

CONFIDENTIAL - FOR PEER-REVIEW ONLY

LLM Conspiracy Persuasion, Sample 2 (#163392)

Created: 02/23/2024 08:53 AM (PT)

This is an anonymized copy (without author names) of the pre-registration. It was created by the author(s) to use during peer-review.
A non-anonymized version (containing author names) should be made available by the authors when the work it supports is made public.

1) Have any data been collected for this study already?

No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?

- H1. GPT-4 can persuade people to reject their favored conspiracy theories.
- H2. Will this intervention work for all types of conspiracy theories?
- H3. Will the effect be moderated by trust in generative AI, age, gender, dispositional overconfidence, political affiliation, race, or education?
- H4. Will the effect extend to behavioral (and related) outcomes, including information-seeking and a suite of behavioral intentions?

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

The key dependent variable is person-specific conspiracy beliefs. It will be measured on the basis of participants' answers to an open-ended question about their favored conspiracy theory. Following each open-ended response, GPT-4 will summarize their question as a declarative, high-level statement of belief (e.g., "JFK was assassinated by the CIA"). Participants will be shown this summary and asked, "On a scale of 0% to 100%, please indicate your level of confidence that this statement is true." This question will be administered pre- and post-manipulation.

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

Participants will be assigned to one of four conditions, one of which is the treatment, and the remaining three are controls.

In the treatment condition, participants will carry out a 3-round conversation with GPT-4. The model will (1) be provided with the participant's chosen conspiracy theory and rationale and (2) be instructed to persuade the participant against their chosen conspiracy belief.

In the control conditions, participants will also carry out a 3-round conversation with GPT-4. However, they will not discuss conspiracy theories. In Control 1, they will discuss their views on, and experiences with, the American medical system. In Control 2, they will discuss their experiences with firefighters. And in Control 3, they will discuss whether they prefer cats or dogs.

The control conditions will be pooled in our analyses.

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

For H1, we will use linear regression to assess changes in belief levels before and after the intervention, controlling for participants initial levels of conspiratorial belief. H1 Model: Post-belief ~ Dummy-coded treatment vs. control + Pre-belief.

For H2, we will extend the linear model described for H1 to include a set of dummy-coded variables indicating the presence of particular conspiracy theories. H4 Model: Post-belief ~ Dummy-coded treatment vs. control × Dummy-coded conspiracy theory type + Pre-belief.

Conspiracy theory types will be determined on the basis of a cluster analysis of text embeddings of the GPT-standardized open-ended conspiracy responses. Particularly, we will generate the text embeddings using OpenAI's text-embedding-3-large model and use the k-means clustering algorithm to group these embeddings. The object of this analysis will be a matrix of cosine similarities between the embeddings. To ascertain the optimal number of clusters, we will use a range of methods, including the silhouette score, gap statistic, and within groups sum of squares.

For H3 we will extend the linear model described for H1 to include each hypothesized moderator and its interaction with the dummy-coded treatment condition (simultaneously).

For H4, we use the behavioral outcome measures as DVs in a set of linear regression models with the dummy-coded treatment condition as the IV.

For all regression-based analyses, p-values and confidence intervals will be computed using robust standard errors.

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

Participants who do not indicate supporting a conspiracy belief will not be included in our analyses. This determination will be made by GPT-4, based on the open-ended question soliciting a conspiracy belief. Similarly, if respondents advance a conspiratorial belief but express skepticism or ambivalence about its veracity (as indicated by a response < 50 / 100 on the pre-treatment scale), they will not be included in our analyses.

We also plan to collect a host of data concerning the accuracy and coherence of participants' responses. Participants determined to be using automated responding (e.g., generative AI), based on being "flagged" by the Roundtable Alias algorithm, or who provide inaccurate responses (failed attention checks) prior to the treatment will be removed from our analyses.

Further, participants who complete the experiment in fewer than 600 seconds, indicating a lack of engagement, will not be included in our analyses. If differential exclusion is observed (i.e., if the "speeders" are disproportionately found in the treatment condition), we will perform sensitivity analyses to see how the results change if the speeders are included.

We will collect a sample of 1000 individuals. 75% will be assigned to the treatment group and 25% will be randomly split across the control groups.

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

We will collect a sample of 1000 individuals. 75% will be assigned to the treatment group and 25% will be randomly split across the control groups.

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

Information-seeking behavior will be assessed on an expiatory basis, as it may be that successful treatment (i.e., debunking of one's conspiratorial belief) sates a need for more information about the conspiracy theory. To assess information-seeking, we plan to use GPT-4 to generate key search terms based on each participant's conversation (in the treatment condition) or original conspiracy theory (in the control condition). These search terms will be designed to identify well-researched, well-sourced news articles about the conspiracy in question (or related events) using the GNews API. The two most relevant such articles will be shown to participants (i.e., headlines, image, description, and source), who will be given the option of following a link to read the full article or proceeding to the next stage of the study. Clicks and time spent on the page will be recorded. Further, we will assess participants' willingness to share the articles on social media. In exploratory analyses, we may examine any relations between source-quality, article topic, and treatment efficacy.

To assess behavioral intentions, we will administer brief self-report items targeting (1) engagement with groups promoting conspiracies, (2) social discussions concerning conspiracy theories, and (3) willingness to engage with petitions or protest movements surrounding conspiracies.

Further, we plan to conduct a series of NLP-based analyses to probe and understand the nature of the intervention.

All hypothesis tests will be two-tailed with an alpha level set at 0.05, and we will report effect sizes and confidence intervals for all findings.