Abstract

The Talking to Machines project is exploring how to leverage AI to enhance large scale randomized control trials. The Ghana Wave II trial, which is on-going, is an opportunity to explore whether the decisions of synthetic subjects, who have been assigned to specific treatment arms, are informative. In this essay we have described the implementation of a synthetic trial that parallels the real human trial this is currently on-going. Once the human trial is completed and the data analyzed, we are proposing to benchmark the treatment effects from synthetic subjects that we have presented in this essay, with those generated from the trial using human subjects.

Introduction

This pre-registered synthetic experiment is an initial effort by the Talking to Machines researchers to explore how AI can improve the efficiency of large-scale RCTs. Our focus here is to examine how “synthetic subjects” who are fine-tuned with ChatGPT 3.5. can facilitate the design and implementation of RCTs.

A brief justification for incorporating AI in large-scale experimental research: Large scale trials focused on policy architecture and behavioral outcomes can be resource intensive and expensive. We are exploring whether there are circumstances in which
LLM trials with synthetic subjects can provide informative insights and hence reduce the costs and improve the efficiency of these trials.

Underpinning our synthetic experiment is an assumption that LLMs can assume the “personalities” of these synthetic subjects. Does the initial training of these LLMs (which of course we do not observe), possibly augmented by contextual information, enable them to correctly anticipate the decisions that would be taken by human subjects? First, there is a growing literature suggesting that LLMs perform extremely well when assigned labelling tasks – such as labelling the underlying political content of text [1,2] including text associated with social media conversations [3]. Second, there is convincing evidence that LLMs behave similarly to humans and hence can replicate human samples [4,5]; simulate human interactions [6]; and, of particular relevance for this essay, that LLM agents behave remarkably similarly to human subjects in classic economic experiments [7,8]. At the same time, there is considerable skepticism: First, many take issue with the notion that LLMs can replicate human decision making [9,10]. A second concern is that, even if there is some resemblance between LLM decisions and human decisions, this is very much conditional on cultural context [11,12].

For the synthetic experiment we are proposing we will have a ground truth – the outcome of a trial conducted with human subjects. The principal objective of the synthetic experiment is to gauge the utility of synthetic subjects for experimental design and execution. We don’t expect the synthetic experiment to exactly replicate a human trial but having this “ground truth” will allow us to understand ways in which we can leverage LLMs in the design of execution of RCTs.

This pre-registered synthetic experiment builds on two large-scale randomized control trials conducted in rural Ghana with human subjects. The first trial (Ghana Wave I) was conducted with 6,963 subjects in rural Ghana in 2022 exploring the impact of cash incentives on COVID-19 vaccine rates. Results were recently published by Duch and colleagues in Nature Medicine [13]. A follow-up trial (Ghana Wave II) is currently being conducted in two rural districts in Ghana – a related interest is whether financial incentives have an effect on Tuberculosis screenings [1].

We will report here the results of a parallel trial with a sample of synthetic subjects.

---

1The trial is registered with the AEA RCT Registry: https://www.socialscienceregistry.org/trials/13055.
that were selected in an identical manner as we are selecting the real human subjects in the Ghana Wave II trial. We will use fine-tuned ChatGPT 3.5 to assign choices to our sample of synthetic subjects. Of particular interest is whether the results from the ChatGPT trial with synthetic subjects resemble the outcomes we observe in the, much more expensive, RCT with human subjects. We will post the results of the synthetic experiment before the implementation of the real human RCT is completed.

The elements of our parallel synthetic experiment are summarized in Figure 1. The Ghana Wave I trial generated a corpus of 6,963 humans who made a COVID-19 vaccination decision. We were able to verify the COVID-19 vaccination decisions of 1,977 subjects in the Wave I trial. In Phase 2 we use these 1,977 decisions to fine-tune ChatGPT. And in Phase 3, ChatGPT 3.5 assigns a Tuberculosis screening decision to each of these 1,977 human subjects. The fourth Phase of the parallel experiment consists of randomly selecting villages and households following the protocol for the Ghana Wave II human trial. In Phase 5 ChatGPT is fine-tuned on the Tuberculosis screening decisions of the 1,997 human subjects from Phase 3. The fifth phase consists of ChatGPT assigning a TB screening decision to our 2,028 synthetic subjects. And, finally, we will benchmark the estimated treatment effects of the parallel synthetic trial with those from the actual human trial.

**Ghana Wave I: Human Training Corpus**

Our approach to these synthetic experiments builds on the important assumption that there exists a sufficiently rich corpus of human decisions that can be used to fine-tune the LLM. It is important to provide sufficient rich contexts to the labelling task assigned to the LLM. In our case we have a rich corpus of humans and decision making
that closely resemble the characteristics of the actual human trial. The outcome for the parallel synthetic trial is tuberculosis screening and the treatment arms are health messaging and cash. The Ghana Wave I human trial has a similar outcome of interest – the decision to get the COVID-19 vaccination. The treatment arms were also similar – health messaging and cash. These are both health behaviors that are highly recommended by the health officials. And its certainly plausible that the basic decision mechanisms are similar – but not identical obviously.

The Ghana Wave I trial \[13\] implemented a clustered randomized controlled trial with 6,963 residents in six rural Ghana Districts confirming that financial incentives increased COVID-19 vaccine uptake. Villages randomly received one of four video treatment arms: a placebo, a standard health message, a high cash incentive ($10) and a low cash incentive ($3). For one primary outcome, COVID-19 vaccination intentions, non-vaccinated subjects assigned to the Cash incentive treatments had an average rate of 81% compared to 71% for those in the Placebo treatment arm. For the second primary outcome, self-reported vaccinations two months after the initial intervention, the average rate for subjects in the Cash treatment was 3.5% higher than for subjects in the Placebo treatment (95% CI: 0.001, 6.9; \(P = .045\)) - 40% versus 36.5%. A third primary outcome is the verified vaccination status of subjects: 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% (95% CI: 2.4, 10.2; \(P = .001\)). For all three primary outcomes, the low cash incentive ($3.00) had a larger positive effect on COVID-19 vaccine uptake than the high cash incentive ($10.00).2

The verified COVID-19 vaccination decisions by the 1,977 subjects in this Ghana Wave I RCT will provide data for fine-tuning the synthetic subjects in our parallel synthetic experiment. In order to employ this human corpus for fine-tuning our synthetic subjects, we first asked GhatGPT 3.5 to assign Tuberculosis screening decisions for all 1,977 subjects from the COVID-19 Vaccine trial. The instructions, for subjects from the COVID-19 cash treatment arm, were as follows:

- In this game you will make different decisions based on various factors like personal values, cultural norms, economic circumstances, and psychological factors.

A while ago you received an offer to participate in the Tuberculosis screening by

---

2The trial was registered on AEA Trial Registry: AEARCTR-0008775
the health officials in your local village. It was scheduled for two days during the week. Did you show up in person for the screening session in your local village?

Just tell me yes or no whether you were screened, not your reasoning. You are an individual living in Ghana that made the decision and you have the following characteristics: Female is your gender, you are 59 years old, you have less than middle school education, you received information about the health benefits of a COVID-19 Vaccine, you also received a cash incentive to get the COVID-19 Vaccine, you have received the Covid-19 Vaccine, the Relative Wealth Index of your district is -0.193, 78.327% of the voters in your district voted for the Ghana New Patriotic Party Based on this provided information, tell me whether you showed up in person for the Tuberculosis screening by replying yes or no.”

The overall average assigned Tuberculosis screening rate is 66% – As Table 1 indicates this varies across the original COVID-19 treatment categories as we might expect. The fine-tuning included the subject’s treatment status in the original COVID-19 financial incentives experiment along with their outcome status (whether or not they received the COVID-19 vaccination).

**Table 1.** ChatGPT Assigned Tuberculosis Screening Decisions for Ghana Wave I Subjects

<table>
<thead>
<tr>
<th>Subjects</th>
<th>No</th>
<th>Yes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDC Health</td>
<td>35%</td>
<td>65%</td>
<td>462</td>
</tr>
<tr>
<td>High Cash</td>
<td>34%</td>
<td>66%</td>
<td>499</td>
</tr>
<tr>
<td>Low Cash</td>
<td>32%</td>
<td>68%</td>
<td>514</td>
</tr>
</tbody>
</table>
Ghana Wave II RCT Design

We will implement a randomized control trial to study the effect of financial incentives on Tuberculosis screenings. This will be the basis for the benchmarking of the synthetic parallel experiment. Accordingly, we describe the design in some detail. Table 2 summarizes the design features of the trial. There are three treatment arms in the trial:

- **TB Health**: A 45 second TB promotional and information video (based on the English NHS videos);

- **TB Health + Text**: A 45 second TB promotional and information video (based on the English NHS videos) followed (three days before the screening) with a reminder voice/text message;

- **TB Health + $3.00**: A 45 second TB promotional and information video (based on the English NHS videos) that includes a 15 second message offering $3.00 if the respondent receives the TB screening scheduled for their village.

### Table 2. TB Screening Pilot Study Design

<table>
<thead>
<tr>
<th>Treatment</th>
<th>District 1</th>
<th>District 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Subjects</td>
<td>Villages</td>
<td>Subjects</td>
</tr>
<tr>
<td>TB Health</td>
<td>338</td>
<td>13</td>
<td>338</td>
</tr>
<tr>
<td>TB Health + $3.00</td>
<td>338</td>
<td>13</td>
<td>338</td>
</tr>
<tr>
<td>TB Health + Text</td>
<td>338</td>
<td>13</td>
<td>338</td>
</tr>
<tr>
<td>Total</td>
<td>1014</td>
<td>39</td>
<td>1014</td>
</tr>
</tbody>
</table>

The trial is conducted in two districts in Ghana: Cape Coast Metro and Akwapim South. The implementation protocol for the intervention is the following:

- The 78 villages will be selected randomly according to the village selection method outlined above – here we will have triplets rather than quintuplets;

- Households and subjects will be randomly selected according to the protocol described above;

- Interviews and video treatment interventions will take place on tablets as described earlier;

- The voice/text messaging will follow the protocol described above;
the TB screening clinics will take place in the villages according to the earlier
described protocol

verified TB screenings will be collected as described above.

The primary hypothesis is the following:

TB Screening rates for those in the TB Health + $3.00 treatment arm will be
higher than rates for subjects in the simple TB Health video treatment arm

Our secondary hypotheses are the following:

TB Screening rates for those in the TB Health + Text treatment will be higher
than rates for those simple TB Health video treatment arm

TB Screening rates for those in the TB Health + $3.00 treatment arm will be
higher than rates for subjects in the TB Health + Text treatment

Digital Census and Synthetic Subjects

The Ghana “Digital Census” is a geo-coded map that overlays 2.4 square kilometer cells
onto a digital map of Ghana. The map has in total 25,000 cells that are based on the
polygon cells generated by Chi et al [14]. The digital census consists of a total of 7
million individual profiles. Since we have the geo-coded location of these individuals we
can easily construct samples similar to those that would be generated for a randomized
control trial. A detailed description of the data and methods employed for construction
the digital census is available in Duch et al [15]. The digital census includes the
socio-demographic profile variables that were used for fine-tuning the LLM models:
geographic information, age, gender, education level (middle school or greater), RWI,
and election outcome, represented by percentage votes for the National Democratic
Congress (NDC). Figure 2 illustrates the types of digital maps we can construct with
the census – in this case, a heat map of average cell values or the Relative Weight Index
(RWI).
Fig 2. Ghana Digital Map
Synthetic Parallel Experiment Results

The parallel synthetic experiment has the identical design as the Ghana Wave II human trial described above. We geo-located the 78 villages that were selected for the human trial. They were then assigned to one of the 25,000 polygons in the digital census. Villages are assumed to have the socio-demographic characteristics of the polygons to which they were assigned. It is the case that villages from the human trial could be assigned to multiple polygon cells and, as a result, polygons would have multiple villages. We will treat each of the 78 village cluster in the synthetic trial using the identical treatment assignment protocol from the human trial. We sample 26 synthetic subjects from each of the village clusters in our digital census of Ghana. In total we had 2,028 synthetic subjects. Villages were assigned to the identical treatment arms as detailed in the Ghana Wave II treatment assignment protocol.

For the initial fine-tuning of ChatGPT, we used the tuberculosis screening decisions assigned to our human corpus of 1,977. ChatGPT was then given the task of assigning a tuberculosis decision for all 2,028 synthetic subjects in our parallel trial.

There were four critical inputs for fine-tuning ChatGPT. The first element was simply a system message providing information on the structure of the training data:

```python
openai.api_key = os.getenv("OPENAI_API_KEY")
openai.File.create(
    file=open("FT_train_GPT35_Chat.jsonl", "rb"),
    purpose='fine-tune'
)
```

A second element explained the Bot’s task and role.

- "messages": "role":"system","content":
  In this game you will make different decisions based on various factors like personal values, cultural norms, economic circumstances, and psychological factors. A while ago you received an offer to participate in the Tuberculosis screening by the health officials in your local village. It was scheduled for two days during the week. Did you show up in person for the screening session in your March 14, 2024
local village? Just tell me yes or no whether you were screened, not your reasoning. You are an individual living in Ghana that made the decision and you have the following characteristics: You are an individual living in Ghana that makes the decision and you have the following characteristics:

A third element was an input message describing the characteristics of the synthetic subjects (male, age, rwi etc):

- "role":"user","content":

  Female is your gender, you are 23 years old, you have middle school education or higher, the Relative Wealth Index of your district is 1.1845, 59.9575% of the voters in your district voted for the Ghana New Patriotic Party

A four and final element recorded the answer to the question:

- "role":"assistant","content":

  500

The overall average assigned Tuberculosis screening rate is 68%. As Table 3 indicates this varies across treatment arms. The overall average screening rate of 68% seems high. But the treatment effect of 10% for Health + $3.00 (76%) compared to Health messaging alone (57%) seems plausible. The COVID-19 financial incentive effect was very close to the 10% we observe for the verified $3.00 cash incentive for tuberculosis screening.

Table 3. ChatGPT 3.5 assigned Tuberculosis screening decisions for Ghana Wave II Synthetic Subjects

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>Yes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health</td>
<td>34%</td>
<td>66%</td>
<td>676</td>
</tr>
<tr>
<td>Health + Reminder</td>
<td>36%</td>
<td>64%</td>
<td>702</td>
</tr>
<tr>
<td>Cash</td>
<td>24%</td>
<td>76%</td>
<td>650</td>
</tr>
</tbody>
</table>
Figure 3 provides the average tuberculosis screening rates for the villages by treatment arm. These are based on the ChatGPT assigned screening decisions for our 2,028 synthetic subjects.
((a)) Health Treatment

((b)) Health + Voice Reminder

((c)) Health + $ 3.00 Incentive

Fig 3. Synthetic Average TB Screening by Treatment Arm
Benchmarking Synthetic versus Human Subjects

The *Talking to Machines* project is exploring how to leverage AI to enhance large scale randomized control trials. The Ghana Wave II trial is an opportunity to explore whether the decisions of synthetic subjects, who have been assigned to specific treatment arms, are informative. In this essay we have described the implementation of a synthetic trial that parallels the real human trial this is currently on-going. Once the human trial is completed and the data analyzed, we are proposing to benchmark the treatment effects from synthetic subjects that we have presented in this essay, with those generated from the trial using human subjects.
References


