

# CONFIDENTIAL - FOR PEER-REVIEW ONLY

## LLM Conspiracy Persuasion, Sample 2 Extension (#165307)

Created: 03/07/2024 11:37 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, political affiliation, race, or education?
- H4. Will the effect extend to behavioral (and related) outcomes, including signing an anti-conspiracy petition and 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. We plan to run this cluster analysis using all data collected for this project (not just this study).

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 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.

60% of participants will be assigned to the treatment group and 40% 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 1300 individuals, in order to achieve a sample of roughly  $n = 1000$  people who supply valid conspiracy theories.

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

This study serves as a follow-up, and slight correction, to virtually-identical study we recently conducted (AsPredicted #163392). We noticed, after running this prior study, that two of the four behavioral intentions questions had problems. Firstly, one the questions was phrased ambiguously, such that did not clarify whether it was referring to behaviors supporting or opposing people's chosen conspiracy (i.e., "If people you knew were going to engage in a protest or action in response to the theory you described, how likely would you be to join in?"). We have revised this item to read as "If people you knew were going to engage in a protest or action in support of to the theory you described, how likely would you be to join in?". Secondly, one of the questions contained a double negative (between the statement and response options) and was stated in the direction opposing several previous questions. To account for this issue, we now present the questions on separate pages and visually emphasize the statement direction using bolding.

Further, to extend and corroborate participants responses on these questions, we now administer a GPT-generated petition, expressing opposition to their chosen conspiracy theory, and provide participants with the option of signing the petition. This will serve as a quasi-behavioral outcome with greater ecological validity.

To account for the nested nature of these two studies, we plan to pool all participant responses from AsPredicted #163392 and the present study for H1:H3 (and all analyses pertinent to those hypotheses). For H4, which pertains to the behavioral outcomes, we will present the results separately.