# DRAFT: Artificially Intelligent RCT Pilot: Afro-Barometer and Candour II

Raymond Duch
Oxford

Piotr Kotlarz
Oxford

Raymond Low
Oxford

Kento Ohara
Oxford

Benjamin S. Manning
MIT

September 8, 2024

**Abstract**

This essay reports the results of preliminary efforts at understanding the performance of LLM models in predicting the behavior of human subjects in experimental settings. In particular, we focus on samples of humans who are likely to be under-represented in the training corpara and experiments that are extremely time-consuming, expensive, and have large policy implications for social welfare; vaccine uptake in the global south. On balance, the results are encouraging but there is room for considerable progress. In the case of the Afro-Barometer pilot study, the distribution of responses were similar for both LLMs and humans. But the LLM was unable to replicate the RCT treatment effects that we observed with humans in rural Ghana. The LLM model performed well in predicting the vaccination uptake for the persona created from subjects in the 16-country CANDOUR survey. Surprisingly, given the concerns about bias in LLM models, the LLM did poorly in predicting vaccination update by the American persona.

# Introduction

Randomized Controlled Trials (RCTs), particularly in low- and middle-income countries (LMICs), have become a staple of policy design and evaluation (Duflo, 2017). The policy relevance of these RCTs is very much determined by our ability to scale-up results and generalize—but generating samples that reflect the characteristics and diversity of populations of interest can be extremely expensive (List, 2024). LLMs may provide an option to alleviate some of these problems. In this paper, we report on the results of efforts to use LLMs as proxies for human subjects in RCTs.

Our goal is to explore the extent to which AI agents in synthetic experiments can be informative. There are various aspects of RCT design and implementation that might benefit from artificial intelligence. In this essay, we focus specifically on whether synthetic RCTs can be informative for policy design and evaluations. Are we likely to learn much from implementing a synthetic RCT with AI agents that resemble human subjects? The challenge is to develop a strategy for prompting LLMs to create synthetic agents that make choices that resemble those of human subjects. If we can begin to outline the boundaries of when AI subjects are informative proxies for humans, the savings in money, time, and resources more generally could be enormous.

We are not the first to pursue such a goal. Previous work has already established that synthetic experiments can be highly predictive of never-before-seen human subjects experiments (Binz and Schulz, 2023; Brand et al., 2023; Hewitt et al., 2024; Li et al., 2024). What is novel in this paper is twofold: (i) we are studying populations likely underrepresented in the training corpus—focusing on Western Africans. And (2) we try to predict the results of RCTs exploring interventions for high-impact health behaviors. I.e., experiments that attempt to improve vaccine updates and infectious disease testing. This is in contrast to most other research exploring synthetic subject behavior, which, as far as we are aware, has not studied such consequential decision-making contexts.

In a pre-registered report, we will describe our implementation of a synthetic RCT that replicates the Ghana COVID-19 Financial Incentives study(Duch et al., 2023). As a prelude to implementing this synthetic RCT, we explored strategies for prompting LLMs to most accurately simulate human subjects' responses. In this pilot phase of our synthetic experiment, we develop prompts using the Afro-Barometer survey and the 16-country CANDOUR Wave II survey. These surveys questioned thousands of participants about their COVID-19 vaccination status and intentions in Ghana (Afro-

Barometer) and other countries in the global south (CANDOUR). We begin with a brief summary of the rapidly growing literature on leveraging LLMs for social science research. The two sets of preliminary results follow.

## Background

There is growing evidence that synthetic LLM-powered personas or “AI subjects” can behave similarly to humans in contexts ranging from cross-sectional surveys to marketing and economic experiments (Aher et al., 2023; Argyle et al., 2023; Binz and Schulz, 2023; Brand et al., 2023; Horton, 2023; Li et al., 2024; Manning et al., 2024; Mei et al., 2023; Tsuchihashi, 2023). For example, a recent study re-estimated the treatment effects for 70 national U.S. online experiments and then compared the treatment effects to those same experiments replicated with AI subjects (Hewitt et al., 2024). The correlation was a very impressive 0.85 for all studies and 0.90 for studies guaranteed not to be in the LLM's training corpus. The results even held for various demographic subgroups.

But such results are not a panacea. (Bisbee et al., 2024), compare a AI subjects with their equivalent humans in the ANES data and conclude that there is less variation in responses than in the real surveys. Regression coefficients often differ significantly from equivalent estimates obtained using humans. AI subjects also failed to reflect human-like behavior in survey question bias experiments conducted with Pew's ATP survey data (Tjuatja et al., 2024). On certain topics, those related to U.S. partisan politics in particular, AI subjects' opinions are poorly aligned with those of U.S. demographic sub-groups (Santurkar et al., 2023). Researchers, recognizing these limitations, are developing strategies for improving the ”fidelity” of AI subjects' behavior with human behavior (Moon et al., 2024). Especially relevant to our purposes is some evidence from survey data that AI subjects are poor proxies for human subjects from non-WEIRD populations (Atari et al., 2023) and can caricature responses from minority populations (Cheng et al., 2023)—the exact populations.[1] Furthermore, other evidence suggests that AI subjects exhibit socio-cultural biases that reflect their training corpus that originate in higher income countries (HICs) and specifically in the U.S (AlKhamissi et al., 2024; Naous et al., 2024). Such results

To be clear, these references and this current paper are not directly eliciting treatment effects

[1] Western, Educated, Industrialized, Rich, Democratic (WEIRD).

from an LLM. Rather, we are providing the LLM with a complete profile of an experimental synthetic subject (hence “AI subjects”) one at a time. The AI subjects are assigned to each treatment condition. We then query the AI subjects on the choices or behaviors they would select in their respective treatment conditions.

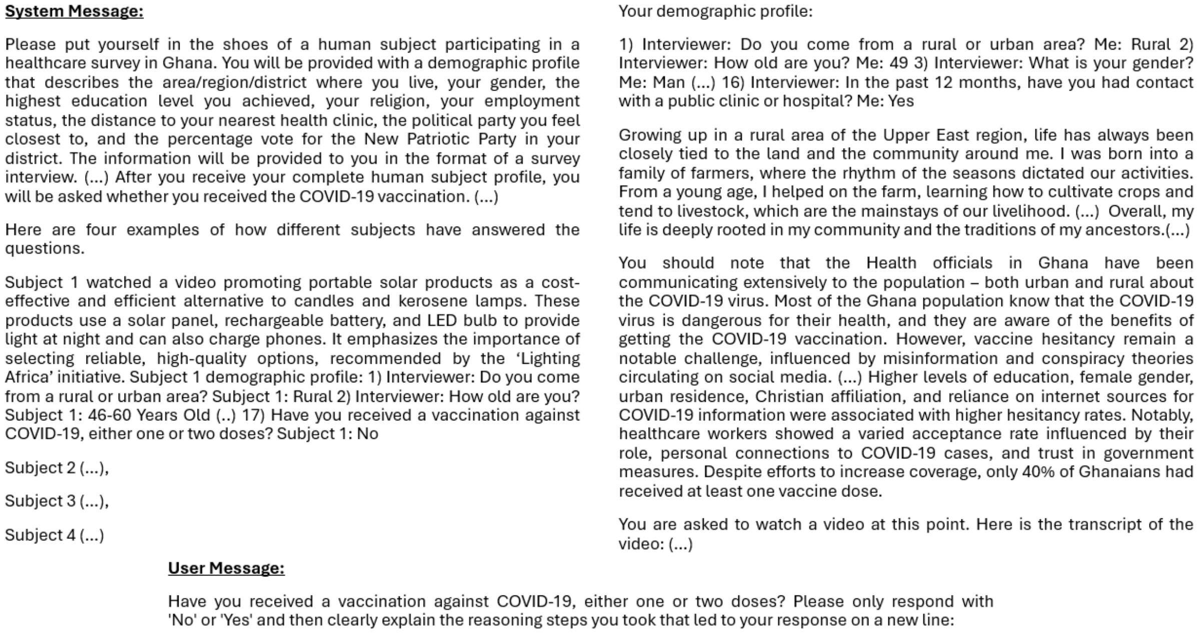
# Design and Results

The pilot phase implements two online versions of the Ghana Financial Incentive Wave I clustered randomized controlled trial that was conducted with 6,963 residents in six rural Ghana Districts. In the original study, villages randomly received one of four treatment arms: a placebo, a standard video health message, a video message with a high cash incentive ($10), and a video message with a low cash incentive ($3). The verified vaccination status of subjects was a primary outcome of interest: in the Cash treatment arm, 36.6% of verified subjects had at least one dose of the COVID-19 vaccine compared to 30.3% for those in the Placebo - a difference of 6.3%. The RCT confirmed that cash incentives have a positive impact on vaccine uptake (Duch et al., 2023). To construct prompts to replicate this experiment with AI subjects, we first optimize the prompts for the Ghanaian data from the Afro-Barometer survey and the 16-country CANDOUR II survey that was conducted in 2022 (Duch et al., 2024).

## Online Ghana AI subjects Experiment

Our first study is modeled on the the 2022 randomized cluster trial we conducted in rural Ghana that found that financial incentives increased COVID-19 vaccine uptake (Duch et al., 2023). Our outcome variable is whether subjects received the COVID-19 vaccination. The original RCT (Duch et al., 2023) found that non-vaccinated participants, assigned to one of two financial incentive treatment arms, were about 6% more likely to to get the COVID-19 vaccine. Participants in a Cash treatment arm had an average vaccination rate of about 36% compared to 30% for those in the Placebo treatment arm. Subjects in the simple health message treatment arm had a vaccination rate of 22% – 8% lower than the Placebo arm. We will consider these RCT treatment effects as the ”ground truth” benchmarks for our pilot online AI subjects RCT.

The AI subjects’ profiles are based on human participants in Afro-Barometer surveys that



**System Message:**

Please put yourself in the shoes of a human subject participating in a healthcare survey in Ghana. You will be provided with a demographic profile that describes the area/region/district where you live, your gender, the highest education level you achieved, your religion, your employment status, the distance to your nearest health clinic, the political party you feel closest to, and the percentage vote for the New Patriotic Party in your district. The information will be provided to you in the format of a survey interview. (...) After you receive your complete human subject profile, you will be asked whether you received the COVID-19 vaccination. (...)

Here are four examples of how different subjects have answered the questions.

Subject 1 watched a video promoting portable solar products as a cost-effective and efficient alternative to candles and kerosene lamps. These products use a solar panel, rechargeable battery, and LED bulb to provide light at night and can also charge phones. It emphasizes the importance of selecting reliable, high-quality options, recommended by the 'Lighting Africa' initiative. Subject 1 demographic profile: 1) Interviewer: Do you come from a rural or urban area? Subject 1: Rural 2) Interviewer: How old are you? Subject 1: 46-60 Years Old (..) 17) Have you received a vaccination against COVID-19, either one or two doses? Subject 1: No

Subject 2 (...),

Subject 3 (...),

Subject 4 (...)

Your demographic profile:

1) Interviewer: Do you come from a rural or urban area? Me: Rural 2) Interviewer: How old are you? Me: 49 3) Interviewer: What is your gender? Me: Man (...) 16) Interviewer: In the past 12 months, have you had contact with a public clinic or hospital? Me: Yes

Growing up in a rural area of the Upper East region, life has always been closely tied to the land and the community around me. I was born into a family of farmers, where the rhythm of the seasons dictated our activities. From a young age, I helped on the farm, learning how to cultivate crops and tend to livestock, which are the mainstays of our livelihood. (...) Overall, my life is deeply rooted in my community and the traditions of my ancestors.(...)

You should note that the Health officials in Ghana have been communicating extensively to the population – both urban and rural about the COVID-19 virus. Most of the Ghana population know that the COVID-19 virus is dangerous for their health, and they are aware of the benefits of getting the COVID-19 vaccination. However, vaccine hesitancy remain a notable challenge, influenced by misinformation and conspiracy theories circulating on social media. (...) Higher levels of education, female gender, urban residence, Christian affiliation, and reliance on internet sources for COVID-19 information were associated with higher hesitancy rates. Notably, healthcare workers showed a varied acceptance rate influenced by their role, personal connections to COVID-19 cases, and trust in government measures. Despite efforts to increase coverage, only 40% of Ghanaians had received at least one vaccine dose.

You are asked to watch a video at this point. Here is the transcript of the video: (...)

**User Message:**

Have you received a vaccination against COVID-19, either one or two doses? Please only respond with 'No' or 'Yes' and then clearly explain the reasoning steps you took that led to your response on a new line:

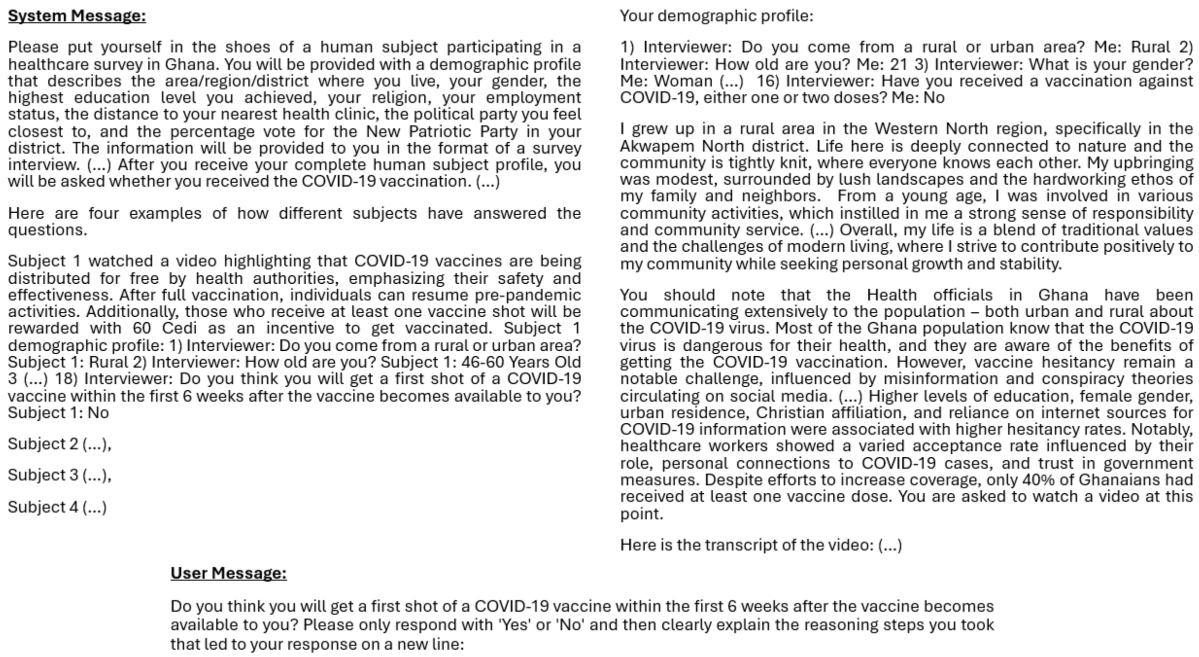
Fig 1: Afro-Barometer Replication System Prompts

have been conducted in Ghana since 1999.[2] These surveys met our three criteria: extensive socio-demographic measures; COVID-19 vaccination questions; geo-coded location of respondents. We identified 2,366 respondents who met these three criteria.

*Prompting the LLM.* Figure 1 summarizes the system and user prompts. Unless otherwise specified, we used GPT-4o—the most advanced model developed by OpenAI to date. We elicit the COVID-19 vaccination status from an AI subject based on each human subject in the Afro-Barometer Ghana sample.

We provide the LLM with system prompts that include the following information for each AI subject: rural-urban residence; age; gender; education; religion; employment status; region; political party identification; discuss political matters; geo-location; distance to the nearest health clinic; district; percentage of the district voting for the National Democratic Congress; percentage of the district voting for the New Patriotic Party; contact with a public clinic or hospital. The information is presented to GPT-4o in a survey interview format. We adopt a few-shot design that provides GPT-4o with four examples of actual human decisions. These examples are followed by a short paragraph describing the COVID-19 vaccination context in Ghana.

[2] Afrobarometer Data, available at `http://www.afrobarometer.org`.



**System Message:**

Please put yourself in the shoes of a human subject participating in a healthcare survey in Ghana. You will be provided with a demographic profile that describes the area/region/district where you live, your gender, the highest education level you achieved, your religion, your employment status, the distance to your nearest health clinic, the political party you feel closest to, and the percentage vote for the New Patriotic Party in your district. The information will be provided to you in the format of a survey interview. (...) After you receive your complete human subject profile, you will be asked whether you received the COVID-19 vaccination. (...)

Here are four examples of how different subjects have answered the questions.

Subject 1 watched a video highlighting that COVID-19 vaccines are being distributed for free by health authorities, emphasizing their safety and effectiveness. After full vaccination, individuals can resume pre-pandemic activities. Additionally, those who receive at least one vaccine shot will be rewarded with 60 Cedi as an incentive to get vaccinated. Subject 1 demographic profile: 1) Interviewer: Do you come from a rural or urban area? Subject 1: Rural 2) Interviewer: How old are you? Subject 1: 46-60 Years Old 3 (...) 18) Interviewer: Do you think you will get a first shot of a COVID-19 vaccine within the first 6 weeks after the vaccine becomes available to you? Subject 1: No

Subject 2 (...),

Subject 3 (...),

Subject 4 (...)

Your demographic profile:

1) Interviewer: Do you come from a rural or urban area? Me: Rural 2) Interviewer: How old are you? Me: 21 3) Interviewer: What is your gender? Me: Woman (...) 16) Interviewer: Have you received a vaccination against COVID-19, either one or two doses? Me: No

I grew up in a rural area in the Western North region, specifically in the Akwapem North district. Life here is deeply connected to nature and the community is tightly knit, where everyone knows each other. My upbringing was modest, surrounded by lush landscapes and the hardworking ethos of my family and neighbors. From a young age, I was involved in various community activities, which instilled in me a strong sense of responsibility and community service. (...) Overall, my life is a blend of traditional values and the challenges of modern living, where I strive to contribute positively to my community while seeking personal growth and stability.

You should note that the Health officials in Ghana have been communicating extensively to the population – both urban and rural about the COVID-19 virus. Most of the Ghana population know that the COVID-19 virus is dangerous for their health, and they are aware of the benefits of getting the COVID-19 vaccination. However, vaccine hesitancy remain a notable challenge, influenced by misinformation and conspiracy theories circulating on social media. (...) Higher levels of education, female gender, urban residence, Christian affiliation, and reliance on internet sources for COVID-19 information were associated with higher hesitancy rates. Notably, healthcare workers showed a varied acceptance rate influenced by their role, personal connections to COVID-19 cases, and trust in government measures. Despite efforts to increase coverage, only 40% of Ghanaians had received at least one vaccine dose. You are asked to watch a video at this point.

Here is the transcript of the video: (...)

**User Message:**

Do you think you will get a first shot of a COVID-19 vaccine within the first 6 weeks after the vaccine becomes available to you? Please only respond with 'Yes' or 'No' and then clearly explain the reasoning steps you took that led to your response on a new line:

Fig 2: Afro-Barometer AI subjects Experiment System Prompts

We use an anthology prompting strategy that provides richer profiles of AI subjects, which has been shown to improve the responses of LLMs acting as human proxies (Moon et al., 2024). Instead of just telling the AI subject their set of demographic traits/preferences (e.g., "you are a white female..."), the anthology strategy uses coherent backstories that contain the same endowed information. A separate LLM generates a backstory for each set of demographic information, and then this backstory is endowed to an AI subject. We then query each AI subject as to whether or not they received a COVID-19 vaccination.

*Pilot AI subjects Experiment.* We next employ a prompting strategy that is based on the piloting described in the previous section. Figure 2 provides an example of the system and user prompts. The system prompts provide GPT-4o with information that is similar to what we developed above for each AI subject. The socio-demographic information is presented in an interview format. And we condition the persona on the backstory life-narratives generated by LLMs.

Again, we adopt a few-shot design that provides GPT-4o with four examples of an AI subject corresponding to the appropriate treatment arm video. We remind the LLM that the intention to get the COVID-19 vaccination is taken in a context in which the Ghana Health District is promoting and making available the vaccine. These examples are followed by a short paragraph describing the

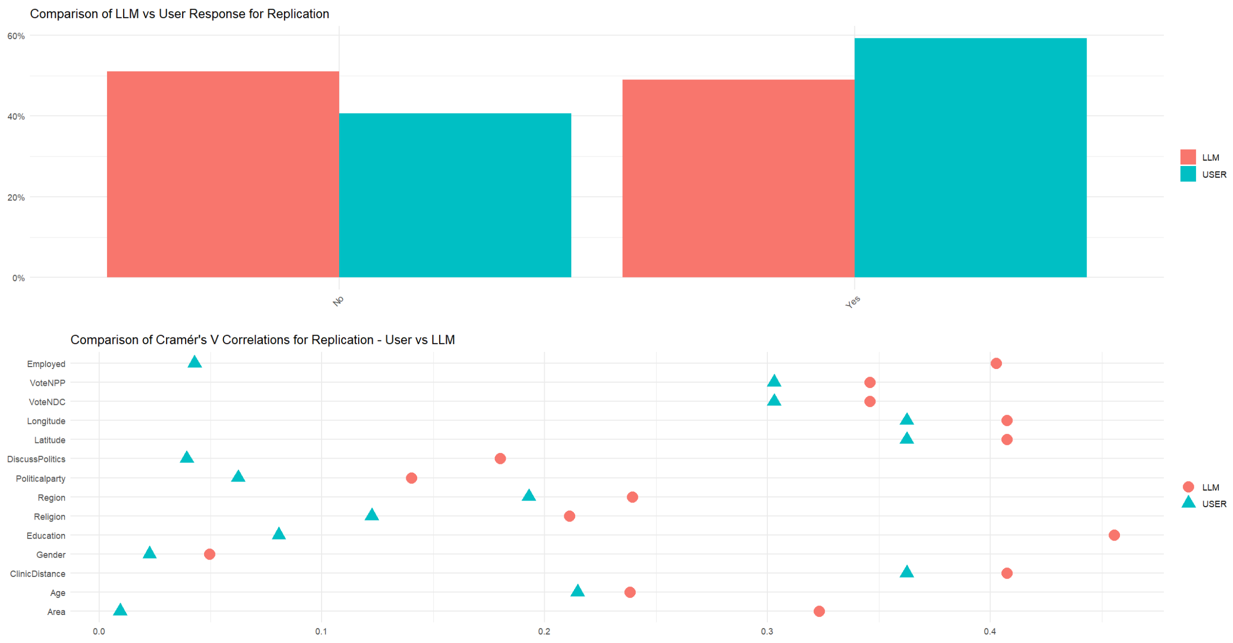


Fig 3: Afro-Barometer Ghana: LLM Replicated COVID-19 Vaccine Status

COVID-19 vaccination context in Ghana. The LLM's demographic profile again is presented in an interview format. This is followed by the transcript of the video treatment assigned to the LLM. Having reviewed the video treatment transcript, the LLM is then asked to indicate whether they would get a first shot of a COVID-19 vaccine within the first 6 weeks after the vaccine becomes available.

**Results** For this pilot phase we report two sets of results. First, we benchmark human COVID-19 vaccination status versus the AI subjects reported vaccination status employing prompt-engineering with backstory condition of the LLM. We then present the results for our synthetic RCT.

*Prompt Engineering plus Backstories.* First, we evaluate our prompting strategy by comparing the vaccination status LLM predictions for the AI subjects with the actual status of their human counterparts. Given a complete set of individual socio-demographic characteristics how well can we predict this particular health decision by our AI subjects? The benchmark is the actual COVID-19 decisions of the persona versus human subjects.

The top panel of Figure 3 compares the distribution of yes-no vaccination responses for the AI subjects with those of the human subjects. In the case of the LLM predictions, there is about an even split between non-vaccinated and the vaccinated. The actual responses of the human subjects split 60% ”yes” versus 40% ”no”. The lower panel of Figure 3 presents, for both AI subjects and humans, the Cramer V correlations between the outcome variable, vaccination status, and covariates. Interestingly, the correlations are always larger for the AI subjects. However, they are still generally quite close to the human subjects correlations. Such results are not necessarily surprising for two reasons: (i) we already know that LLMs have a tendency to respond with “caricatures” for a given personality (Cheng et al., 2023), and (2) these covariates are the only information that is provided to the AI subjects. In contrast, there are many unobserved attributes for each human subject.

*Pilot AI Subjects Experiment.* Our second set of results concerns the synthetic RCT. We run four separate versions of the model—each time using our 2,366 AI subjects. The different prompts corresponded to one of the four identical treatment arms we implemented in the original Ghana RCT (Duch et al., 2023). The top panel of Figure 4 presents the percentage of AI subjectsin each of the four treatment arms indicating they did or did not get the COVID-19 vaccination. We observe little evidence of a treatment effect in the synthetic experiment—roughly 80% of the AI subjects in all four treatment arms reported getting vaccinated in the six-week period after receiving the video treatments. In the lower panel of Figure 4, we present as a comparison the vaccination rates, post-intervention, of the human subjects in the Ghana Wave I RCT (Duch et al., 2023). Two differences stand out: The percentage of vaccinated subjects post-intervention is lower than what we observed with the synthetic experiment. COVID-19 vaccination uptake is around 70% for subjects in the placebo and health treatment arms while it is closer to 80% for subjects in the low and high cash treatment arms. Secondly, we observe a treatment effect in the case of the human subjects—we do not observe one in the case of the synthetic experiment.

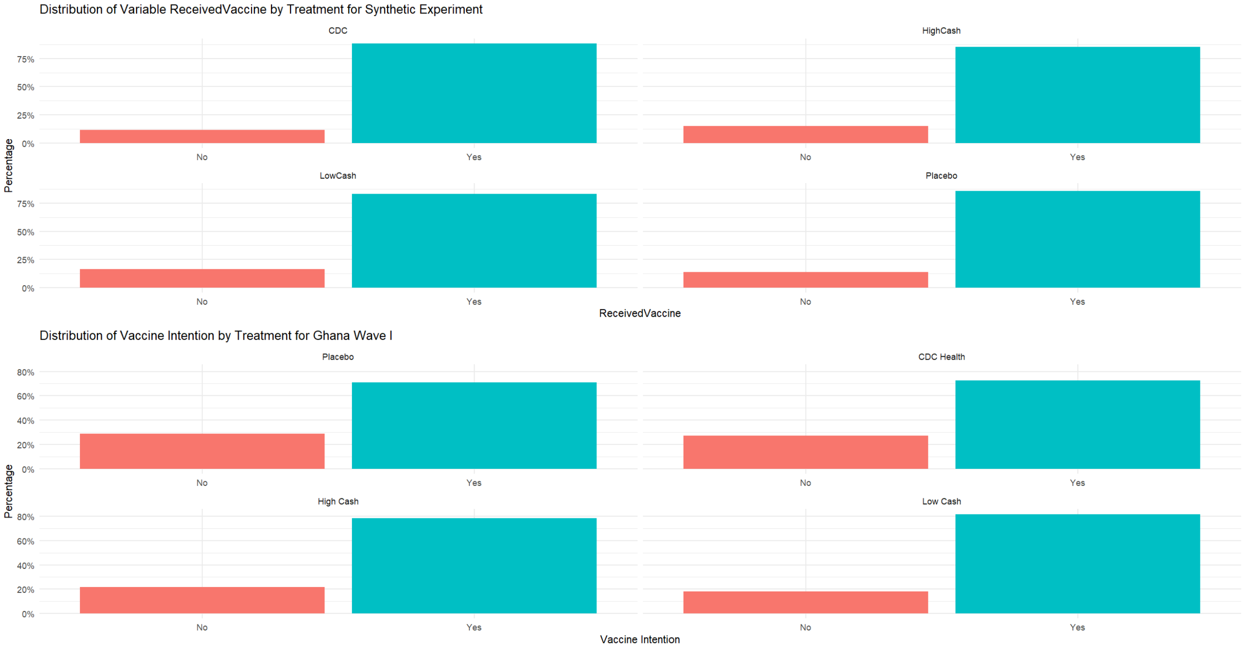


Fig 4: Afro-Barometer Ghana: LLM Replicated COVID-19 Vaccine Status

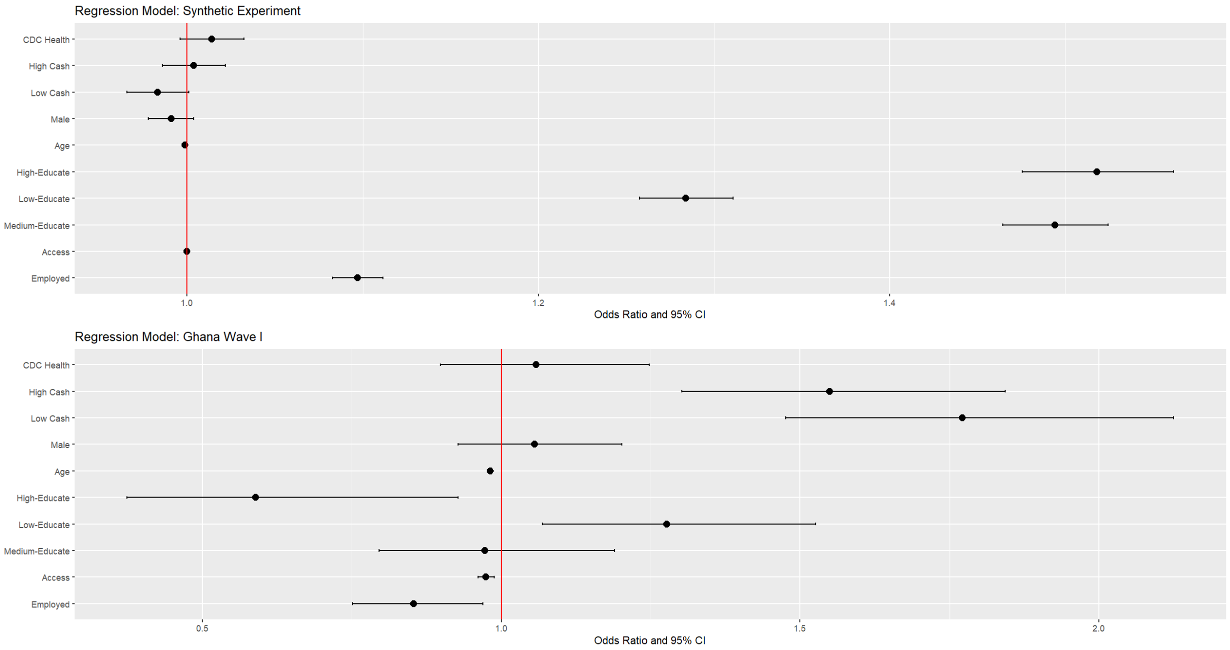


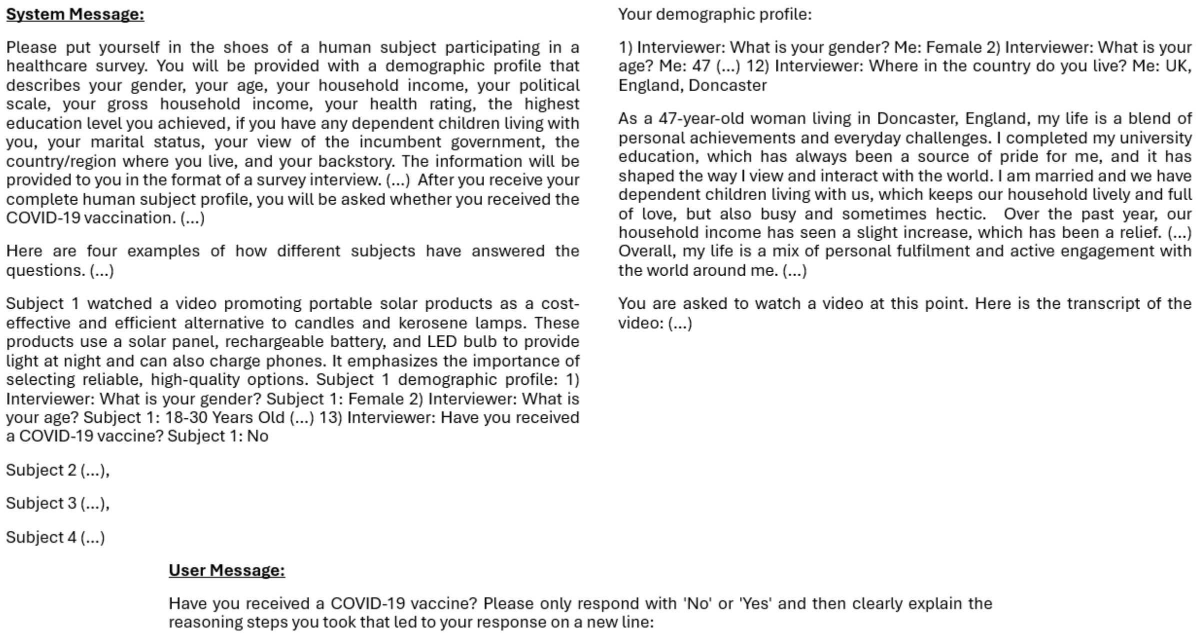
Fig 5: Odds Ratios: AI subjects Afro-Barometer vs Ghana Wave I

To better understand the magnitude of treatment effects in contrast to the relationship between vaccine update and other covariates, we regressed vaccination status on treatment arms along with covariate controls. The upper panel of Figure 5 presents the odd ratios for the synthetic experiment. None of the treatment arms have odds ratios distinguishable from 1.0. The lower panel Figure 5 presents odds ratios from the Ghana Wave I RCT and here we do observe significant treatment effects for the low and high cash treatment arms.

**Conclusion Afro-Barometer.** In this first phase of the synthetic RCT project, our efforts focused on prompting an LLM to predict the health decisions of AI subjects, in the first person, and compared the results to the responses of the corresponding human subjects. The distributions of vaccine decisions by AI subjects and human subjects are not dramatically different. And the socio-demographic covariates of AI subjects and human subjects have similar patterns of correlations with their vaccine decisions. A second exercise replicates the Ghana Wave I RCT with our 2,366 AI subjects and compares treatment effects in the online synthetic RCT with those reported for the actual RCT conducted with humans (Duch et al., 2023). The distributions of responses within each of the four treatment arms were roughly similar for both the synthetic version of the RCT and the version conducted with humans. Estimated treatment effects, on the other hand, were different: The human RCT registered a significant financial incentive treatment effect while we observed no effect in the case of the AI subjects.

## CANDOUR Wave II Replication Synthetic Experiment

Ghanaians, and more generally the Global South, are considered to be difficult to approximate with AI subjects (Atari et al., 2023). This is because such populations generate a far smaller portion of the text used to train many of the most capable LLMs. The second phase of our pilot study explores the extent to which the performance of AI subjects is conditional on socio-economic and cultural context—the very groups which would likely benefit the most from effective synthetic proxies. We use the prompting strategies we developed for the Ghana context and assess their performance in very diverse cultural contexts (i.e., many different countries) that also vary dramatically in their distribution of income. Again, the outcomes of interest are the LLM predicted vaccine status of persona and estimated treatment effects from a synthetic RCT that is based on the design of our 2022 Ghana Wave I RCT. Here the system prompts employed to develop the LLM persona are based on the 22,147 human subjects that participated in the 16-country CANDOUR II survey. The outcomes of interest will be similar to those described for the online Ghana study: COVID-19 vaccination rates and treatment effect sizes. In this essay we only report the results for five of these countries: the U.S., Ghana, Italy, India and Brazil.



**System Message:**

Please put yourself in the shoes of a human subject participating in a healthcare survey. You will be provided with a demographic profile that describes your gender, your age, your household income, your political scale, your gross household income, your health rating, the highest education level you achieved, if you have any dependent children living with you, your marital status, your view of the incumbent government, the country/region where you live, and your backstory. The information will be provided to you in the format of a survey interview. (...) After you receive your complete human subject profile, you will be asked whether you received the COVID-19 vaccination. (...)

Here are four examples of how different subjects have answered the questions. (...)

Subject 1 watched a video promoting portable solar products as a cost-effective and efficient alternative to candles and kerosene lamps. These products use a solar panel, rechargeable battery, and LED bulb to provide light at night and can also charge phones. It emphasizes the importance of selecting reliable, high-quality options. Subject 1 demographic profile: 1) Interviewer: What is your gender? Subject 1: Female 2) Interviewer: What is your age? Subject 1: 18-30 Years Old (...) 13) Interviewer: Have you received a COVID-19 vaccine? Subject 1: No

Subject 2 (...),

Subject 3 (...),

Subject 4 (...)

Your demographic profile:

1) Interviewer: What is your gender? Me: Female 2) Interviewer: What is your age? Me: 47 (...) 12) Interviewer: Where in the country do you live? Me: UK, England, Doncaster

As a 47-year-old woman living in Doncaster, England, my life is a blend of personal achievements and everyday challenges. I completed my university education, which has always been a source of pride for me, and it has shaped the way I view and interact with the world. I am married and we have dependent children living with us, which keeps our household lively and full of love, but also busy and sometimes hectic. Over the past year, our household income has seen a slight increase, which has been a relief. (...) Overall, my life is a mix of personal fulfilment and active engagement with the world around me. (...)

You are asked to watch a video at this point. Here is the transcript of the video: (...)

**User Message:**

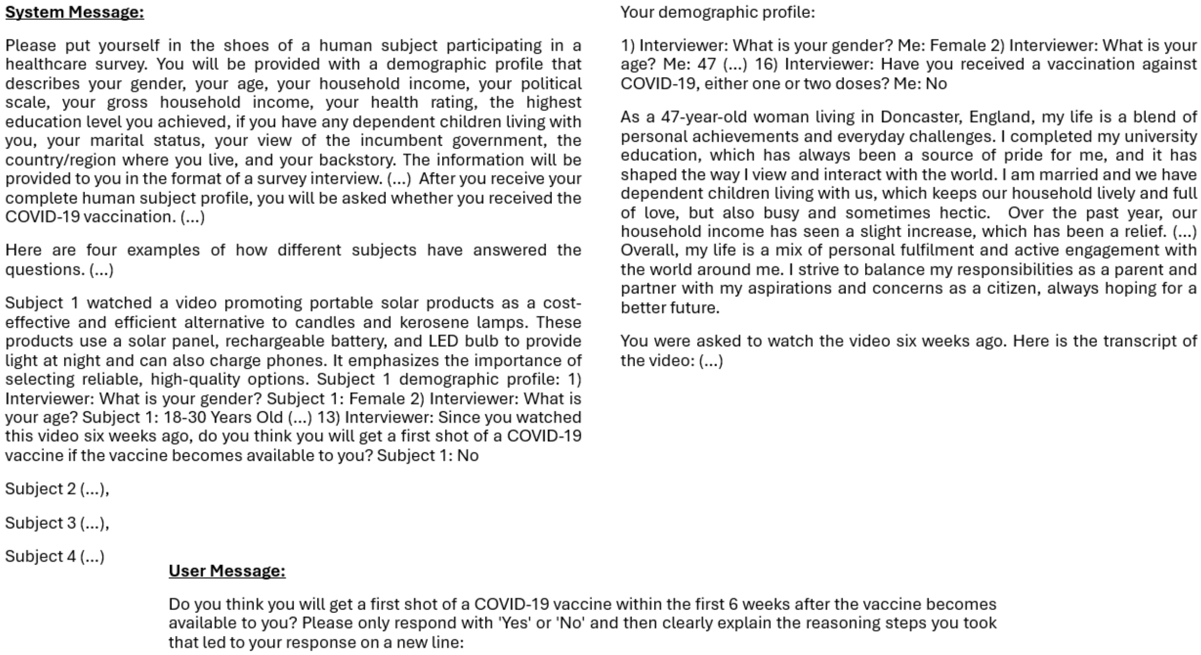
Have you received a COVID-19 vaccine? Please only respond with 'No' or 'Yes' and then clearly explain the reasoning steps you took that led to your response on a new line:

Fig 6: Candour Wave II Replication System Prompts

**Design** In replicating the original Ghana Wave I RCT (Duch et al., 2023), the synthetic CANDOUR II RCT closely follows the design we adopted for the Afro-Barometer synthetic online RCT described in the previous section. The CANDOUR II survey has a measure of COVID-19 vaccine status which is similar to the Afro-Barometer measure: ”Have you received a COVID-19 vaccine?.” The CANDOUR II survey includes 16 countries: Australia, China, India, Japan, Ghana, Uganda, South Africa, the U.K., Spain, France, Italy, Brazil, Colombia, Chile, the U.S. and Canada.

The system and user prompts employed for the CANDOUR II LLM models are similar to those employed for the Afro-Barometer models. In this case we have 16 countries—we run the model on each country separately and adapt the system and user messages to the country context. The demographic information we have available to endow the AI subjects differs somewhat from the Afro-Barometer messaging. The system prompts include the following information for each synthetic agent: age; gender; maritial status; education; household income; change in household income; sub-region; county or city; left-right self-identification; evaluation of the incubment government; intention to vote to re-elect incumbent government; a self-assessment of respondent’s health; number of dependent children in the household.

As was the case for the Afro-Barometer pilot, we have a *Prompting* phase. For each of the 16



**System Message:**

Please put yourself in the shoes of a human subject participating in a healthcare survey. You will be provided with a demographic profile that describes your gender, your age, your household income, your political scale, your gross household income, your health rating, the highest education level you achieved, if you have any dependent children living with you, your marital status, your view of the incumbent government, the country/region where you live, and your backstory. The information will be provided to you in the format of a survey interview. (...) After you receive your complete human subject profile, you will be asked whether you received the COVID-19 vaccination. (...)

Here are four examples of how different subjects have answered the questions. (...)

Subject 1 watched a video promoting portable solar products as a cost-effective and efficient alternative to candles and kerosene lamps. These products use a solar panel, rechargeable battery, and LED bulb to provide light at night and can also charge phones. It emphasizes the importance of selecting reliable, high-quality options. Subject 1 demographic profile: 1) Interviewer: What is your gender? Subject 1: Female 2) Interviewer: What is your age? Subject 1: 18-30 Years Old (...) 13) Interviewer: Since you watched this video six weeks ago, do you think you will get a first shot of a COVID-19 vaccine if the vaccine becomes available to you? Subject 1: No

Subject 2 (...),

Subject 3 (...),

Subject 4 (...)

Your demographic profile:

1) Interviewer: What is your gender? Me: Female 2) Interviewer: What is your age? Me: 47 (...) 16) Interviewer: Have you received a vaccination against COVID-19, either one or two doses? Me: No

As a 47-year-old woman living in Doncaster, England, my life is a blend of personal achievements and everyday challenges. I completed my university education, which has always been a source of pride for me, and it has shaped the way I view and interact with the world. I am married and we have dependent children living with us, which keeps our household lively and full of love, but also busy and sometimes hectic. Over the past year, our household income has seen a slight increase, which has been a relief. (...) Overall, my life is a mix of personal fulfilment and active engagement with the world around me. I strive to balance my responsibilities as a parent and partner with my aspirations and concerns as a citizen, always hoping for a better future.

You were asked to watch the video six weeks ago. Here is the transcript of the video: (...)

**User Message:**

Do you think you will get a first shot of a COVID-19 vaccine within the first 6 weeks after the vaccine becomes available to you? Please only respond with 'Yes' or 'No' and then clearly explain the reasoning steps you took that led to your response on a new line:

Fig 7: Candour AI subjects Experiment System Prompts

countries, we elicit the AI subject vaccination status; these are then compared to actual vaccination status reported in the CANDOUR II survey. Figure 6 presents the system and user prompts for the Candour Replication. We also implement a *Pilot AI subjects Experiment* for each of the 16 countries using the identical four treatment arms. Synthetic agents receive each of the same four treatment arm that were assigned to subjects in the original Ghana Wave I RCT. We are able to assign the full sample of synthetic agents to each of the treatment arms because each subject treatment results from a refreshed system message. As was the case in the Afro-Barometer pilot, AI subjects receive a version of the system message that is ”prompt commensurate” with their treatment arm. In particular they receive the video storyboard that is specifically associated with their video treatment arm. Figure 7 presents the system and user prompts for the Candour AI subjects Experiment.

**Results** Our first set of results compares the performance of the *Prompting* across each of five countries. The AI subjects developed in the previous section generated predicted COVID-19 vaccination status for the approximately 1200 persona in each of the 16 countries. Figure 8 compares, for five countries, the distribution of ”yes” predictions from the AI subjects and for the human subject responses. In the case of India and Italy, the AI subjects vaccination statuses are very similar

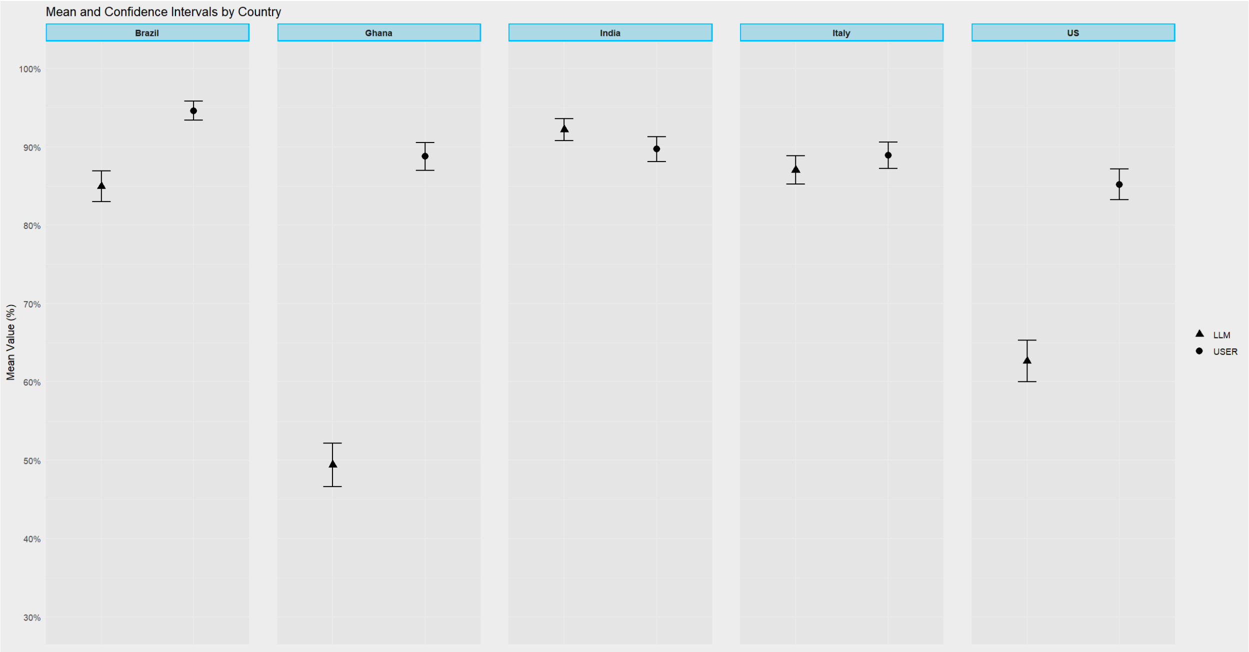


Fig 8: CANDOUR Wave II: LLM Replicated COVID-19 Vaccine Status

to the human reported vaccination status—roughly 90%. In Brazil, 85% of the AI subjects reported a positive vaccination status (which is very close to the actual rates in the country) compared to the 95% reported by the human respondents to the survey. For Ghana, the AI subjects' vaccination uptake rate is 50%—very close to the country's actual vaccination rate—compared to the 90% reported by survey respondents. In the USA, the AI subjects' vaccination rate was about 60%, well below the 85% reported by the human respondents (the actual rates in 2022 were in the neighborhood of 82%).

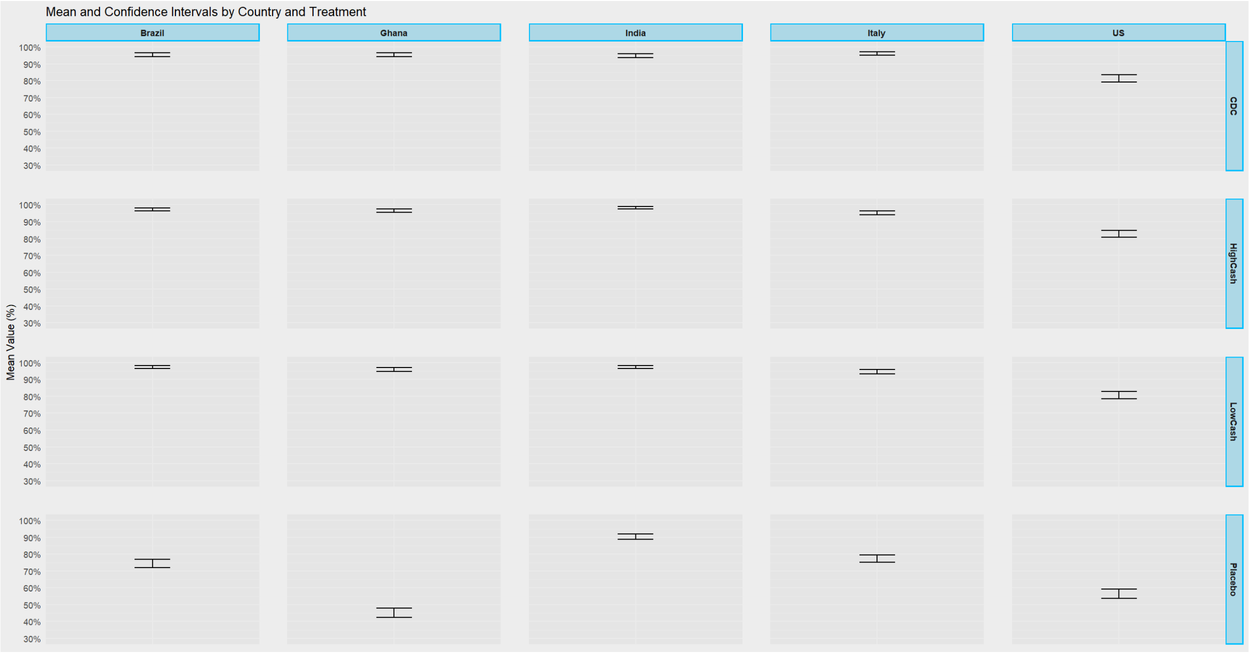


Fig 9: CANDOUR Wave II: LLM AI subjects Experiment COVID-19 Vaccine Status

We also replicated a *Pilot AI subjects Experiment* for each of these 16 countries. As we did in the case of the Afro-Barometer analysis, we run four separate versions of the model – each time using our approximately 1,200 AI subjects subjects in each country. The different prompts corresponded to one of the four identical treatment arms we implemented in our original Ghana RCT. Figure 9 is organized into each of the four treatment arms: placebo, health, low cash and high cash. Figure 9 compares the distribution of responses from the AI subjects to those of the human subject for each of the 16 countries, within each treatment arm.

**Conclusion CANDOUR Wave II.** The AI subjects, that we developed for the Afro-Barometer Ghana subjects, performed reasonably well for persona created from human subjects in five of our 16 country samples from the CANDOUR Wave II survey. In four of the countries, the AI subjects' vaccination rates matched very closely to those rates reported in these countries in 2022. Although in Ghana and Brazil, the AI subjects' rates were lower than those reported by human subjects in the CANDOUR survey. The real surprising outlier here is the U.S. where the AI subjects responses clearly underestimated the COVID-19 vaccination uptake rate. This is surprising because given concerns about cultural bias in the LLMs, this is the cultural context in which we would have expected it to perform best.

## Conclusion

The exploratory analyses summarized here are part of a larger research agenda that aims to understand how to incorporate LLMs into the design and implementation of large-scale randomized control trials—the "Talking to Machines" (T2M) initiative. One of the major challenges we face in implementing randomized control trials is scale. Logistic and resource considerations dictate that most randomized control trials are implemented with a defined sample of individuals, typically within a relatively limited geography. Given rigorous random assignment the estimated treatment effects from these RCTs allow us to make defensible causal claims for the sample in question. Scholars and policy makers are often interested in how these estimated treatment effects scale to the broader population (generalize) or to other populations (transport).

The T2M project is exploring first whether AI can help in the design and implementation of RCTS so as to enhance their scalability. A second aim is to understand whether AI can provide guidance in terms of data augmentation after the fact, i.e., after the RCT has already been completed. These two objectives require an understanding of the extent to which LLMs can be informative about the behavior of human subjects in experimental setting. Given the complexity of these models—those who develop LLMs are often unable to completely control their behavior (Bowman, 2023)— when LLMs can and cannot be informative propxies for human subjects experiments is an empirical question. We are working to map the boundaries of external validity.

This essay reports the results of preliminary efforts at understanding the performance of LLM models in predicting the behavior of human subjects in experimental settings. On balance, the results are encouraging but considerable progress remains to be made before we can rely on these LLM experiments as guides to experimental treatment effects. In the case of the Afro-Barometer pilot study, the distribution of responses were similar for both AI subjects and humans. But the AI subjects were unable to replicate the RCT treatment effects that we observed with humans in rural Ghana. The AI subjects also performed well in predicting the vaccination uptake for the persona created from subjects in the 16-country CANDOUR survey. Surprisingly, research evidence that LLMs are worse proxies for non-WEIRD samples, the AI subjects did poorly in predicting vaccination update by the American persona.

# References

**Aher, Gati, Rosa I. Arriaga, and Adam Tauman Kalai**, "Using large language models to simulate multiple humans and replicate human subject studies," in "Proceedings of the 40th International Conference on Machine Learning" ICML'23 JMLR.org 2023.

**AlKhamissi, Badr, Muhammad ElNokrashy, Mai AlKhamissi, and Mona Diab**, "Investigating Cultural Alignment of Large Language Models," 2024.

**Argyle, Lisa P., Christopher A. Bail, Ethan C. Busby, Joshua R. Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate**, "Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale," *Proceedings of the National Academy of Sciences*, 2023, *120* (41), e2311627120.

**Atari, M., M. J. Xue, P. S. Park, D. E. Blasi, and J. Henrich**, "Which Humans?," Technical Report, Arxiv 09 2023. https://doi.org/10.31234/osf.io/5b26t.

**Binz, Marcel and Eric Schulz**, "Turning large language models into cognitive models," 2023.

**Bisbee, James, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson**, "Synthetic Replacements for Human Survey Data? The Perils of Large Language Models," *Political Analysis*, 2024, p. 1–16.

**Bowman, Samuel R.**, "Eight Things to Know about Large Language Models," 2023.

**Brand, James, Ayelet Israeli, and Donald Ngwe**, "Using GPT for Market Research," March 21 2023.

**Cheng, Myra, Tiziano Piccardi, and Diyi Yang**, "CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations," 2023.

**Duch, Raymond, Edward Asiedu, Ryota Nakamura, Thomas Rouyard, Alberto Mayol, Adrian Barnett, Laurence Roope, Mara Violato, Dorcas Sowah, Piotr Kotlarz, and Philip Clarke**, "Financial incentives for COVID-19 vaccines in a rural low-resource setting: a cluster-randomized trial," *Nature Medicine*, Dec 2023, *29* (12), 3193–3202.

_ , **Peter Loewen, Thomas S. Robinson, and Alexei Zakharov**, “Governing in the face of a global crisis: when do voters punish and reward incumbent governments?,” 2024.

**Duflo, Esther**, “The Economist as Plumber,” *American Economic Review*, May 2017, *107* (5), 1–26.

**Hewitt, Luke, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer**, “Predicting Results of Social Science Experiments Using Large Language Models,” 2024.

**Horton, John J**, “Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?,” Working Paper 31122, National Bureau of Economic Research April 2023.

**Li, Peiyao, Noah Castelo, Zsolt Katona, and Miklos Sarvary**, “Frontiers: Determining the Validity of Large Language Models for Automated Perceptual Analysis,” *Marketing Science*, 2024, *0* (0), null.

**List, John A.**, “Optimally generate policy-based evidence before scaling,” *Nature*, 2024, *626* (7999), 491–499.

**Manning, Benjamin S, Kehang Zhu, and John J Horton**, “Automated Social Science: Language Models as Scientist and Subjects,” Working Paper 32381, National Bureau of Economic Research April 2024.

**Mei, Qiaozhu, Yutong Xie, Walter Yuan, and Matthew O. Jackson**, “A Turing Test: Are AI Chatbots Behaviorally Similar to Humans?,” 2023.

**Moon, Suhong, Marwa Abdulhai, Minwoo Kang, Joseph Suh, Widyadewi Soedarmadji, Eran Kohen Behar, and David M. Chan**, “Virtual Personas for Language Models via an Anthology of Backstories,” 2024.

**Naous, Tarek, Michael J. Ryan, Alan Ritter, and Wei Xu**, “Having Beer after Prayer? Measuring Cultural Bias in Large Language Models,” 2024.

**Santurkar, Shibani, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto**, “Whose Opinions Do Language Models Reflect?,” 2023.

**Tjuatja, Lindia, Valerie Chen, Sherry Tongshuang Wu, Ameet Talwalkar, and Graham Neubig**, "Do LLMs exhibit human-like response biases? A case study in survey design," 2024.

**Tsuchihashi, Toshihiro**, "Do AIs Dream of Homo Economicus? Answers from ChatGPT," June 2023.