. In our pilot studies, where we generated 10 text completions per vignette, we noticed that the effects were not consistent across iterations. We decided to significantly increase the number of text completions for each vignette to 500, which will lead to approximately 2500 times the number of pairwise cosine similarity measurements for each intersectional group and hence more stable results.

8) **Other.** Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)

We will perform an additional round of analyses where we test the robustness of the results to additional pre-processing steps on the ChatGPT-generated text. In these analyses, we will remove all mentions of race / ethnicity (words like “African”, “Asian”; “Hispanic”, and “White”) and replace all mentions of gender (“man” and “woman”) with the general “person” to control for the extent to which these words contribute to the homogeneity of texts.

9) **Name.** Give a title for this AsPredicted pre-registration. Suggestion: use the name of the project, followed by a study description.

Group status affects variability of group representations in LLM-generated text

10) **Type of study**

Experiment

11) **Data source**

Other: ChatGPT