Pre-Registration for Racial Stereotyping in Multimodal LLMs (Study 2)

Feb 23, 2024

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?

* Does GPT-4 describe African Americans as more homogeneous than White Americans?
* What kind of stereotypes does GPT-4 reproduce when writing texts about African Americans as opposed to White Americans, and vice versa?
* Does prototypicality of African and White American faces relate to stereotypes that GPT-4 produces?
* Is the extent to which prototypicality relates to stereotypes that GPT-4 produces greater for African Americans than for White Americans?

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

HOMOGENEITY OF GPT-4-GENERATED TEXT

* To measure the homogeneity of text generated by the multimodal LLM (i.e., GPT-4), we will compute the pairwise cosine similarity between sentence embeddings of the collected text.
* Sentence embeddings represent the meanings of sentences as numerical vectors. We used Sentence-BERT models, a family of models fine-tuned on models like BERT and RoBERTa to yield sentence embeddings that are suited for similarity assessment (Reimers & Gurevych, 2019). We select three pre-trained Sentence-BERT models that perform best on the sentence encoding task: all-mpnet-base-v2, all-distilroberta-v1, and all-MiniLM-L12-v2. These pre-trained models were assessed on a variety of sentence-encoding-related tasks such as Semantic Textual Similarity (STS), Paraphrase Detection, Question-Answering, and Text Classification. We will use the sentence-transformers package in python (python 3.11.4; Reimers & Gurevych, 2019) to encode the collected text into sentence embeddings.
* Then, we will compute the pairwise cosine similarity between the sentence embeddings of texts generated for each 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. This measure will quantify the extent to which texts about a group are similar to each other.

PREVALENCE OF STEREOTYPES

* To discover attributes that the model commonly associates with African and White Americans, we will use structural topic models to identify common themes in the text and use theta values as measure of topic prevalence. Theta values correspond to the proportion of a document that belongs to the latent topics inside the text corpus.

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

We do not assign participants to conditions, but we will ask GPT-4 to write stories about all 186 images of African and White American males with neutral faces (93 for each group) in the Chicago Face Database (Ma et al., 2015).

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

**Variable Structure**

* **Race**: factor variable indicating the race of the individual in the image stimuli used to calculate the cosine similarity value (i.e., African or White American).
* **Mean Prototypicality**: The mean prototypicality value of the image stimuli used to calculate the cosine similarity value. These prototypicality values are from the Chicago Face Database and were participants’ ratings of how Eurocentric/Afrocentric the faces were (Ma et al., 2015). As two texts were used to calculate cosine similarity, the mean prototypicality value was either calculated using the prototypicality value of one image (if cosine similarity value was calculated using texts generated in response to the same image stimuli) or two different images.
* **Race \* Mean Prototypicality**: The interaction term between race and mean prototypicality. This term allows us to examine whether prototypicality of the image stimuli has a greater effect on the cosine similarity value for African Americans than White Americans.
* **Pair ID**: Unique ID of the pair of image stimuli used to calculate the cosine similarity value.
* **Prototypicality**: The prototypicality value of the image stimuli used to generate text.

**Homogeneity Bias**

We will fit mixed-effects models to evaluate the following hypotheses. The mixed-effects models will look like the following equation:

Depending on the hypothesis in question, predictor variables will include Race, Mean Prototypicality, and their interactions. Pair ID will be entered as the random intercept as we expect the cosine similarity baseline to vary depending on the image pair, but we do not expect the magnitude and direction of the effects of Race and/or Mean Prototypicality to vary depending on the image pair.

**The Race and Mean Prototypicality Terms**

*We hypothesize that cosine similarity between GPT-4-generated texts about African Americans will be greater than White Americans.*

*We hypothesize that cosine similarity between GPT-4-generated texts from image stimuli with higher mean prototypicality will be greater than those from image stimuli with lower mean prototypicality.*

* We will fit a mixed-effects model including Race and Mean Prototypicality as the two predictor variables.

**The Interaction Term (Race \* Mean Prototypicality)**

*We hypothesize that the effect of race on the homogeneity of GPT-4-generated texts will be greater for African Americans than White Americans.*

* We will fit a mixed-effects model including Race, Mean Prototypicality, and their interaction as predictor variables.

**Trait Associations**

We will fit a single Structural Topic Model (STM) to evaluate the following hypotheses.

We will first use the searchK function of the stm package in R (R Version 4.3.1) to determine the optimal number of topics (K) to identify in the collected text and use the value to fit the optimal model. The STM will look like the following equation where Race, Prototypicality, and their interactions are included as predictors to model the prevalence of all K topics.

**The Race Term**

*We hypothesize that different topics will be more prevalent for African Americans than White Americans and for White Americans than African Americans.*

* We will look at the Race term of the STM using the estimateEffect function of the stm package.

**The Prototypicality Term**

We do not expect prototypicality of the images to relate to prevalence of topics in general.

**The Interaction Term (Race \* Prototypicality)**

*We hypothesize that the effect of race on the prevalence of topics will be greater for faces with higher prototypicality values.*

* We will look at the Race \* Prototypicality term of the STM using the estimateEffect function of the stm package.

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

Previous work studying the same bias in LLMs have noted that LLMs would sometimes avoid answering writing prompts and generate responses like, “I'm sorry, but I can't assist with that request.” or “I’m sorry, but I cannot create content that may be offensive or perpetuate stereotypes.” We will exclude these observations from the analysis.

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 use all 93 images of African and White American men with neutral faces included in the Chicago Face Database and collect 50 text completions per image. We collected the same number of text completions per image as we did in Study 1.

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

Exploratory Analyses

* We will focus on specific features within the generated text (e.g., names) to see if GPT-4 associates certain groups with these features and test whether certain features are disproportionately associated with one group.

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

Racial Stereotyping in Multimodal LLMs (Study 2)

10) **Type of study**

Experiment

11) **Data source**

Other: GPT-4 (GPT-4 with Vision)