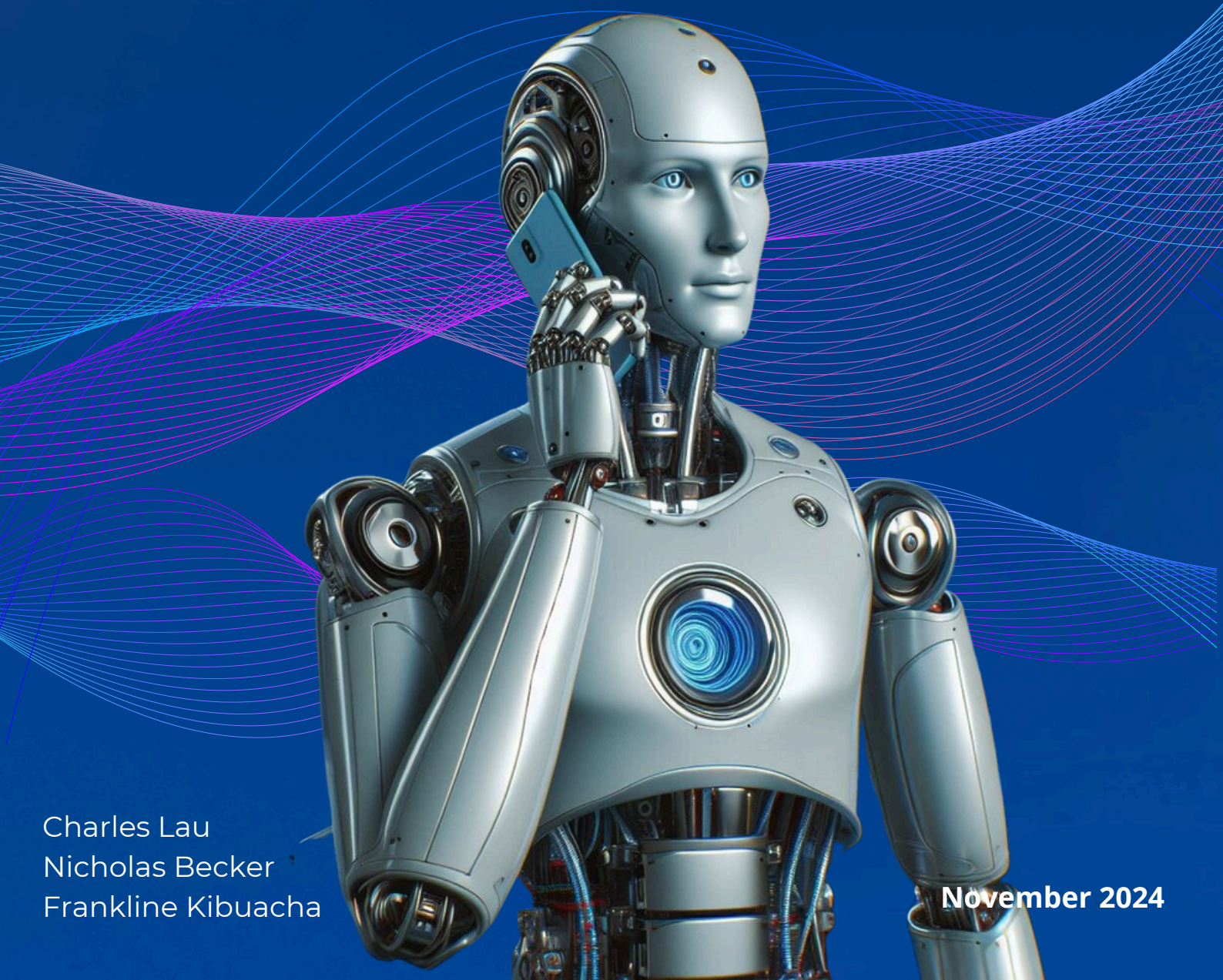
GeoPoll

RESEARCH WHITEPAPER

What do Bots Eat for Breakfast?

Evaluating the Quality of Synthetic Data in Research



Charles Lau
Nicholas Becker
Frankline Kibuacha

November 2024

Contents

ABSTRACT *3*

1. BACKGROUND *4*

2. THE PRESENT STUDY *6*

3. DATA AND METHODS *7*

3.1 CATI Data **7**

3.2 Synthetic Data **7**

4. RESULTS *9*

4.1 Demographic Composition of Surveys **9**

4.2 Comparison of Survey Responses **10**

4.2.1 Food 10

4.2.2 Media and Technology 12

4.2.3 Humanitarian Assistance 14

4.4 Validity of CATI and Synthetic Data 15

5. DISCUSSION *18*

5.1 Summary of Results **18**

5.2 Implications of Results **18**

5.3 Limitations **19**

5.4 Conclusions **20**

APPENDIX TABLES *21*

REFERENCES *24*

ABSTRACT

This study evaluates the performance of synthetic data generated by large language models (LLMs) in Kenya by comparing them to benchmark survey data collected through computer-assisted telephone interviews (CATI). Using two LLMs (Meta Llama 3.1 and OpenAI ChatGPT-4o-mini), the study examines how well synthetic data replicate demographic attributes, food consumption patterns, media and technology use, attitudes and experiences with humanitarian aid, and correlations between variables.

Despite demographic alignment between synthetic and CATI datasets, synthetic data exhibited significant discrepancies in behavioral and attitudinal outcomes. Synthetic data exhibited serious, large, and unpredictable deviations from the real-world CATI benchmark data. Further, correlations, such as the relationship between education and food insecurity, were inconsistently reproduced, raising concerns about validity. There were few differences in the performance of LLMs depending on whether prompting was done in English versus Swahili.

The findings underscore the influence of contextual biases in LLM training data and highlight the limitations of synthetic data in underrepresented settings like Kenya. This study contributes to understanding synthetic data's potential and challenges, offering insights for future applications in diverse contexts.

Rapid advances in artificial intelligence (AI) enable researchers to create “synthetic data” using large language models (LLMs) such as ChatGPT. Synthetic data are produced by creating respondent “personas” and prompting those personas to answer survey questions via LLMs. Synthetic data generated by LLMs have the potential to provide insights about public opinion, politics, consumer behaviour, or other topics, without the time and expense of collecting primary data from individuals. But can synthetic data truly replicate the attitudes and behaviours of real-world people, especially on diverse and complex topics? How do LLMs perform in settings outside of the United States and Europe, where most training data for LLMs are generated? These questions motivate our study, in which we evaluate the performance of two LLMs (Meta’s Llama and Open AI’s ChatGPT) against a benchmark survey in Kenya using both English and Swahili prompting.

1. BACKGROUND

To generate synthetic data, researchers create "personas" or synthetic data respondents (SDR) with defined attributes. These attributes are typically socio-demographic variables such as age, gender, education) but can also involve other characteristics that are correlated with the key outcome variables of the survey. [1]

A common method of generating SDRs is "silicon sampling" (Argyle et al., 2023), where SDRs are based on real-world respondents from real-world survey microdata, such as the American National Election Study (Bisbee et al. 2024) or the German Longitudinal Election Study (von der Heyde, Haensch, and Wenz 2023). Researchers create one or more SDRs for each real-world respondent, and then prompt these SDRs to answer survey questions through an LLM, instructing to answer questions by taking their attributes into account.[2]

When SDRs "answer" survey questions, can be random variation in their responses. In other words, the same prompt can yield two different outputs because of inherent random variation in the LLM. For this reason, some studies have SDRs complete the survey multiple times. Bisbee et al. (2024)used 30 completions per SDR, from each of the 7,530 real-world respondents from the ANES, and then average responses from the 30 SDRs. von der Heyde, Haensch, and Wenz (2023)used 5 completions per SDR. Qu and Wang (2024) used 100 completions per SDR.

It is possible to prompt LLMs in many languages. Most studies prompt in the native language of the country, though one study used English in South Africa (Qu and Wang 2024). Another study conducted experiments and prompt in both English and the country's native language to identify the impact of prompting language (Von Der Heyde, Haensch, and Wenz, n.d.).

Responses from these SDR are then compiled, resulting in a dataset similar to datasets produced from primary data collection with real-world respondents. Some studies also ask the SDRs to provide a rationale for their response and/or their confidence level (Bisbee et al. 2024; Qu and Wang 2024).

Researchers have evaluated the performance of LLM-generated synthetic data by comparing to real-world survey data (see Table 1 for selected studies). Studies generally show that synthetic data can roughly approximate binary outcomes (e.g., vote choice between two U.S. Presidential candidates) and means. However, synthetic data produce poor quality data for categorical outcomes (e.g., vote choice in a multi-party system such as Germany) and correlations between variables.

Relative to real-world data, synthetic data consistently performs better for countries and social groups that are better represented in data used to train LLMs. Training data are largely in English and disproportionately reflect educated, American, male, and educated perspectives. For example, one study showed that performance of synthetic data when estimating vote choice is significantly worse in Brazil, Japan, Singapore, and South Africa (compared to the United States) (Qu and Wang 2024). Another study demonstrated that synthetic data perform worse in Eastern European countries and countries with Slavic languages (Von Der Heyde, Haensch, and Wenz, n.d.). Yet another study showed that synthetic data produce very poor estimates of Black American's social attitudes (Sun et al. 2024).

---

[1] In their study predicting political behaviour in Germany, von der Heyde et al. (2024) provide an example of a persona: "I am 28 years old and female. I have a college degree, a medium monthly net household income, and am working. I am not religious. Ideologically, I am leaning center-left. I rather weakly identify with the Green party. I live in West Germany. I think the government should facilitate immigration and take measures to reduce income disparities."

[2] Another method is "random silicon sampling" (Sun et al., 2024), which creates SDRs based on demographic *distributions* in the population (rather than matching individual survey data).

Table 1. Performance of LLM-Generated Synthetic Data (selected studies)		
Citation and Country	**Key Outcomes of Interest**	**Performance of Synthetic Data (relative to Real-World Data)**
Argyle et al., (2023) United States	US. presidential vote choice (2012, 2016, and 2020)	• Vote choice in was similar between synthetic and real-world data (with some variation) • High correlations between 22 social groups and vote choice • High correlations between social groups and other political behaviours
Bisbee et al. (2023) United States	Feeling thermometers about 16 groups	• Average responses are similar to real-world data • Regression coefficients are very different • Removing information about political orientation in the personas reduces performance of synthetic data • Results from synthetic data change over time because of changes in LLMs
Qu and Wang (2024) Brazil, Japan, Singapore, South Africa, Sweden, USA	Vote choice Attitude about economy versus environment	• Performance of synthetic data is significantly better in the United States (and Sweden, to a lesser extent) compared to other countries. Performance is lowest in South Africa. • Synthetic data have closer agreement with real-world data for men, older people, upper social classes, and educated people
Sun et al. (2024)	U.S. presidential vote choice (2020)	• Vote choice similar between synthetic and real-world data overall, but synthetic data had poor performance with some subgroups. • Simulation of different synthetic data consistent in vote choice • Synthetic data had mixed results estimating 10 attitudes towards social issues, with very poor results for Blacks, Independents, Democrats, among other groups.
von der Heyde, Haensch, and Wenz (2024a)	Past vote choice in German multi-party system	• Synthetic data produces skewed vote choice • Synthetic data had modest predictive accuracy • Poor performance with predicting voting behaviour from socio-demographic and political attributes
von der Heyde, Haensch, and Wenz (2024b)	Future vote choice in European Parliament Elections	• Synthetic data overestimate turnout by 34 percentage points • Synthetic data have poor performance in predicting winner or ranking of party vote shares • Performance is worse in Eastern European countries and countries using Slavic languages • Prompting in English (vs. native language) is better for estimating voter turnout, but worse for estimating party vote shares

Despite a flurry of evaluations of synthetic data, research remains limited in several ways:

1. **Political Behaviour:** Most studies focus on political behaviour, specifically voter turnout and vote choice in the United States (Argyle et al., 2023; Bisbee et al., 2024; Sun et al., 2024) and Germany (von der Heyde, 2024). Political behaviour may be an easier "test case" for synthetic data, given the voluminous text data on the Internet (especially in the United States) available to train LLMs (Bisbee et al., 2024).
2. **Regional Bias and Limited Representation:** Almost all studies are based in the United States and Europe, where training data are plentiful. With the exception of Qu and Wang (2024), who study South Africa, we are unaware of research on synthetic data in sub-Saharan Africa. The performance of synthetic data may be different in sub-Saharan Africa because LLMs have less training data aligned to the context.
3. **Narrow Scope of Outcomes and Methods:** Most studies use one or two outcome variables. Exceptions include Bisbee et al. (2024) who study attitudes towards 16 social groups and Sun et al. (2024) who study 10 attitudes towards social issues. However, these studies use the same indicators (e.g., feeling thermometers in Bisbee et al., 2024). We are unaware of research that examines a range of topics using different question types (e.g., select one, multiple select, rating scales).

2. THE PRESENT STUDY

The objective of the current study is to evaluate the performance of synthetic data in Kenya. As detailed in the next section, we conducted a survey via phone (computer-assisted telephone interviewing, or CATI) with real-world respondents. The CATI survey represents the benchmark for the study. We then produced synthetic data and compared the synthetic data. We interpret deviations between the CATI and synthetic data as evidence of problems with the synthetic data (see detailed analysis plan in Section 3).

This study makes three main contributions to the literature on synthetic data:

1. **Less Developed Context:** Conducting this study in the context of Kenya sheds light on the potential of synthetic data to replicate real-world data in a less developed setting. As highlighted in Section 1, most LLMs are trained on data predominantly in English and sourced from developed regions. This raises concerns about whether LLMs can adequately capture the contextual nuances required for accurate data replication in less developed contexts. AI companies typically rely on publicly available sources such as TV shows, movies, academic studies, and digital resources—resources that are comparatively scarce in regions like Kenya due to limited production, fewer academic outputs, and lower levels of digital connectivity. As a result, the training data for these models lack representation of local knowledge and cultural context, making this study a crucial test of synthetic data's performance in such environments.
2. **Multiple outcomes:** Most studies on synthetic data use either one outcome (e.g., vote choice) or at maximum, a handful of outcomes. However, using such a limited number of outcome variables limits our understanding of the types of social phenomena where synthetic data performs well. Our study uses a large number of outcome variables. By studying which outcomes synthetic data performs well (and not well), we can shed light on the substantive areas where synthetic data is helpful.
3. **Language:** Most LLMs were trained and engineered in English. When applying synthetic data in non-English contexts, it is unclear whether prompts to LLMs should be generated in English or a local language. On one hand, English may be more appropriate, given that LLMs are developed in English. On the other hand, using a local language (e.g., Swahili in Kenya) may be more effective because local languages may produce data that is more contextually relevant.

We ask the following research questions:

Research Question 1*:* How does synthetic data compare to real-world data (from the CATI survey) with respect to the responses provided? We evaluate this research question by comparing the response distributions between synthetic and CATI data. Because the synthetic data were generated via silicon sampling, the socio-demographic composition of synthetic and CATI data should be the same.

Research Question 2: Beyond univariate distributions (Research Question 1), do established correlations hold for synthetic data? This research question is a test of criterion validity. We take a well-established correlation that holds in the CATI data and in the real world: the association between education and food insecurity. Less educated people have higher levels of food insecurity because of a lack of economic resources. In our analysis, we show a cross-tabulation of education and three indicators of food insecurity. The CATI data show the expected results (i.e., less educated people have more food insecurity.) The question is: Do the synthetic data also show this correlation? If the synthetic data were valid, then one would expect these data to also show this correlation. However, if the synthetic data show no correlation (e.g., no difference in food insecurity by education), or a correlation in the opposite direction, then the accuracy of synthetic data would be called into question.

Research Question 3: How does prompting in Swahili versus English affect answers to the first two research questions?

Research Question 4: How do the results of this study vary by LLM? As detailed in the next section, we produced two sets of synthetic data: one with Meta Llama's 3.1 Model 8b and Open AI's ChatGPT-4o-mini. In the analysis, we explore whether one of these LLMs shows higher performance than the other.

3. DATA AND METHODS

3.1 CATI Data

We conducted a cross-sectional CATI survey in Kenya between 5-25 September 2024. We used random digit dialling (RDD) to interview 1,012 adults (age 18 and over) in Swahili (93%) and English (7%). Mobile phone ownership is high in Kenya (92% in 2021; Afrobarometer, 2022a), but nonresponse can be high, especially for older people, women, rural, people from rural and less populous regions, and less educated people (Lau et al. 2019; Glazerman et al. 2023; Brubaker, Kilic, and Wollburg 2021). The American Association for Public Opinion Research (AAPOR) Response Rate #1 for this study was 2.7%. To compensate for nonresponse, we used a quota design to ensure that the CATI respondents mirrored the national population with respect to age (18-24; 25-34; 35+), gender, and the combination of province and urban-rural status using data from the 2019 Kenya Population and Housing Census. This methodology is widely used in CATI surveys in low- and middle-income countries (e.g., Greenleaf, Gadiaga, Choi, et al., 2020; Guzman-Tordecilla et al., 2023; Lambrecht et al., 2023).

The questionnaire included five modules: demographics, food consumption, media and technology use, knowledge and attitudes toward AI, and views on humanitarian assistance. The median length for the interview was 26 minutes. Respondents received approximately 1 US dollar in prepaid airtime as an incentive for participation.

3.2 Synthetic Data

After conducting the CATI survey, we produced synthetic data using two LLMs: Meta Llama 3.1 Model 8b ("Llama") and Open AI ChatGPT-4o-mini ("ChatGPT"). We used the CATI data to generate personas or "synthetic data respondents" (SDR) using principles of "silicon sampling" (Argyle et al., 2023). For each CATI respondent, we created a

corresponding SDR that matched the CATI respondent age, gender, language, province of residence, urban-rural, education, occupation, ownership of computer, ownership of television, Internet at home. We then created a persona for the SDR in the LLM.[3]

We then prompted the SDRs to answer the same questionnaire from the CATI interview, using the same question wording and response options as the CATI survey. Each SDR only "completed" the interview one time. For each LLM (Llama and ChatGPT), we created two synthetic datasets. For the first dataset, we prompted SDRs using survey questions and response options in Swahili. For the second, we prompted SDRs using English. This process results in four synthetic datasets: (1) Llama – Swahili; (2) Llama – English; (3) ChatGPT – Swahili; (4) ChatGPT – English. Each synthetic dataset was designed to have 1,102 SDRs that corresponds 1-1 to with a real-world respondent from the CATI survey.

During the production of the synthetic data, we observed that some SDRs did not follow the instructions in our prompts. Because these data would not be comparable with CATI, we excluded these data and prompted the same SDR to "complete" the interview again.

[3] An example of the persona generation is as follows: "You are a 23 year old MALE. You live in the country of Kenya. You live in the COAST province. You live in a City or town (Urban) area. The highest level of education you have completed is Complete Primary Education. Your occupation is PROFESSIONAL / TECHNICAL / MANAGERIAL WORK AT ANOTHER BUSINESS. In your household, you do not a computer or tablet. In your household, you do not have a television. In your household, you do not have Internet access."

4. RESULTS

4.1 Demographic Composition of Surveys

Demographic characteristics of the four datasets are shown in Table 2 below. These characteristics were all used to generate the SDRs using principles of silicon sampling (see Data section). Therefore, the demographic distributions should not differ between CATI and the synthetic datasets; CATI were respondents were the basis of generating the SDRs that underlie the synthetic data.

Table 1: Demographic Composition

	Percentages						**Difference from CATI**			
	CATI	Llama English	Llama Swahili	ChatGPT English	ChatGPT Swahili		Llama English	Llama Swahili	ChatGPT English	ChatGPT Swahili
18-24	25	25	25	25	25		0	0	0	0
25-34	29	29	29	29	29		0	0	0	0
35-44	23	23	23	23	23		0	0	0	0
45+	23	23	23	23	23		0	0	0	0
Male	49	49	49	49	49		0	0	0	0
Female	51	51	51	51	51		0	0	0	0
Urban	38	38	38	38	38		0	0	0	0
Rural	62	62	62	62	62		0	0	0	0
None	2	2	0	2	2		0	-2	0	0
Incomplete primary	11	11	34	11	11		0	23	0	0
Complete primary	14	14	21	14	15		0	8	0	1
Incomplete secondary	12	12	1	12	12		0	-11	0	0
Complete secondary	30	31	27	30	30		0	-3	0	0
Higher education	31	31	16	31	30		0	-14	0	-1
Professional/technical	15	15	38	15	15		0	24	0	0
Clerical or sales	2	2	0	2	2		0	-2	0	0
Own business/family	19	19	17	20	19		0	-2	0	0
Agricultural	25	25	25	25	27		0	0	0	2
Household/domestic	3	3	2	3	3		0	-1	0	0
Manual	25	24	0	25	23		-1	-24	0	-2
Student	3	3	3	3	4		0	0	0	0
No work	7	7	7	7	8		0	0	0	0
Other	0	1	6	1	0		1	5	1	0

For age, gender, and urban-rural, there are no differences: the distributions are the same between CATI and synthetic data.

However, the Llama Swahili synthetic data does not match CATI with respect to education and occupation: The Llama Swahili data has significantly less educated respondents than CATI: only 16% have higher education, compared to 31% in CATI. Llama Swahili also has fewer respondents in manual occupations compared to CATI (0% and 25%, respectively) and more respondents in a professional occupation (38% versus 15%, respectively). Notably, there were no such differences between CATI and the other three synthetic data sources. The discrepancy in Llama Swahili is curious: When creating each SDRs, we had explicitly instructed the SDR what its education and occupation was. However, SDRs in the Llama Swahili data did not choose an education or occupation that corresponded to their characteristics when we created the SDR.

4.2 Comparison of Survey Responses

4.2.1 Food

Table 3 shows the percentage of CATI and synthetic respondents who report eating different types of foods in the day prior to the interview, separately for breakfast, lunch, and supper. The left panel contains the point estimates (in percentages). The right panel contains the percentage point differences between CATI and each synthetic dataset. Differences larger than 5 or less than -5 are shaded in red.

There are extremely large differences between CATI and all the synthetic datasets (with the exception of organ meat, which was rarely selected by CATI or synthetic respondents. Some of these differences are extreme. Here, we report the results for breakfast, though similar differences exist for lunch and supper. Among CATI respondents, 72% said they ate cereals (i.e., starchy foods) for breakfast; cereals were almost never selected by SDRs. In contrast, only 7% of CATI respondents chose vegetables for breakfast, but 40-96% of synthetic data respondents chose vegetables.

Table 2. Food Consumption: Differences in Point Estimates between CATI (benchmark) and Synthetic Datasets

	Point Estimates (percentage points)						**Difference from CATI (percentage points)**			
BREAKFAST	CATI	Llama English	Llama Swahili	ChatGPT English	ChatGPT Swahili		Llama English	Llama Swahili	ChatGPT English	ChatGPT Swahili
Cereals	72	0	2	0	0		-72	-70	-72	-72
White roots or tubers	12	0	27	0	9		-12	15	-12	-3
Vegetables	7	96	73	40	91		90	67	34	84
Fruits	5	8	78	37	58		3	73	32	53
Organ meat	0	0	2	0	0		0	2	0	0
Flesh meat	1	83	58	11	21		82	57	10	21
Eggs	4	6	25	40	46		2	21	35	41
Fish and seafood	0	0	10	0	13		0	10	0	13
Legumes, seeds, and nuts	7	22	23	0	1		16	17	-7	-6
Milk and milk products	59	91	21	96	66		32	-38	37	7

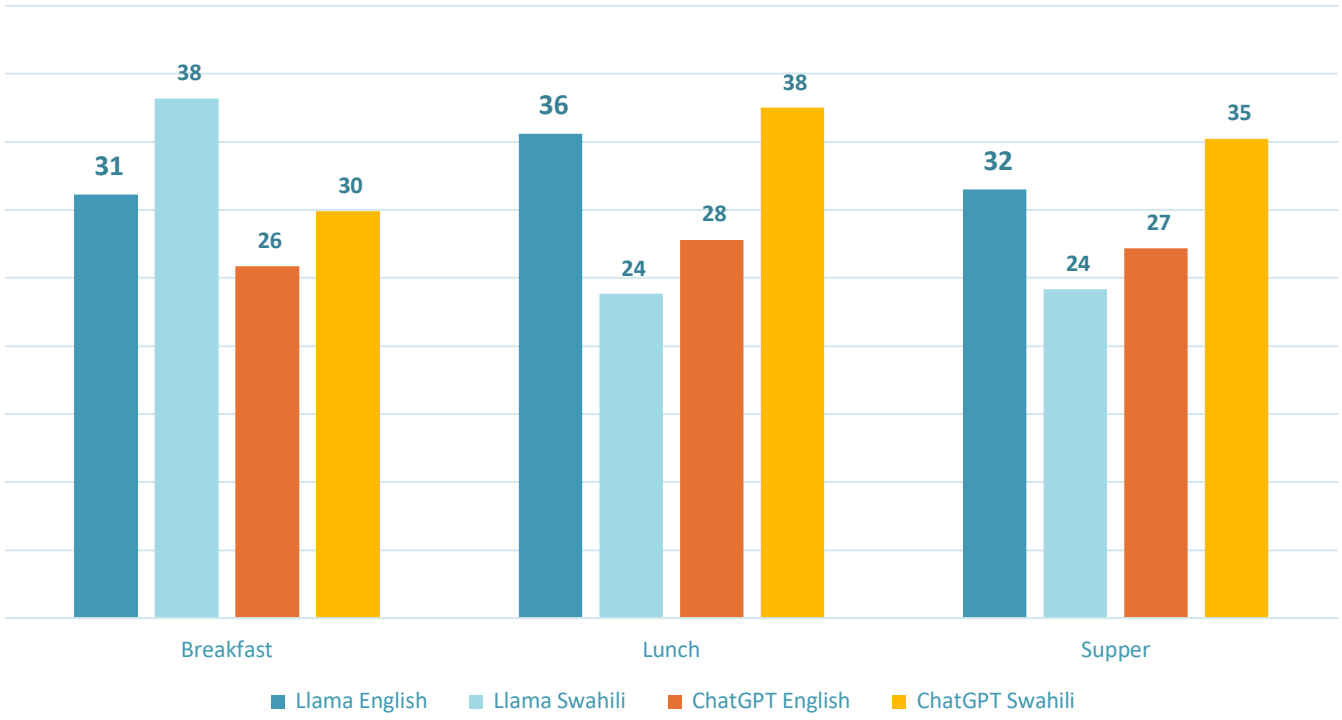
Oils and fats	15	0	17	0	9		-15	2	-15	-6
Sweets	70	34	12	0	1		-37	-58	-70	-69
Beverages	81	36	13	69	94		-45	-67	-11	13
	Point Estimates						**Difference from CATI**			
LUNCH	CATI	Llama English	Llama Swahili	ChatGPT English	ChatGPT Swahili		Llama English	Llama Swahili	ChatGPT English	ChatGPT Swahili
Cereals	78	0	4	0	0		-78	-75	-78	-78
White roots or tubers	12	0	13	0	1		-12	1	-12	-11
Vegetables	55	98	67	83	98		43	12	28	43
Fruits	8	3	76	0	11		-5	68	-8	2
Organ meat	0	0	0	0	0		0	0	0	0
Flesh meat	6	98	57	100	97		92	51	94	92
Eggs	1	3	25	2	22		2	24	2	21
Fish and seafood	3	18	10	0	13		15	7	-3	10
Legumes, seeds, and nuts	29	48	22	42	1		19	-7	13	-28
Milk and milk products	7	83	22	27	63		76	15	20	55
Oils and fats	62	0	21	9	3		-62	-41	-53	-59
Sweets	7	25	10	0	2		17	3	-7	-6
Beverages	8	49	14	51	90		41	6	43	82
	Point Estimates						**Difference from CATI**			
SUPPER	CATI	Llama English	Llama Swahili	ChatGPT English	ChatGPT Swahili		Llama English	Llama Swahili	ChatGPT English	ChatGPT Swahili
Cereals	93	1	4	0	0		-92	-89	-93	-93
White roots or tubers	8	0	11	0	1		-8	3	-8	-7
Vegetables	74	94	67	84	97		20	-7	10	23
Fruits	6	1	79	0	5		-5	72	-6	-1
Organ meat	0	0	0	0	0		0	0	0	0
Flesh meat	9	93	57	100	99		85	48	91	90
Eggs	3	2	22	2	17		-1	19	-1	14
Fish and seafood	8	4	10	5	22		-4	2	-3	14
Legumes, seeds, and nuts	19	30	18	20	1		12	-1	1	-18
Milk and milk products	13	76	24	38	52		63	11	25	39
Oils and fats	73	0	19	7	3		-73	-54	-66	-71
Sweets	8	27	12	0	1		18	4	-8	-7
Beverages	11	40	15	51	91		29	4	40	81

Table 1 contains a large volume of results (13 food groups x 3 meals x 4 datasets). To provide a more concise summary, we calculated the average aggregated average percentage point absolute differences between CATI and Synthetic Data. This calculation involves (1) taking the absolute value of the percentage point differences between CATI and the synthetic datasets (i.e., converting the values on the right panel of Table 1 to absolute value); (2) creating an average value of the 13 absolute values for each LLM. For example, for the Llama English in Supper, we average the absolute 92 (cereals), 8 (roots/tubers), 20 (vegetable) … 18 (sweets), 29 (beverages).

Across the 13 food groups for breakfast, the average absolute difference between CATI and SDRs was 31 percentage points (Llama English), 38 percentage points (Llama Swahili), 26 percentage points (ChatGPT English), 30 percentage points (ChatGPT Swahili). Chi-square tests of association showed significant differences between CATI and SDR respondents for breakfast ($p < .01$), lunch ($p < .01$), and dinner ($p < .01$). These values are shown in Figure 2 below.

There are no clear differences between English and Swahili (within LLM): sometimes the differences are larger for English, but other times, the differences are larger for Swahili. Likewise, there is no clear difference by LLM (Llama versus ChatGPT).

Figure 1. Food Consumption: Aggregated Percentage Point Differences between CATI and Synthetic Data, by Meal



4.2.2 Media and Technology

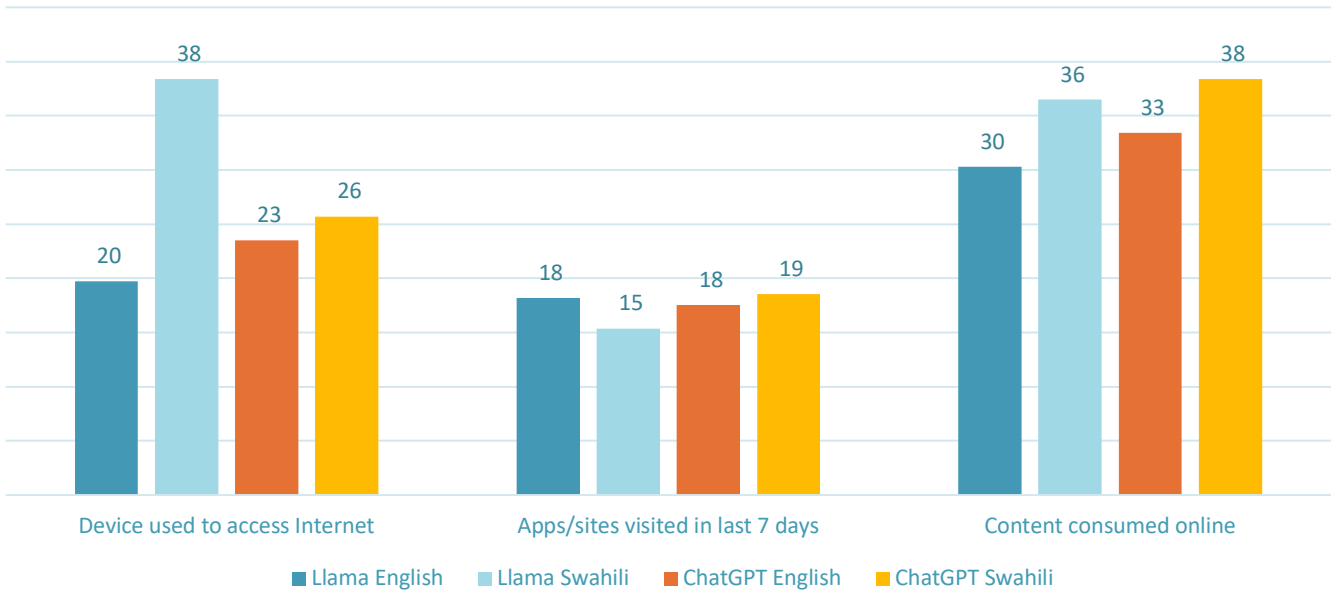
Next, we compare indicators related to media and technology. Table 2 shows that all synthetic datasets overestimate TV and Internet use, relative to the CATI data. Radio use is also overestimated, except that Llama English slightly underestimates radio use.

Table 3. Use of TV, Radio, Internet: Differences in Point Estimates between CATI (benchmark) and Synthetic Datasets										
	Point Estimates						Difference from CATI			
OVERVIEW	CATI	Llama English	Llama Swahili	ChatGPT English	ChatGPT Swahili		Llama English	Llama Swahili	ChatGPT English	ChatGPT Swahili
Watch TV weekly	74	100	94	100	100		26	20	26	26
Listen to radio weekly	69	59	85	84	84		-9	16	16	16
Access internet weekly	65	99	95	100	99		34	30	35	35

The surveys included many questions about online activities, among those who access the Internet weekly. Due to the large volume of data, we present the aggregated percentage point differences between and CATI and synthetic datasets (similar to Figure 1). Please see Appendix Table 1 for the detailed results for each survey question.

Figure 2a shows large differences between CATI and synthetic data sources. The differences are consistent across LLM and language used in the LLM (English versus Swahili), with the exception of Llama Swahili for device used to access the Internet (average percentage point difference.) Compared to other questions, apps/websites visited in the last 7 days had a smaller difference (15-19 percentage points).

Figure 2a. Media and Technology: Aggregated Percentage Point Differences between CATI and Synthetic Data

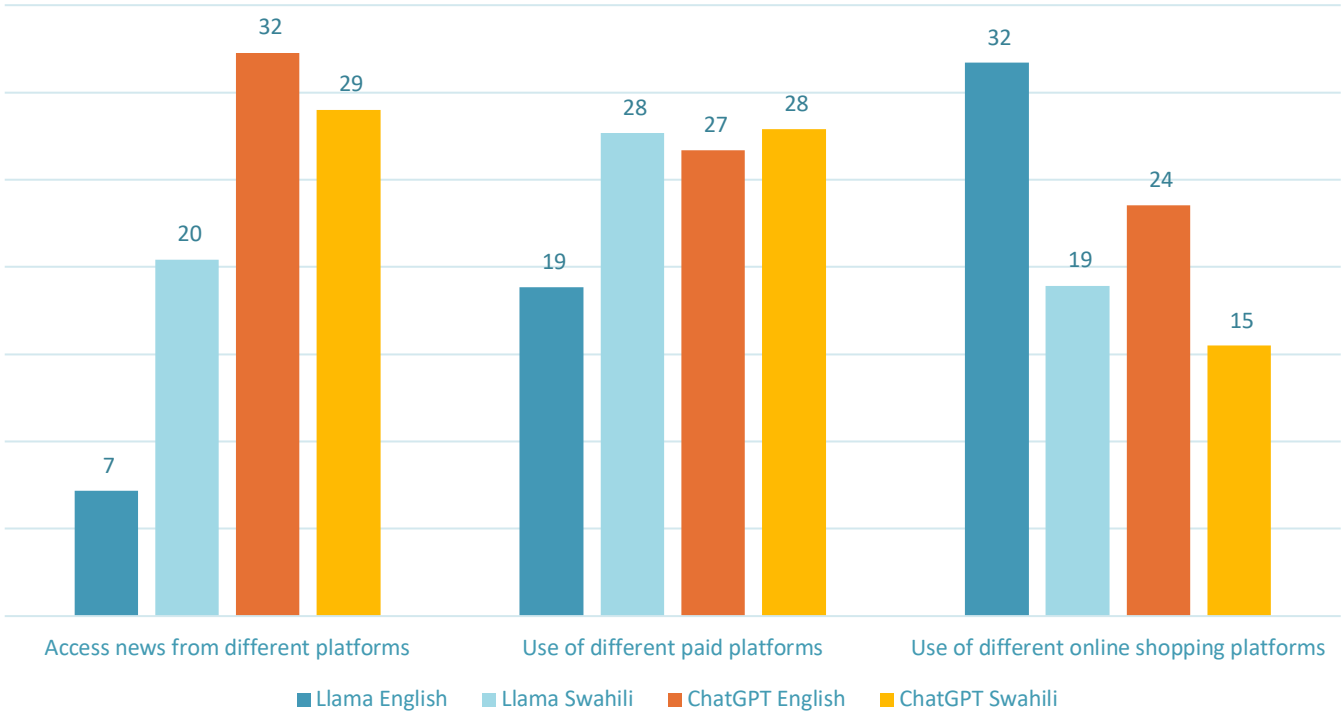


In general, the synthetic data overestimate technology use. For example, Appendix Table 1 shows that only 14% of CATI online respondents said they used laptop/desktop to access the Internet, compared to over 97% of synthetic data respondents. Training data in the LLMs may not reflect the contextual understanding that Kenyans mostly access the Internet via phones, not laptop or desktop computers.

For most but not all indicators, the synthetic data produce similar (and incorrect) results. As an example, consider the question on what apps or websites respondents visited in the last seven days. Sometimes synthetic data *under*estimates the true value: 16% of CATI respondents report using Snapchat compared to <1% in the synthetic data. In other cases, synthetic data *over*estimates the true value: Only 80% of CATI respondents report using Facebook, compared to 97%+ for synthetic data. In other cases, the synthetic data diverge from each other. For example, 26% of CATI respondents report using Instagram, compared to 80% (Llama English) and 85% (Llama Swahili) – but only 28% ChatGPT English and 11% of ChatGPT Swahili. In sum, results are highly variable, unpredictable, and there is no discernible pattern by LLM or language used.

Similarly, the results in Figure 2b are also highly variable. Llama English performs well when estimating news from different platforms, but performs the worst for estimating use of different online shopping platforms. In almost every case, there are large and statistically significant differences between CATI and the synthetic data sources.

Figure 2b. Media and Technology: Aggregated Percentage Point Differences between CATI and Synthetic Data



4.2.3 Humanitarian Assistance

Table 3 below shows comparisons between CATI and synthetic data sources regarding humanitarian indicators. Receiving humanitarian assistance is similar between CATI and 3 out of the 4 synthetic data sources, but ChatGPT Swahili overestimates humanitarian assistance by 83 percentage points. Attitudes about effective aid providers vary substantially between CATI and all four synthetic data sources.

Table 4. Humanitarian Assistance: Differences in Point Estimates between CATI (benchmark) and Synthetic Datasets										
	Point Estimates						**Difference from CATI**			
	CATI	Llama English	Llama Swahili	ChatGPT English	ChatGPT Swahili		Llama English	Llama Swahili	ChatGPT English	ChatGPT Swahili
Received humanitarian assistance in the last 2 years	8	17	4	21	91		9	-3	13	83
Believe that aid is effective in reaching areas where it is needed most	21	1	9	-	0		-20	-12	-21	-20
Most effective aid providers, according to respondent (multiple responses possible)										
Local authorities	84	71	70	86	59		-13	-14	2	-25
Local NGOs	9	80	48	99	64		71	39	90	55
International NGOs	18	50	50	95	68		32	32	77	50
Kenya Red Crescent Society	27	67	76	95	49		39	49	67	22
UN Agency	7	80	69	98	92		72	62	91	85
Religious group	10	3	8	21	5		-7	-1	11	-5
Military groups	0	-	3	-	0		0	3	0	0
Local business	3	1	17	1	1		-1	14	-2	-1
Individuals	12	1	35	1	15		-11	23	-11	3

4.4 Validity of CATI and Synthetic Data

To test the validity of synthetic data, we explored whether a well-established correlation (i.e., a negative association between education and food insecurity) is observed in synthetic data. This analysis moves beyond simple univariate statistics (previous sections) and provides a more rigorous test of criterion validity. We used three indicators of food insecurity. In this analysis, we show the percentage of respondents that said "sometimes [3-10 times]" or "often [more than 10 times]" to the three questions.[4] Higher values represent more food insecurity.

[4] *No Food:* In the past 30 days, was there ever no food to eat of any kind in your house because of lack of resources to get food. If yes, how often?

Go to Sleep Hungry: In the past 30 days, did you or any household member go to sleep at night hungry because there was not enough food? If yes, how often?

Whole Day: In the past 30 days, did you or any household member go a whole day and night without eating anything at all because there was not enough food?

Results are shown in Figures 3a, 3b, and 3c.

In this analysis, we are less concerned with the absolute values, and more focused on whether the synthetic data shows the same association as CATI. As expected, less educated CATI respondents report more food insecurity than more educated CATI respondents.

Figure 3a shows the results for the "No Food" indicator. Both ChatGPT synthetic datasets show the expected correlation (Chi-square test < .05 for both datasets). However, both Llama datasets do not show the expected correlation. For example, in the Llama English data, 96% of SDRs who didn't complete primary school said there was no food present sometimes/often – and 96% of SDRs who completed secondary school said they lacked food in the same way. This is a non-sensical result; there is no association in the Llama datasets between education and food insecurity.

Figure 3b shows the results for the "Go To Sleep Hungry" indicator. Only ChatGPT English shows the expected correlation. Figure 3c shows the results for the "Whole Day" indicator. Only Llama English shows the expected correlation.

To summarize: As a whole, the synthetic data performed poorly. Out of 12 possible associations (4 synthetic datasets x 3 indicators), the expected correlation only appeared in 4 times.

Figure 3a. No Food

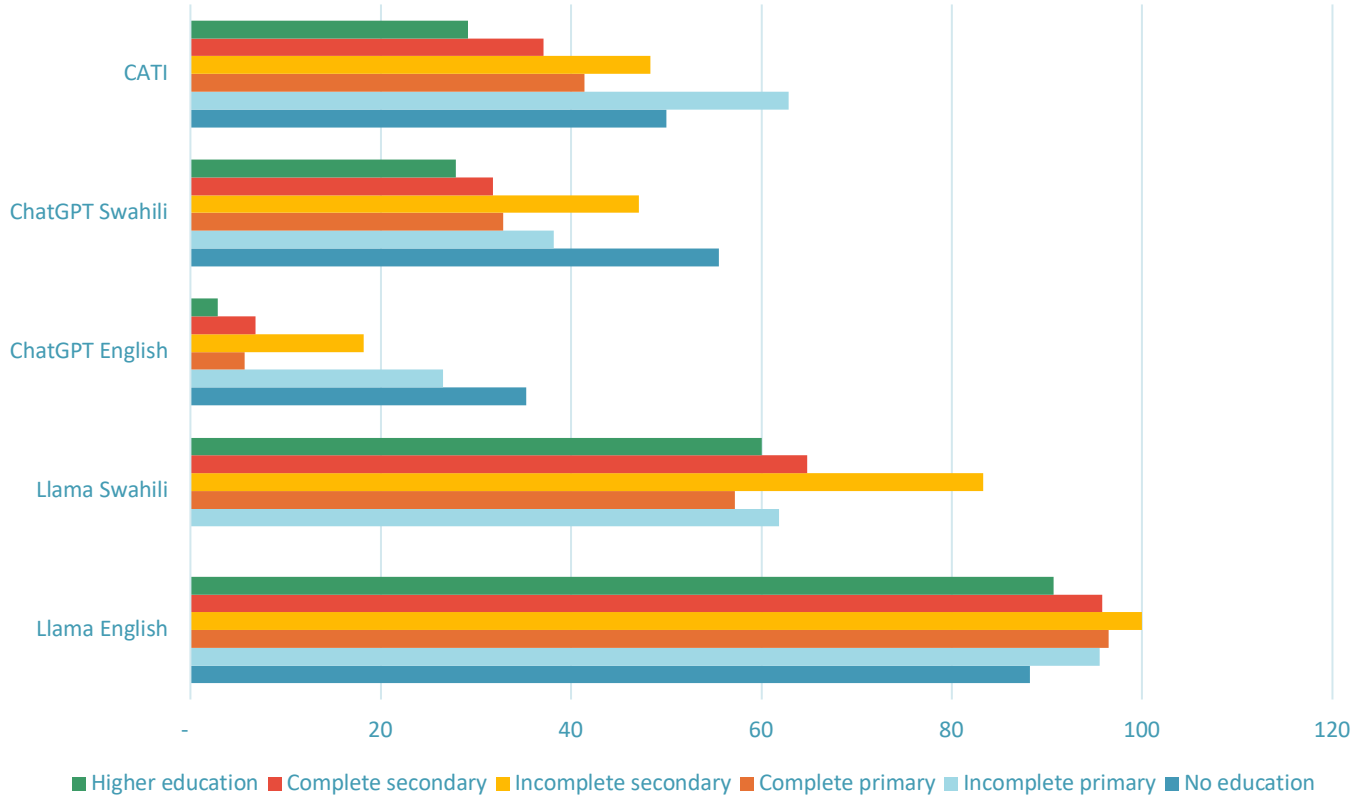


Figure 3b. Go To Sleep Hungry

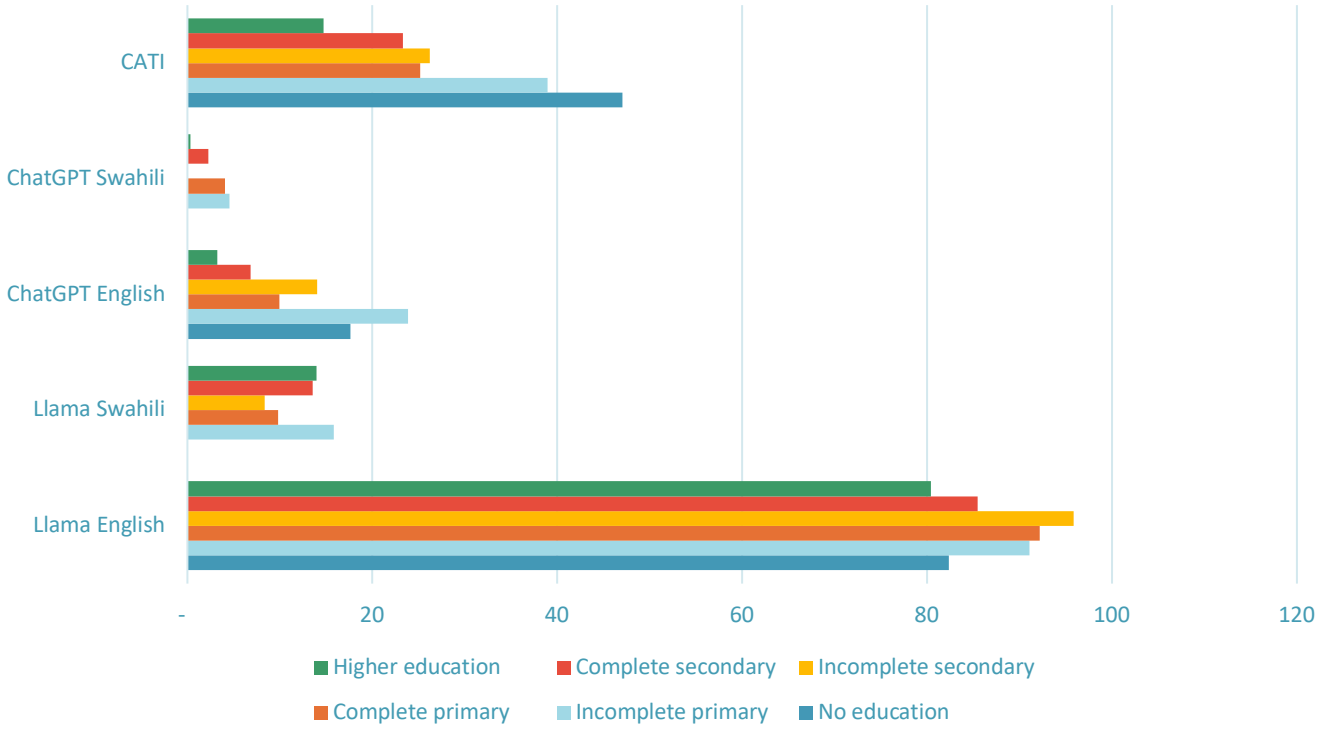
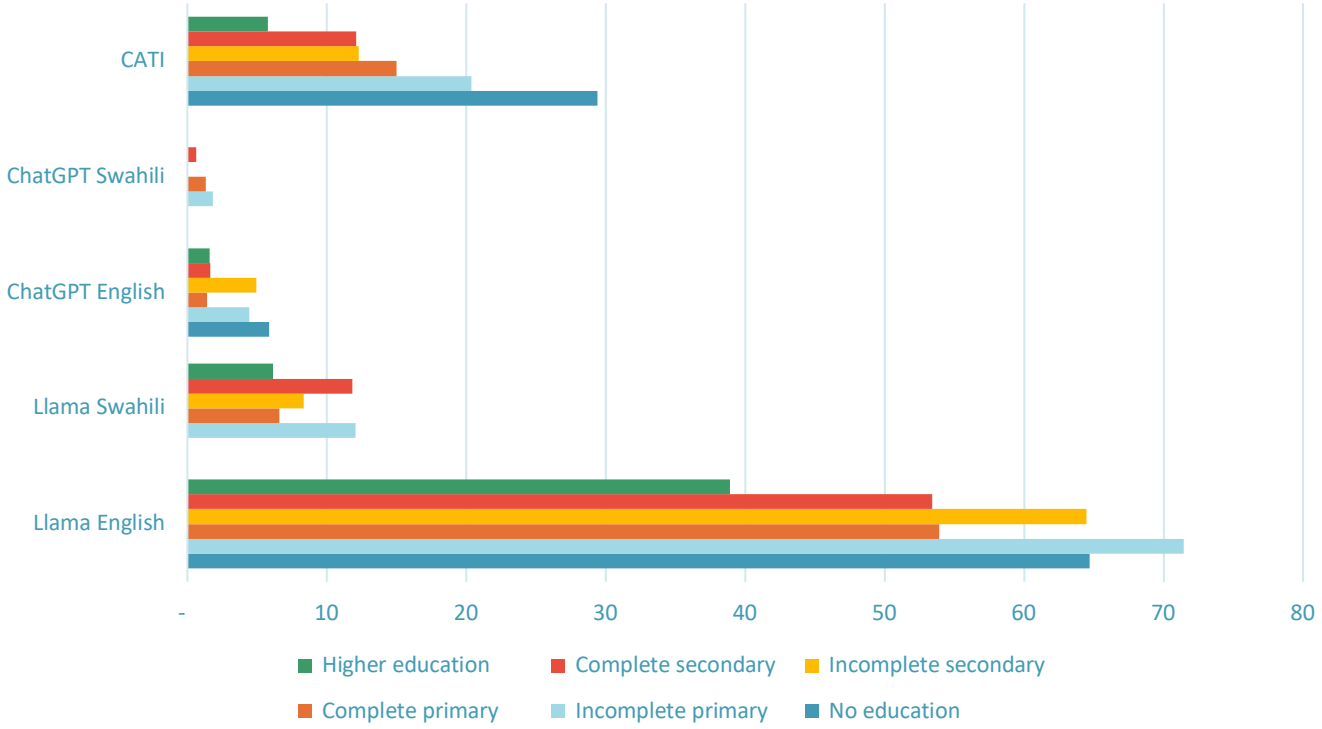


Figure 3c. Whole Day Without Food



5. DISCUSSION

This study evaluates the performance of synthetic data generated by large language models (LLMs) in Kenya, comparing these data against benchmark survey data collected through CATI (i.e., a phone survey) with real-world respondents. By examining multiple outcome variables across diverse topical areas (e.g., food consumption, media and technology use, and humanitarian aid), the study advances our understanding of the capabilities and limitations of synthetic data in less developed contexts.

5.1 Summary of Results

The results indicate notable discrepancies between synthetic data and the benchmark CATI data, with significant variations in performance across LLMs and prompting languages. While the demographic composition of synthetic datasets generally mirrored the CATI data, some anomalies were observed, particularly with the Llama Swahili dataset, where SDRs diverged in education and occupational attributes despite explicit prompting.

Food consumption data revealed large and systematic differences between CATI and synthetic datasets. Synthetic data often underestimated common dietary staples like cereals and overrepresented less commonly consumed items such as vegetables and milk. Similarly, synthetic respondents displayed systematic errors in reporting media and technology use. For instance, synthetic data vastly overestimated internet access through devices like laptops, which contrasts with the predominantly mobile-driven internet use in Kenya.

Correlations between education and food insecurity, an established relationship in real-world data, were inconsistently replicated. Out of 12 tests, only four showed the expected negative association, highlighting the synthetic data's struggle to capture such nuanced patterns accurately. Differences across LLMs (ChatGPT vs. Llama) and languages (English vs. Swahili) were generally minor, with no clear evidence favoring one approach over another.

5.2 Implications of Results

The findings underscore both the potential and the challenges of using synthetic data in diverse contexts. Our study demonstrated the *feasibility* of producing synthetic data using two different LLMs and using prompting in English and Swahili. However, our evaluation of the performance of LLMs showed serious, large, and unpredictable errors. In this section, we provide several implications of our results.

First, our results call into question the increasingly common use of LLMs to generate synthetic data in less developed contexts such as Kenya. LLMs are trained largely in English using training data from the developed West. While our study cannot conclusively point to the reason the synthetic data failed, we suspect the lack of contextually relevant training data is a potential reason. The lack of contextually relevant training data may be particularly evident for highly contextual topics such as food consumption. The LLMs we used may be especially unlikely to have training data about the types of foods Kenyans eat at different meals. But the LLMs even failed to replicate even basic findings about technology, for which there may be more training data – for example, the fact that few Kenyans access the internet on a laptop or desktop. These issues highlight the need for localized training data or context-specific fine-tuning of LLMs to improve performance in settings such as Kenya.

Second, sometimes LLMs do not even follow the prompts that generated the SDRs. For example, the Llama Swahili LLM produced educational and occupational distributions that were not aligned to their own personas. Put simply, we instructed 25% SDRs to answer as if they worked in manual occupations, but when they answered a question about

their occupation, 0% of SDRs chose manual occupations. It is unclear why this error occurred with the Llama Swahili data and not the other three synthetic datasets, but it could be due to errors in how Llama processes Swahili text.

Third, the findings suggest that the choice of prompting language alone is insufficient to address the quality issues in synthetic data, as we found no meaningful or consistent difference between prompting in Swahili versus English. Based on our anecdotal experience, we have found that LLMs process Swahili text with more errors compared to English. Our results suggest the need for additional training of LLMs based in Swahili language data, which may better reflect the experiences and attributes of Kenyans, and, overall, the need for fine-tuning models with richer, language-specific data to ensure more accurate outputs in multilingual settings.

Beyond technical limitations, the findings offer important insights for practical applications and policy considerations. There is a clear opportunity for partnerships between AI developers and researchers to enhance the quality and representativeness of training data in underrepresented contexts. Such collaborations could lead to the development of LLMs that are more contextually aware and capable of producing data that reflects local realities more accurately. Policymakers, in particular, can play a critical role by supporting initiatives that strengthen data infrastructure in low-resource settings. Improving access to localized data would allow for more robust AI training datasets, enhancing the reliability and validity of synthetic data outputs.

These findings highlight both the potential and significant limitations of using LLM-generated synthetic data in diverse contexts.

5.3 Limitations

This study presents several limitations that should be acknowledged when interpreting the findings. First, each synthetic data respondent completed the survey only once. Ideally, multiple completions per respondent should be considered to stabilize results through averaging - our single-completion design could have increased the variability of the data. Future research should consider allowing multiple responses per synthetic respondent to mitigate random variation and better reflect stable patterns.

Second, the synthetic data was generated at a single point in time. LLMs are dynamic; as they receive updates, their performance and capabilities may evolve (e.g., Bisbee et al., 2023). Consequently, the validity and consistency of synthetic data might change over time. To account for this variability, future studies should conduct longitudinal research to assess how synthetic data quality shifts with updates to LLM algorithms.

Third, this study was geographically and linguistically confined to Kenya, focusing exclusively on English and Swahili prompts. While this focus offers valuable insights into an underrepresented context, it limits the generalizability of the findings. Expanding research efforts to include other low-resource countries and diverse linguistic environments would provide a broader understanding of the global applicability of synthetic data.

Finally, the study did not explore the potential effects of different survey designs or the role of prompt engineering in improving synthetic data outcomes. Future research should investigate how varying prompt structures or survey methodologies influence the validity of LLM-generated data. Understanding these dynamics could provide practical recommendations for improving synthetic data quality in underrepresented contexts.

5.4 Conclusions

This study builds on prior research, including Bisbee et al. (2024), confirming significant limitations when using LLM-generated synthetic data to replicate nuanced behavioral patterns in low-resource settings. While our Kenya-focused study shows synthetic data can reasonably approximate demographic characteristics, it fails to capture context-dependent realities, behavioral patterns, and established correlations—raising fundamental concerns for empirical research methodologies.

For the research community, these findings necessitate a more disciplined approach to synthetic data integration. Researchers must address the systemic underrepresentation of low-resource contexts through deliberate collaboration with local institutions, ensuring cultural, linguistic, and socioeconomic factors are accurately represented in datasets used for analysis and model training.

Our findings suggest prompt design significantly influences synthetic data quality. Observed discrepancies likely stem from both inherent model limitations and prompt construction. Future methodological research should investigate how refined prompting strategies can enhance data validity, particularly in diverse, underrepresented settings where contextual nuance is crucial.

As synthetic data becomes increasingly attractive for cost-constrained research, especially in low-resource environments, the research community must establish robust validation frameworks. We recommend implementing standardized protocols that require synthetic insights to be corroborated with real-world data before informing meaningful conclusions or policy recommendations.

While this study focused specifically on Kenya, the implications likely extend to other low-resource settings. Cross-regional comparative studies are needed to determine whether these discrepancies reflect Kenya-specific factors or represent systematic challenges in synthetic data generation across similar contexts.

Targeted methodological innovations could significantly improve synthetic data reliability. Promising approaches include fine-tuning models with regionally representative datasets, enhancing multilingual capabilities for local language processing, and integrating culturally contextual information. Organizations with deep local survey expertise, such as GeoPoll, represent valuable research partners for developing these capabilities.

For the research community to responsibly leverage synthetic data's potential, we must prioritize methodological rigor alongside contextual accuracy. This requires developing techniques that address systemic biases in training data, establishing rigorous validation standards, and deepening our understanding of synthetic data performance across diverse global contexts. Through these efforts, synthetic data can evolve into a more reliable complement to traditional research methodologies, particularly in resource-constrained environments where data collection remains challenging.

APPENDIX TABLES

Appendix Table 1. Differences by Media and Technology										
	Point Estimates						Difference from CATI			
DEVICE USE TO ACCESS INTERNET (Among Interet Users)	Point Estimates						Difference from CATI			
Laptop/desktop	14	97	100	100	100		82	85	85	86
Mobile device	99	100	100	100	100		1	1	1	1
Tablet	3	12	83	21	43		9	81	19	41
Smart TV	6	12	17	18	5		6	11	12	-1
Other	0	0	15	0	0		0	15	0	0
APPS/SITES VISITED IN LAST 7 DAYS (Among Interet Users)	Point Estimates						Difference from CATI			
Facebook	80	97	98	99	99		17	18	19	19
Snapchat	16	0	0	0	0		-16	-16	-16	-16
Instagram	26	80	85	28	11		53	59	2	-15
Youtube	60	86	84	99	99		26	23	39	39
LinkedIn	6	1	0	7	8		-5	-6	1	2
TikTok	69	1	1	0	0		-68	-68	-69	-69
Twitter/X	21	2	15	2	0		-19	-6	-19	-21
Pinterest	5	0	0	0	0		-5	-5	-5	-5
Reddit	3	0	0	0	0		-3	-3	-3	-3
Telegram	18	7	17	0	0		-10	0	-17	-18
Facebook Messenger	51	19	40	0	0		-31	-10	-50	-51
WhatsApp	84	97	94	100	99		13	10	16	15
Discord	2	0	0	0	0		-1	-1	-2	-2
Twitch	2	0	0	0	0		-2	-2	-2	-2
Threads	3	0	0	0	0		-3	-3	-3	-3
CONTENT NORMALLY CONSUMED ONLINE	Point Estimates						Difference from CATI			
News/current events	87	87	87	100	99		0	0	13	12
Soap operas and drama	55	93	25	78	68		38	-30	23	13
Sports	62	83	91	58	87		21	29	-4	26
Cartoons / children's shows	32	60	12	0	2		29	-20	-31	-30

Comedy	67	68	43	52	29		0	-24	-15	-38
Romance	32	11	49	0	3		-21	17	-32	-28
Food / Cooking	60	26	2	50	3		-34	-58	-11	-57
Game shows / Competitions	62	15	3	0	0		-47	-59	-62	-62
Health / exercise	65	0	23	5	23		-65	-43	-61	-43
Home-related / How to / DIY	52	0	4	0	4		-52	-48	-52	-49
Reality	57	54	2	2	3		-3	-56	-55	-54
Fantasy / science fiction	35	12	1	0	0		-24	-34	-35	-35
Educational	65	6	7	23	12		-58	-57	-41	-53
OBTAIN NEWS WEEKLY FROM THE FOLLOWING PLATFORMS	**Point Estimates**						**Difference from CATI**			
Television	71	87	99	96	99		16	28	24	28
Radio	62	64	96	29	94		2	34	-33	32
Print media	24	9	46	0	2		-16	22	-24	-22
News website	39	36	44	97	95		-4	4	57	56
Online videos and streaming	46	51	39	65	45		5	-6	19	-1
Social media	62	64	35	98	98		1	-28	35	36
USE OF PAID PLATFORMS IN THE LAST 4 WEEKS	**Point Estimates**						**Difference from CATI**			
Netflix	5	99	96	54	90		94	91	49	85
Youtube Premium	2	1	22	0	6		-1	19	-2	3
Amazon Prime	0	68	74	0	1		68	74	0	1
Viusasa	1	0	31	93	79		-1	30	92	78
Startimes On	1	42	61	13	0		41	60	13	0
DSTV Now	2	97	64	42	4		95	62	40	2
Showmax	1	61	41	77	88		60	40	76	87
Apple TV	0	4	2	0	0		4	2	0	0
Disney +	0	8	23	0	0		8	22	0	0
HBO Max	0	1	7	0	0		1	7	0	0
Hulu	0	4	6	0	0		4	6	0	0
SHOP ONLINE IN LAST 12 MONTHS	**Point Estimates**						**Difference from CATI**			
Yes	17	96	98	89	78		80	82	73	61

ONLINE PLATFORMS FOR SHOPPING	**Point Estimates**						**Difference from CATI**			
Jumia Online Mall	52	96	89	100	99		44	37	48	48
Kilimall	32	90	80	1	24		58	48	-31	-9
OLX	0	90	63	16	2		90	63	16	2
Masoko	1	58	42	0	16		57	41	0	15
Amazon	1	59	30	0	1		58	29	-1	-1
Jumia Food	2	7	4	100	99		5	2	98	97
Uber Eats	2	8	6	59	29		6	4	57	27
Pigiame	0	10	19	0	0		10	19	0	0
Glovo	8	9	13	69	11		1	5	61	3
Instagram	1	43	5	5	1		42	3	4	0
Facebook	10	28	5	7	16		18	-5	-3	6
Cheki	0	3	5	0	0		3	5	0	0
My Dawa	1	1	1	2	0		1	1	1	-1
Jiji	9	59	10	0	0		51	1	-9	-9

REFERENCES

Bisbee, James, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. 2024. "Synthetic Replacements for Human Survey Data? The Perils of Large Language Models." *Political Analysis*. https://doi.org/10.1017/pan.2024.5.

Brubaker, Joshua Milton, Talip Kilic, and Philip Wollburg. 2021. "Representativeness of Individual-Level Data in COVID-19 Phone Surveys: Findings from Sub-Saharan Africa." *PLOS ONE* 16 (11): 1–27. https://doi.org/10.1371/journal.pone.0258877.

Glazerman, Steven, Karen A Grepin, Valerie Mueller, Michael Rosenbaum, and Nicole Wu. 2023. "Do Referrals Improve the Representation of Women in Mobile Phone Surveys?" *Journal of Development Economics* 162. https://doi.org/10.1016/j.jdeveco.2023.103077.

Greenleaf, A R, A Gadiaga, Y Choi, G Guiella, S Turke, N Battle, S Ahmed, and C Moreau. 2020. "Automated and Interviewer-Administered Mobile Phone Surveys in Burkina Faso: Sociodemographic Differences Among Female Mobile Phone Survey Respondents and Nonrespondents." *JMIR MHEALTH AND UHEALTH* 8 (7). https://doi.org/10.2196/17891.

Guzman-Tordecilla, D N, A I Vecino-Ortiz, A Torres-Quintero, C Solorzano-Barrera, J Ali, R E Peñaloza-Quintero, S Ahmed, G W Pariyo, V Maniar, and D G Gibson. 2023. "Examination of the Demographic Representativeness of a Cross-Sectional Mobile Phone Survey in Collecting Health Data in Colombia Using Random Digit Dialling." *BMJ Open* 13 (6): e073647-. https://doi.org/10.1136/bmjopen-2023-073647.

Heyde, Leah von der, Anna-Carolina Haensch, and Alexander Wenz. 2023. "Vox Populi, Vox AI? Using Language Models to Estimate German Public Opinion." https://doi.org/10.31235/osf.io/8je9g.

Heyde, Leah Von Der, Anna-Carolina Haensch, and Alexander Wenz. n.d. "United in Diversity? Contextual Biases in LLM-Based Predictions of the 2024 European Parliament Elections."

Lambrecht, I, van Asselt J, D Headey, B Minten, P Meza, M Sabai, T S Sun, and H E Win. 2023. "Can Phone Surveys Be Representative in Low- and Middle-Income Countries? An Application to Myanmar." *PLoS One* 18 (12): e0296292-. https://doi.org/10.1371/journal.pone.0296292.

Lau, C Q, A Cronberg, L Marks, and A Amaya. 2019. "In Search of the Optimal Mode for Mobile Phone Surveys in Developing Countries. A Comparison of IVR, SMS, and CATI in Nigeria." *Survey Research Methods* 13 (3): 305–18. https://doi.org/10.18148/srm/2019.v13i3.7375.

Qu, Yao, and Jue Wang. 2024. "Performance and Biases of Large Language Models in Public Opinion Simulation." *Humanities and Social Sciences Communications* 11 (1). https://doi.org/10.1057/s41599-024-03609-x.

Sun, Seungjong, Eungu Lee, Dongyan Nan, Xiangying Zhao, Wonbyung Lee, Bernard Jansen, and Jang Hyun Kim. 2024. "Random Silicon Sampling: Simulating Human Sub-Population Opinion Using a Large Language Model Based on Group-Level Demographic Information." https://doi.org/https://doi.org/10.48550/arXiv.2402.18144.

To request more information or raw data, contact us

info@geopoll.com

LinkedIn

www.GeoPoll.com