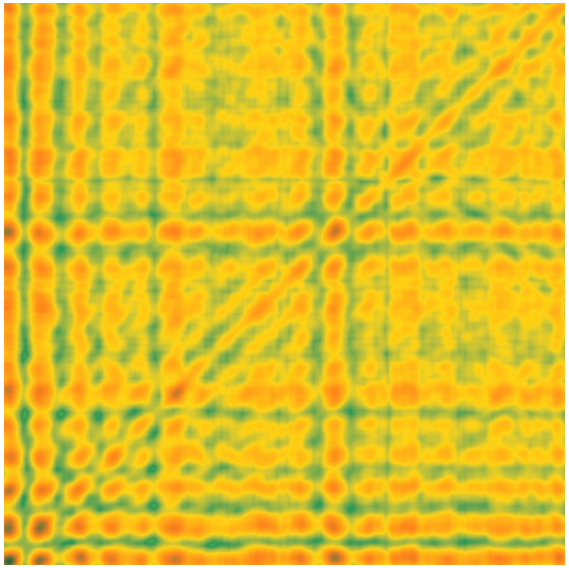
Synthetic Survey Data? It's Not Data

Chris Chapman, PhD

February 2026
Sawtooth Webinar Series



Preliminary

Chris

- **Psychologist** (*Clinical* + *Personality* PhD; also *Philosophy* MA)
- Co-chair, **Quant UX Con** , President, Quant UX Association — join us, November 2026!
- Write the **Quant UX Blog** , plus books about research and Quant UX
- Intersection with **Marketing** ; former President, AMA Insights Council

2000-2011	2012-2022	2022 - 2024	2021 -
General UXR, Microsoft 3 years **Windows** 7 years **Hardware** 1 year **Bing**	**Quant UXR, Google** 2 years **Ads** 1 year **Social Impact** 4 years **Cloud** 2 years **Chrome OS** 1 year **Next Billion Users**	**Industrial Design Research, Amazon** 2 years **Lab 126** (Devices) – **Echo** speakers – **Kindle** readers – **Fire** tablets – Cross-product research	**Quant UX Association** Co-founder Non-profit association organizes annual **Quant UX Conference** (1500+ attendees, 70+ countries)

What is “ synthetic [survey] data ”?

Sometimes it’s used to mean simulated data generically (e.g., from custom code)

Today, I’m using it to mean:

> *Survey **responses** obtained **from an LLM system***
> *That are used as **alternatives or additions** to data obtained from **people***

For instance:: an LLM takes our survey, and we analyze the responses

Some vendors are pushing this hard as a new business for data / augmentation

Overall Outline

01

Synthetic data is never *a priori* "good enough"

02

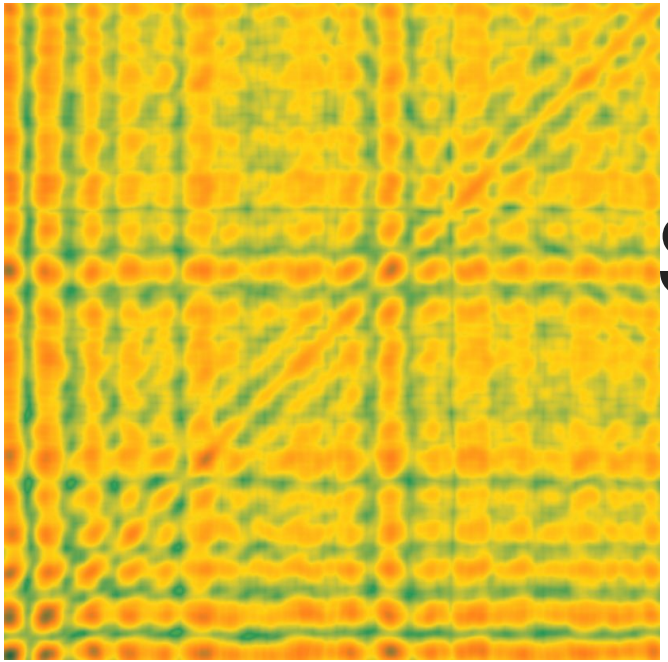
Synthetic data cannot be evaluated using statistical inference

03

Synthetic data misunderstands the point of surveys

04

"Rebuttals" & Counter-rebuttals



01

Synthetic Data is never *a priori* “good enough”

Two paths to find out what people think

Path 1

Identify your question

Ask people

Two paths to find out what people think

Path 1

- Identify your question
- Ask people

Path 2

- Compile 10B digital docs
- Build an LLM model
- Hire people to train it
- Create a way to query it
- Build 100s of data centers
- Collect some human data
- Use data as “digital twins”
- Write API code for the LLM
- Identify your question
- Ask the LLM for each “twin”

Two paths to find out what people think

Path 2

Compile 10B digital docs	Collect some human data
Build an LLM model	Use data as “digital twins”
Hire people to train it	Write API code for the LLM
Create a way to query it	Identify your question
Build 100s of data centers	Ask the LLM for each “twin”

There is no scientific or theoretical reason to expect this will work

It can fail in many ways:

- Biased data
- Biased training
- Poor algorithms
- Poor corpus coverage
- Poor “twin” coverage
- Poor prompting
- Change over time

The core problem

Path 2

Compile 10B digital docs	Collect some human data
Build an LLM model	Use data as “digital twins”
Hire people to train it	Write API code for the LLM
Create a way to query it	Identify your question
Build 100s of data centers	Ask the LLM for each “twin”

With an infinity of possible questions and changing real world conditions, there is no way to prove in advance that this process will generalize.

So what can we do with this path ?

Without any a priori expectation, the options are:

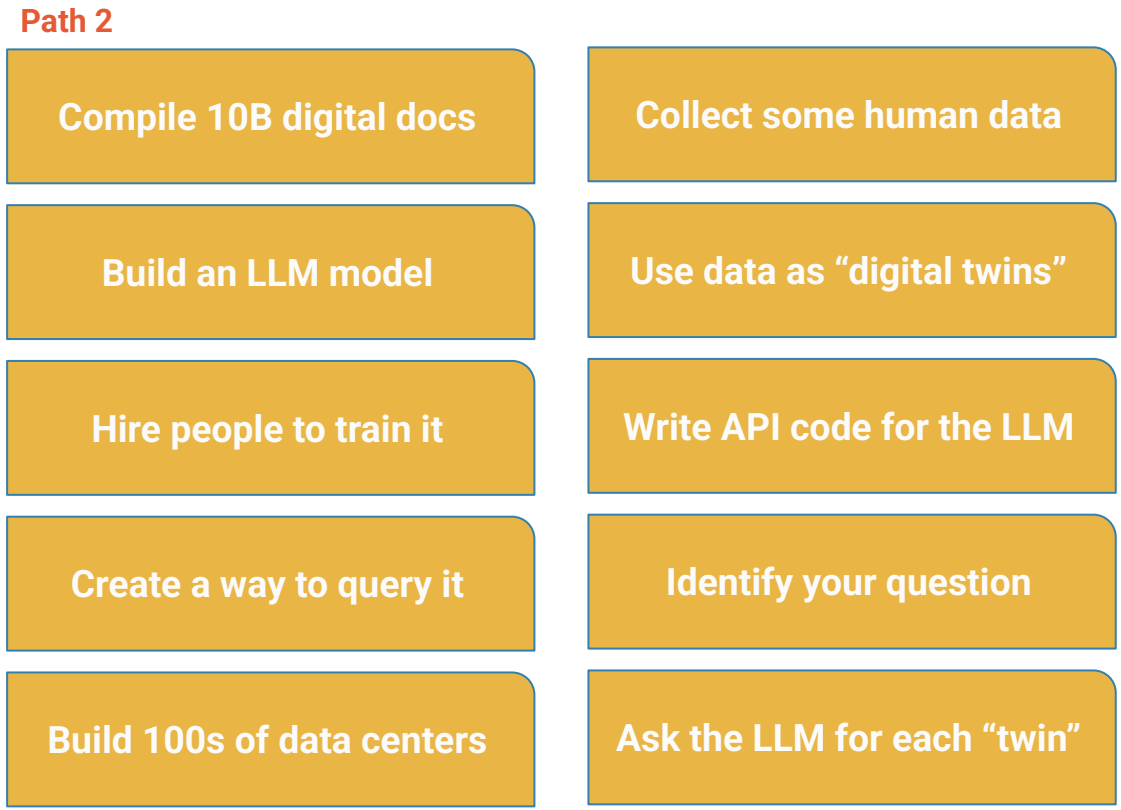
1. **Use it for trivial things**

   OR

2. **Prove it works for every project**

   OR

3. **Ignore logic and just believe it**



Two paths to find out what people think

Without any a priori expectation, the options are:

1. ~~Use it for trivial things~~

   ~~*OR*~~

2. **Prove it works for every project**

   ~~*OR*~~

3. ~~Ignore logic and just believe it~~

Path 2

Compile 10B digital docs	Collect some human data
Build an LLM model	Use data as “digital twins”
Hire people to train it	Write API code for the LLM
Create a way to query it	Identify your question
Build 100s of data centers	Ask the LLM for each “twin”

Two paths to find out what people think

Without any a priori expectation, the options are:

1. ~~Use it for trivial things~~

   ~~*OR*~~

2. **Prove it works for every project**

   ~~*OR*~~

3. ~~Ignore logic and just believe it~~

Requires data … so why not just use that data?

Path 2

Compile 10B digital docs	Collect some human data
Build an LLM model	Use data as “digital twins”
Hire people to train it	Write API code for the LLM
Create a way to query it	Identify your question
Build 100s of data centers	Ask the LLM for each “twin”

“But wait, it does work!”

First of all, a sanity check

- *Suppose I can predict people's real world preferences using synthetic data …*

First of all, a sanity check

- *Suppose I can predict people's real world preferences using synthetic data …*

- **I would predict people's real world preferences for stock market trades 💰💰💰**

First of all, a sanity check

- *Suppose I can predict people's real world preferences using synthetic data …*

- **I would predict people's real world preferences for stock market trades 💰💰💰**

⇒ *No one is known to be doing this successfully*
⇒ *LLM vendors sell 'data' rather than doing any such thing*

Inference: synthetic data can't generally predict preferences

No, no, still wait, it does work!

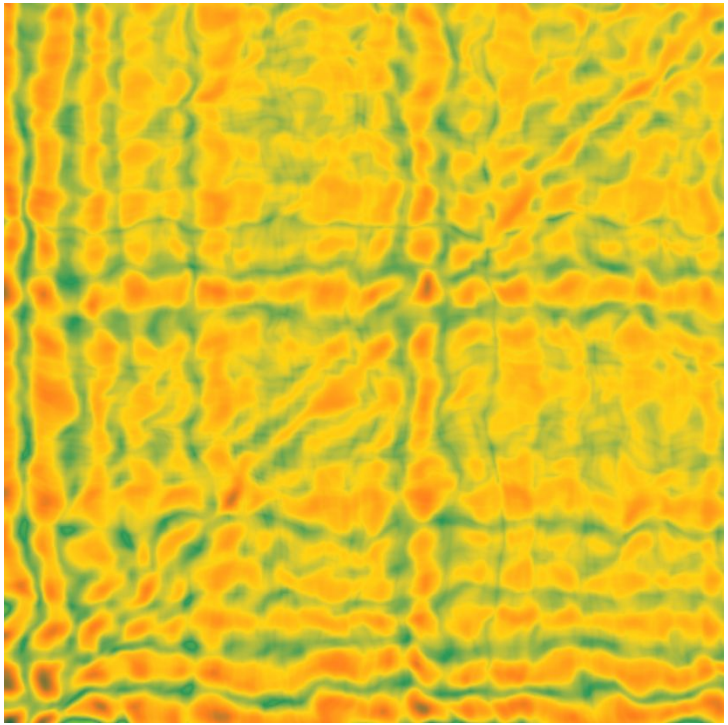
A vendor showed me a study.

A vendor, you say ?

… too easy! *But I'll bet they did what I describe next …*

02

It cannot be evaluated using statistics



Typical Process to “Compare LLMs to Human Surveys”

01

Collect or — more often — reuse a human sample

- *Often*: a sample of convenience
- … or left over from some previous study
- … or collected generically

Typical Process to “Compare LLMs to Human Surveys”

01	Collect or — more often — reuse a **human sample**	• *Often*: a sample of convenience • … or left over from some previous study • … or collected generically
02	Create “**digital twins**”	• Create a generic prompt template • Use the human data above to fill in the prompt template

Typical Process to “Compare LLMs to Human Surveys”

01	Collect or — more often — reuse a **human sample**	• *Often*: a sample of convenience • … or left over from some previous study • … or collected generically
02	Create “**digital twins**”	• Create a generic prompt template • Use the human data above to fill in the prompt template
03	Ask the LLM to “**take the survey**” for each twin	• Use the template to prompt responses to your survey

Typical Process to "Compare LLMs to Human Surveys"

01	Collect or — more often — reuse a **human sample**	• *Often*: a sample of convenience • … or left over from some previous study • … or collected generically
02	Create "**digital twins**"	• Create a generic prompt template • Use the human data above to fill in the prompt template
03	Ask the LLM to "**take the survey**" for each twin	• Use the template to prompt responses to your survey
04	**Compare** results from the LLM to human data	• Compare those LLM responses to … data from #1 above; … or a new sample of people

Problems!

01	Collect or — more often — reuse human sample	• *Often*: a sample of convenience • … or left over from some previous study • … or collected generically
02	Create “digital twins”	• Create a generic prompt template • Use the human data above to fill in the prompt template
03	LLM “takes the survey” for each twin	• Use the template to prompt responses to your survey
04	Compare results from the LLM to human data	• Compare those LLM responses to … data from #1 above; … or a new sample of people

Was it cherry-picked?
Does it generalize to your problem?

Is the template valid?
Are the data points relevant & complete?

LLM “taking a survey” != any human process

Statistical assumptions don’t apply …

Example LLM prompt* for a digital twin

" Your job is now to act as a substitute for a human respondent.

I am going to give you a persona to adopt, a question to answer, and a set of responses to answer with.

You must answer the question in the way you think the given persona would answer, using only one of the given responses, verbatim.

Here is your persona: <persona>

Here is the question: <question>

Here are your answer options: <options>

Now please return just the text of the response to that question that you think is most likely given the question and persona."

* From: Morris & Verasight Data Team (2025), *Your Polls on ChatGPT*. At https://report.verasight.io/synthetic-sampling

Example

" Your job is now to act as a substitute for a human respondent.

I am going to give you a persona to adopt, a question to answer, and a set of responses to answer with.

You must answer the question in the way you think the given persona would answer, using only one of the given responses, verbatim.

Here is your persona: <persona>

Here is the question: <question>

Here are your answer options: <options>

Now please return just the text of the response to that question that you think is most likely given the question and persona."

Matched as a digital twin to one respondent in a human data set

You are a [age] year old [gender] of [hisp/race] race/ethnicity.

Your education level is [edu], and you make [$] US dollars per year.

You live in [state] in the [region] region of the United States.

In terms of political parties, you identify more as a [ideology] and vote more with the [party] party.

From: Morris & Verasight Data Team (2025), *Your Polls on ChatGPT*. At https://report.verasight.io/synthetic-sampling

Example

" Your job is now to act as a substitute for a human respondent.

I am going to give you a persona to adopt, a question to answer, and a set of responses to answer with.

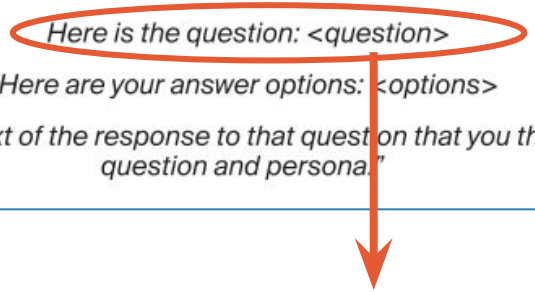
You must answer the question in the way you think the given persona would answer, using only one of the given responses, verbatim.

Here is your persona: <persona>

Here is the question: <question>

Here are your answer options: <options>

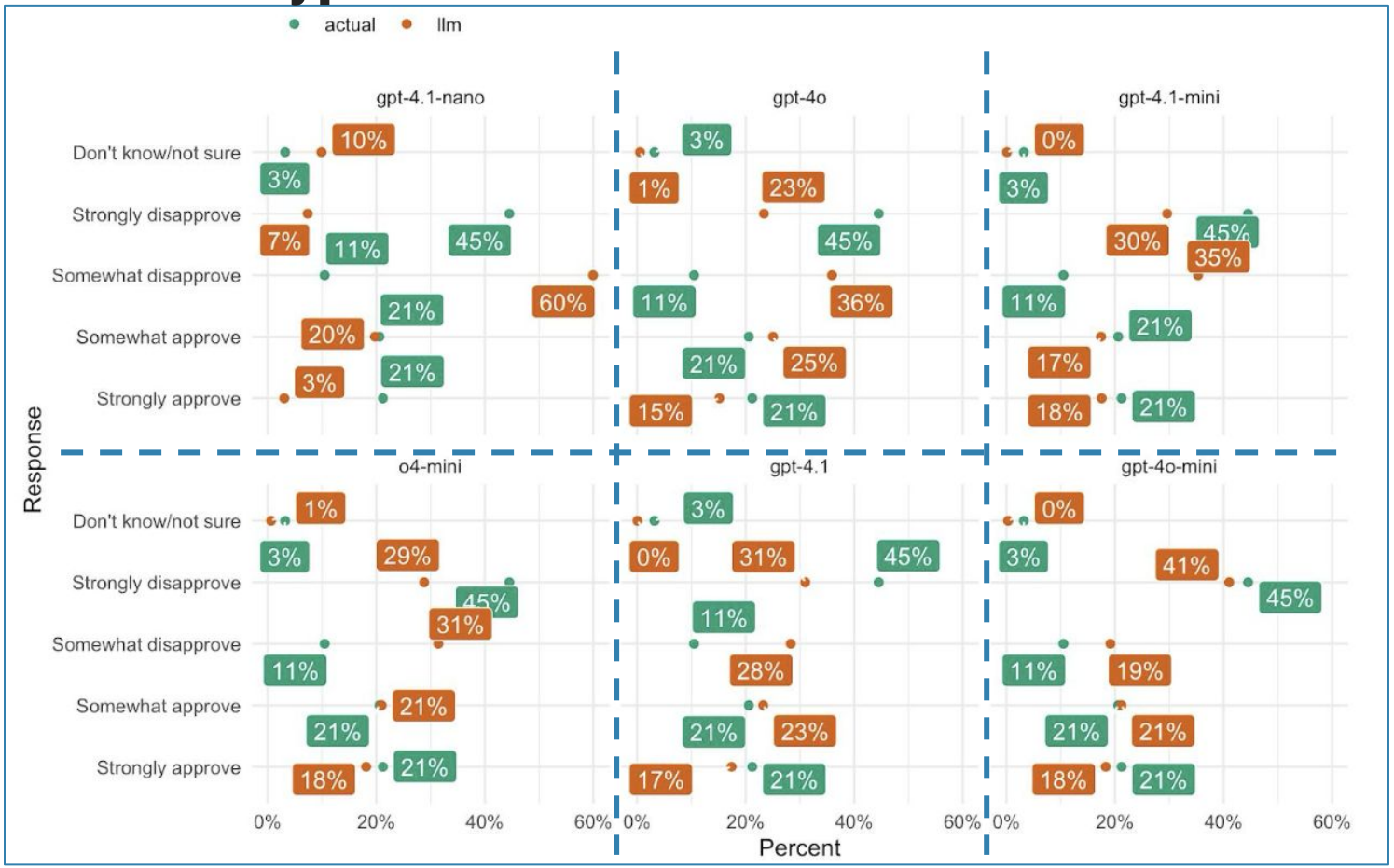
Now please return just the text of the response to that question that you think is most likely given the question and persona."



""Do you approve or disapprove of the way Donald Trump is handling his job as president?"

From: Morris & Verasight Data Team (2025), *Your Polls on ChatGPT*. At https://report.verasight.io/synthetic-sampling

Some typical results — Morris et al – across 6 LLM models

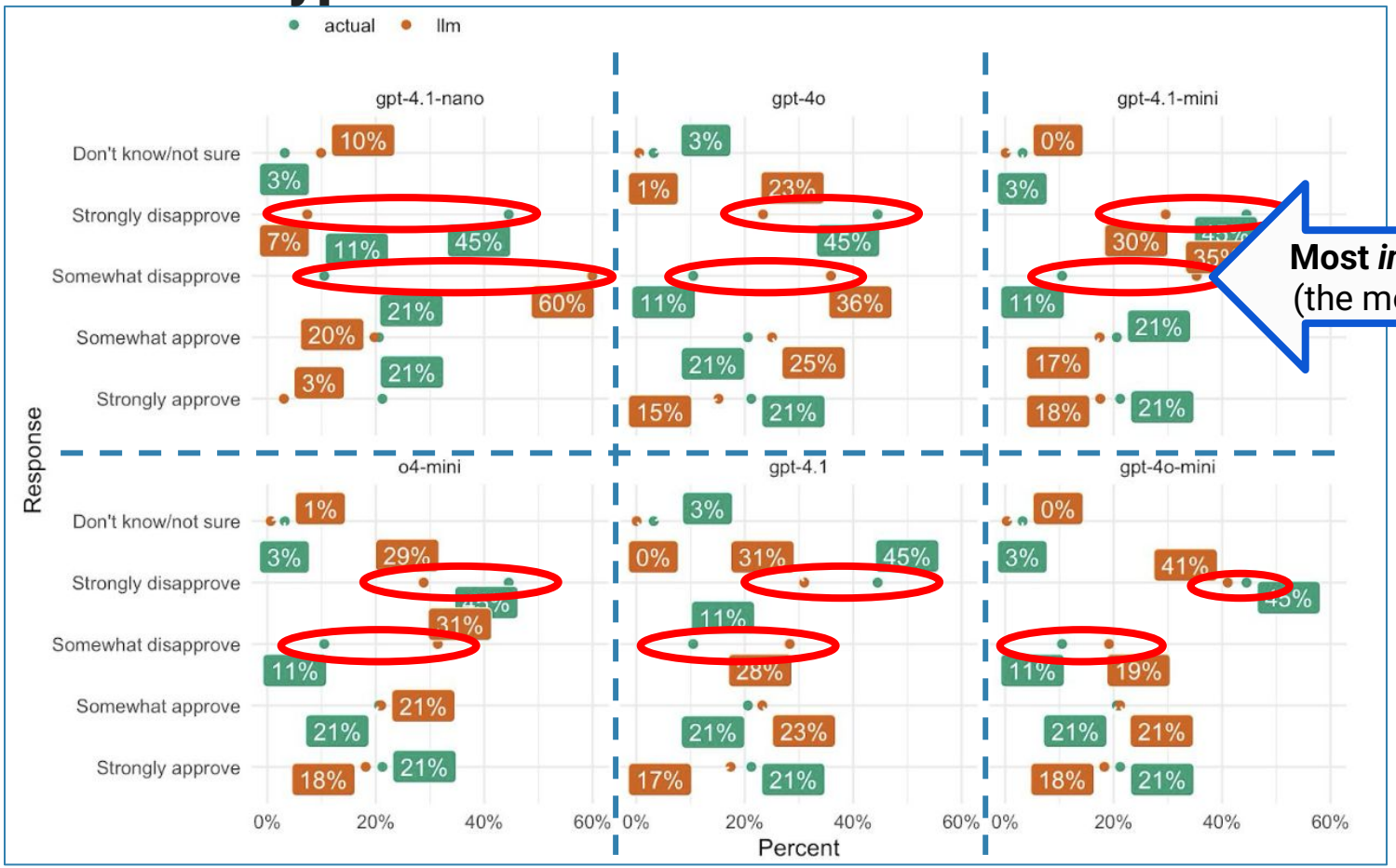


Legend

- **Actual human data**
- **LLM model results**

Morris & Verasight Data Team (2025), *Your Polls on ChatGPT*. At https://report.verasight.io/synthetic-sampling

Some typical results

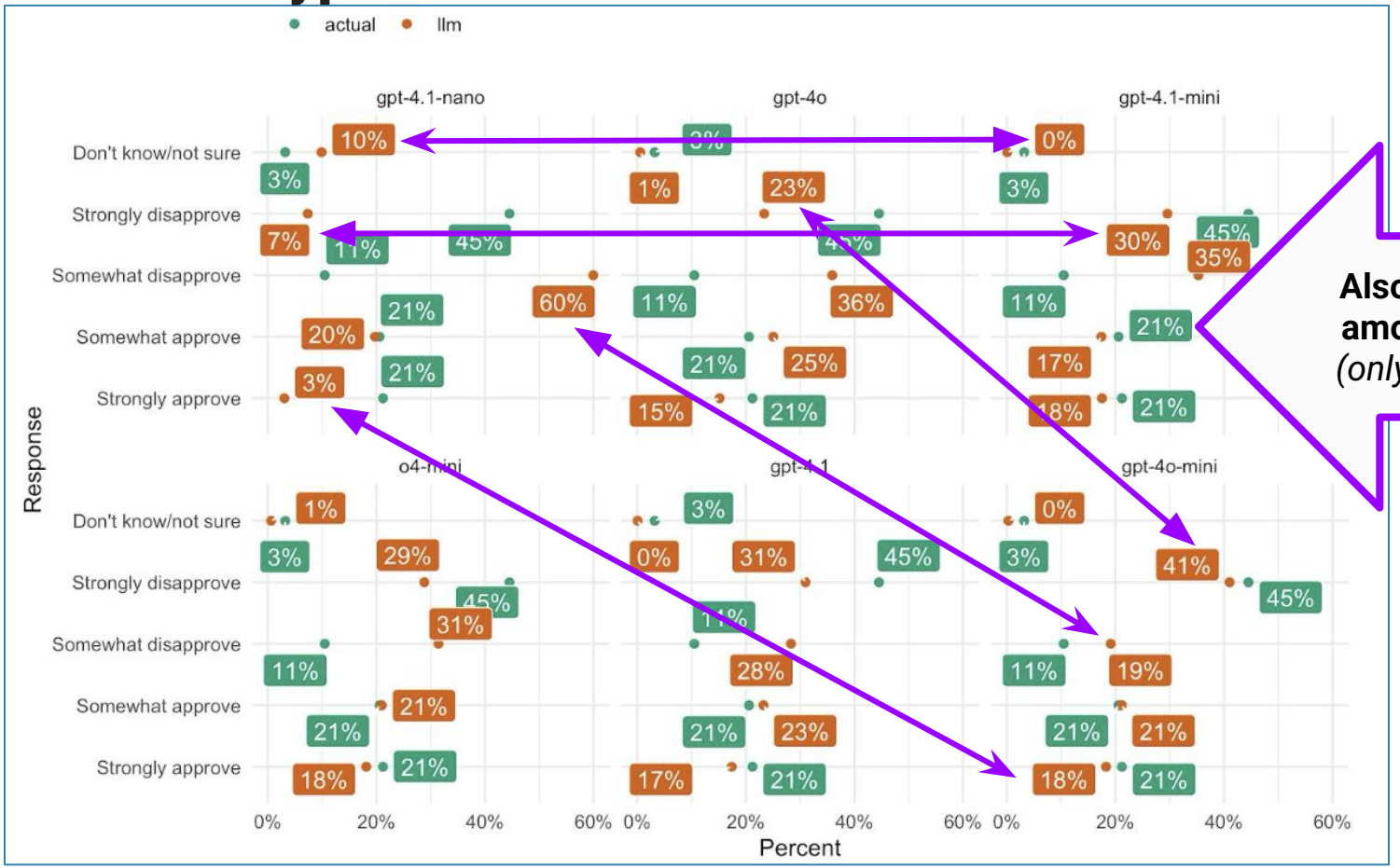


Most *incorrect* on negatives
(the most important points)

Legend
- **Actual human data**
- **LLM model results**

Morris & Verasight Data Team (2025), *Your Polls on ChatGPT*. At https://report.verasight.io/synthetic-sampling

Some typical results

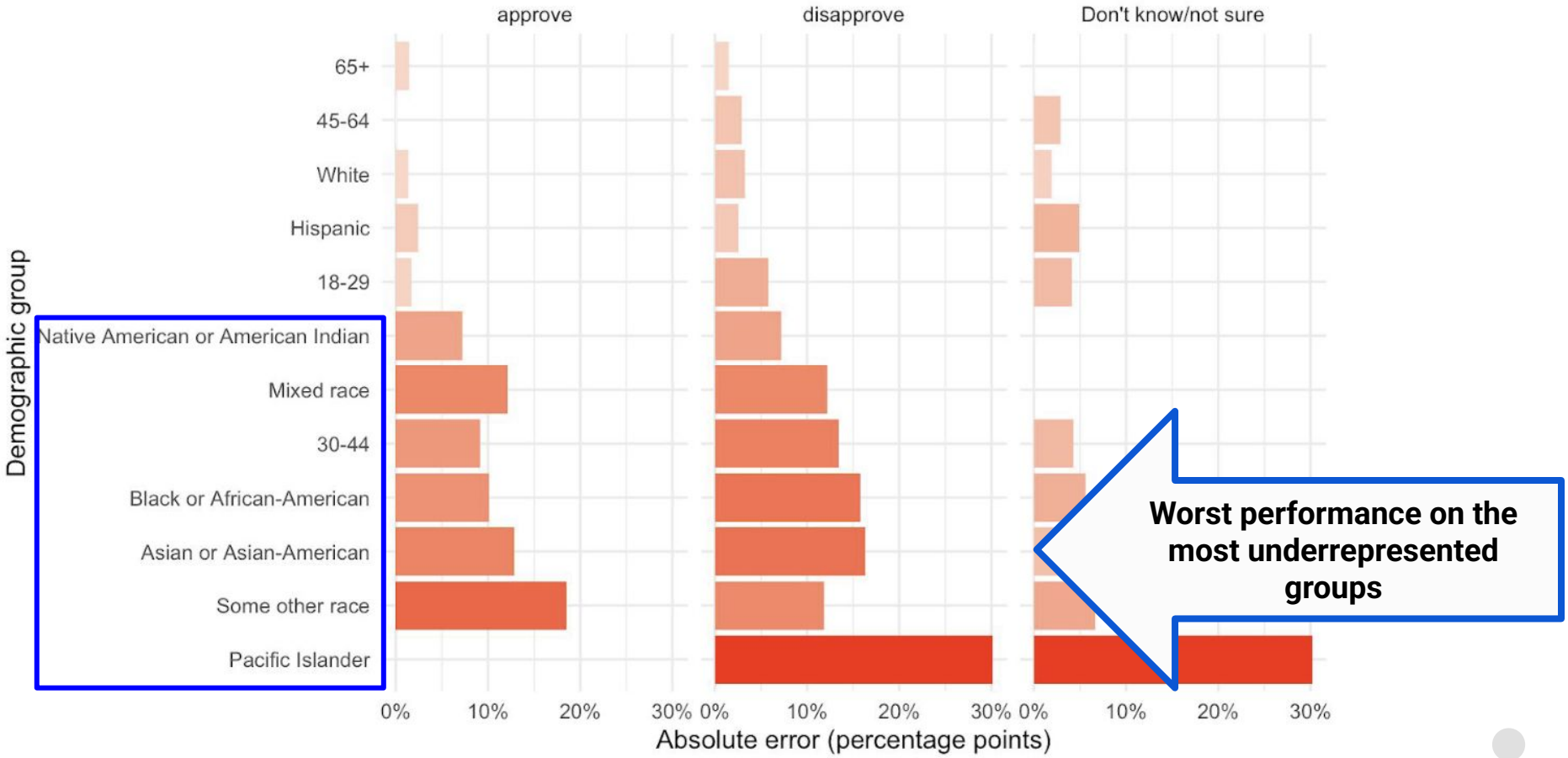


Also high disagreement among the LLM models
(only a few are indicated)

Legend
- **Actual human data**
- **LLM model results**

Morris & Verasight Data Team (2025), *Your Polls on ChatGPT*. At https://report.verasight.io/synthetic-sampling

Some typical results



Morris & Verasight Data Team (2025), *Your Polls on ChatGPT*. At https://report.verasight.io/synthetic-sampling

Brief discussion

- Those data concern perhaps the ***most polled, most written-about, most prominent, most sampled, most trained-upon*** survey topic in the world — presidential ratings in the US

- If an LLM can't perform well there, *why would we expect it to perform well for surveys when*:

  The topic is novel
  There is no prior training data
  There is no prior art on the questions to ask

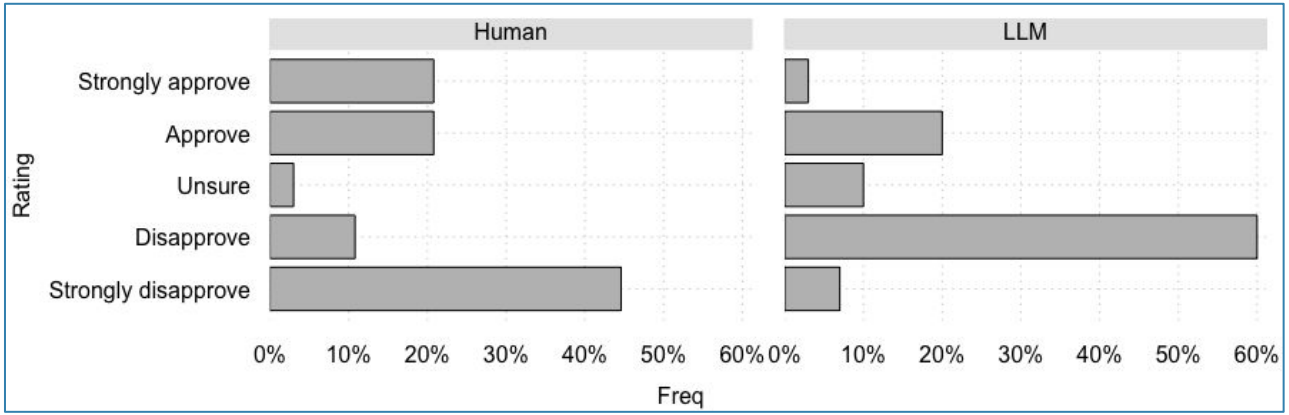
?

This example: Morris & Verasight Data Team (2025), *Your Polls on ChatGPT*. At https://report.verasight.io/synthetic-sampling

… let’s continue the main story …

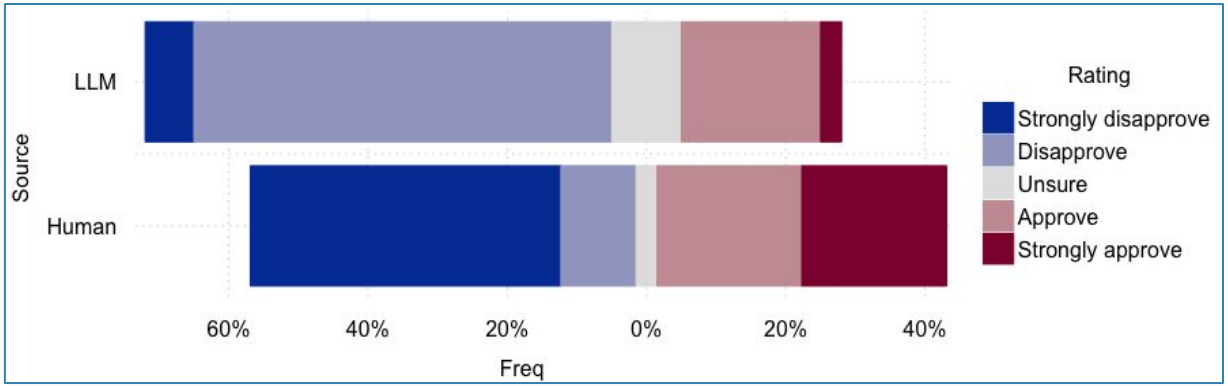
Some other typical analyses

- Many analysts might do something like these:



***Distributions not only unalike but* inverted!**

***Net over/under is wrong by* 30 points!**

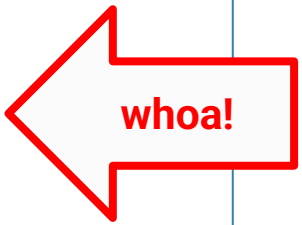


See R code appendix for the code!

Some other typical analyses

- *And analysts often do a statistical test like this:*

```
> N <- 1500
> round(morris.prop * N, 2)
                       Human LLM
Strongly disapprove      669 105
Disapprove               162 900
Unsure                    45 150
Approve                  312 300
Strongly approve         312  45
>
> # chisquare test
> chisq.test(morris.prop * N)

        Pearson's Chi-squared test

data:  morris.prop * N
X-squared = 1180.3, df = 4, p-value < 2.2e-16
```



See R code appendix for the code!

But what is this really saying?

```
        Pearson's Chi-squared test

data:  morris.tab
X-squared = 1183.2, df = 4, p-value < 2.2e-16
```

But what is this really saying?

```
        Pearson's Chi-squared test

data:  morris.tab
X-squared = 1183.2, df = 4, p-value < 2.2e-16
```

It says (somewhat paraphrasing stats language):

*"Assuming these are **random samples**,*

*with a **null hypothesis of no difference**,*

*then the Human and LLM **data distributions are different.**"*


Oh
nooooooo
!

Uh oh …

Assuming these are random samples,

with a null hypothesis of no difference,

then the Human and LLM data distributions are different.

So, which part is* random sampling *for LLM synthetic data ?

- The **corpus** of training data
- The selected human data **trainers**
- The resulting trained **model**
- The choice of **LLM provider**
- The current LLM **algorithm**
- The **prompt** you're using
- Your **human data** for digital twins
- Your code to construct **digital twins**
- The **data points** you include for twins
- The autocorrelated *(?)* **sequential queries**

At this point, perhaps you're thinking …

Yeah, but
no sample is ever "random"

At this point, perhaps you're thinking …

Yeah, but
no sample is ever "random"

First, that ***doesn't logically imply anything*** *about LLMs*

… and it sets up a main point of Section 3. Stay tuned!

… again let’s continue the story …

Uh oh uh oh …

Assuming these are random samples,

with a null hypothesis of no difference,

then the Human and LLM data distributions are different.

***Why should we assume* no difference?**

- Different **sources**, machine vs. human
- **Queried** in different ways
- Using different survey **formats**
- At different **times**
- Based on different sets of **information**
- Using different **systems** to respond
- With differing **motivations**
- … and **shown** in many studies to differ

Uh oh uh oh …

***Why should we assume* no difference?**

- Different **sources**, machine vs. human
- **Queried** in different ways
- Using different survey **formats**
- At different **times**
- Based on different sets of **information**
- Using different **systems** to respond
- With differing **motivations**
- … and **shown** in many studies to differ

⇒ The correct assumption is "different!"

Assuming these are random samples,

with a null hypothesis of no difference,

then the Human and LLM data distributions are different.

Uh oh uh oh …

***Why should we assume* no difference?**

- Different **sources**, machine vs. human
- **Queried** in different ways
- Using different survey **formats**
- At different **times**
- Based on different sets of **information**
- Using different **systems** to respond
- With differing **motivations**
- … and **shown** in many studies to differ

⇒ *The correct assumption is "**different**!"*

So — unfortunately — *statistical hypothesis testing is uninterpretable* (or if you prefer, "invalid")

Assuming these are random samples,

with a null hypothesis of no difference,

then the Human and LLM data distributions are different.

When two sets of things have wildly different data generating processes …

… it is nonsense to compare them statistically

Presumptively **Similar**

(and comparison makes sense)

Images from Google searches for non-licensed, openly usable images

Presumptively Similar
(and comparison makes sense)

Presumptively Different
(and comparison is nonsense)

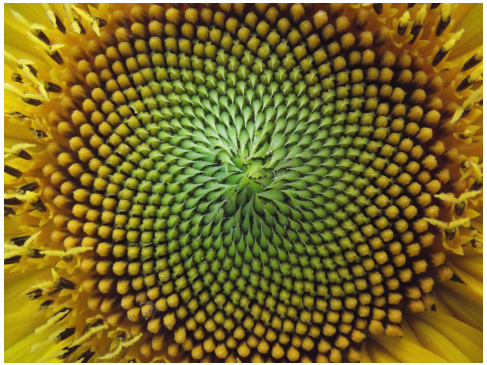
Images from Google searches for non-licensed, openly usable images

Presumptively Similar
(and comparison makes sense)

Presumptively Different
(and comparison is nonsense)

Images from Google searches for non-licensed, openly usable images

Presumptively Similar
(and comparison makes sense)



Images from Google searches for non-licensed, openly usable images

Presumptively Different
(and comparison is nonsense)

```
> fib <- function(x = 0, y = 1, stopval = 5000000) {
+   s <- x + y
+   c(s, if (s + y < stopval) fib(y, s))
+ }
> fib()
 [1]       1       2       3       5       8      13
 [7]      21      34      55      89     144     233
[13]     377     610     987    1597    2584    4181
[19]    6765   10946   17711   28657   46368   75025
[25]  121393  196418  317811  514229  832040 1346269
[31] 2178309 3524578
```

Presumptively Similar
(and comparison makes sense)

Presumptively Different
(and comparison is nonsense)

Images from Google searches for non-licensed, openly usable images

Presumptively **Similar**
(and comparison makes sense)

Presumptively **Different**
(but how would we know?)

Images from Google searches for non-licensed, openly usable images

Presumptively Different

Path 1

- Identify your question
- Ask people

vs.

Path 2

- Compile 10B digital docs
- Build an LLM model
- Hire people to train it
- Create a way to query it
- Build 100s of data centers

- Collect some human data
- Use data as “digital twins”
- Write API code for the LLM
- Identify your question
- Ask the LLM for each “twin”

To recap so far …

- Humans and LLMs …

  … have **incomparable data generating processes**
  … and **different sampling** mechanisms and concepts
  ... with a presumptively **false null hypothesis**

To recap so far …

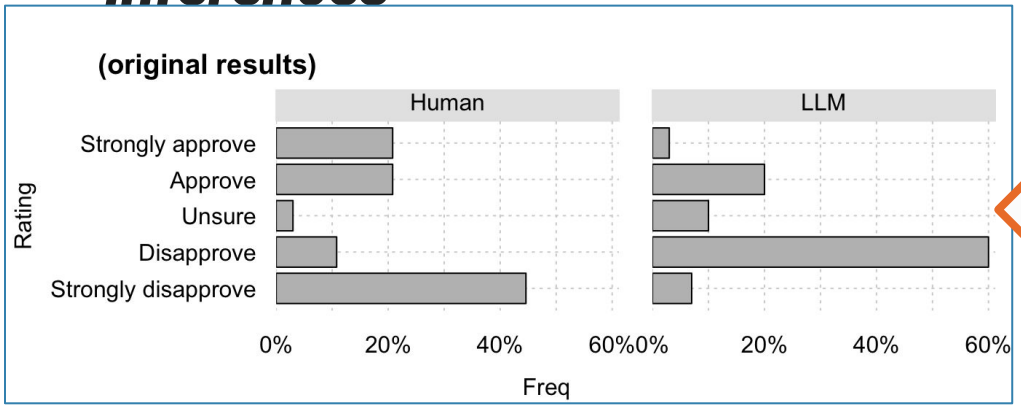
- Humans and LLMs …

  … have **incomparable data generating processes**
  … and **different sampling** mechanisms and concepts
  ... with a presumptively **false null hypothesis**

- **QED, statistical inferential testing between human and LLM samples is uninterpretable**

  ⇒ *One can do **math** on such data, but not **statistical inference***

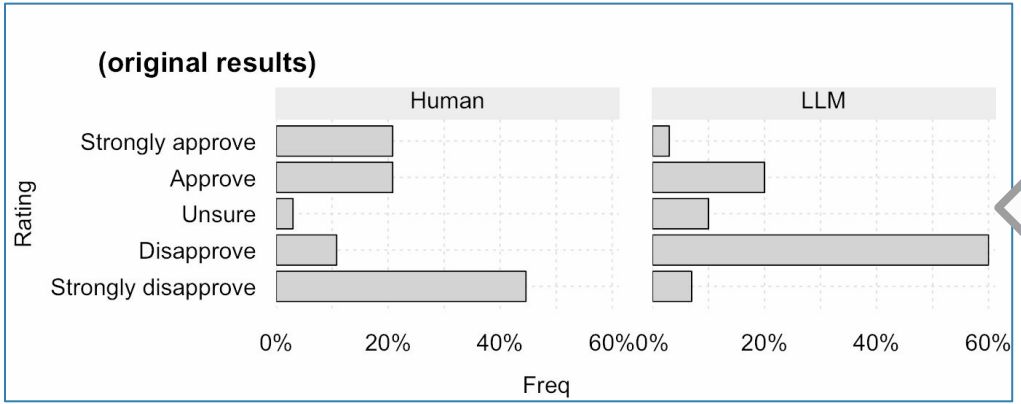
To be clear, the conclusion applies to all human-LLM inferences



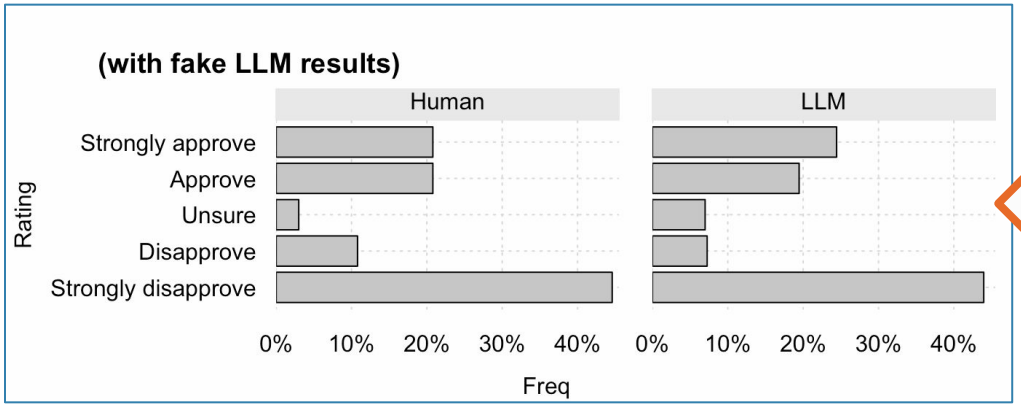
The **difference is N=1**, **uninterpretable**, and **doesn't generalize** to other samples / problems

… "But we've seen evidence that the results are similar!"

What if the results had been highly comparable?

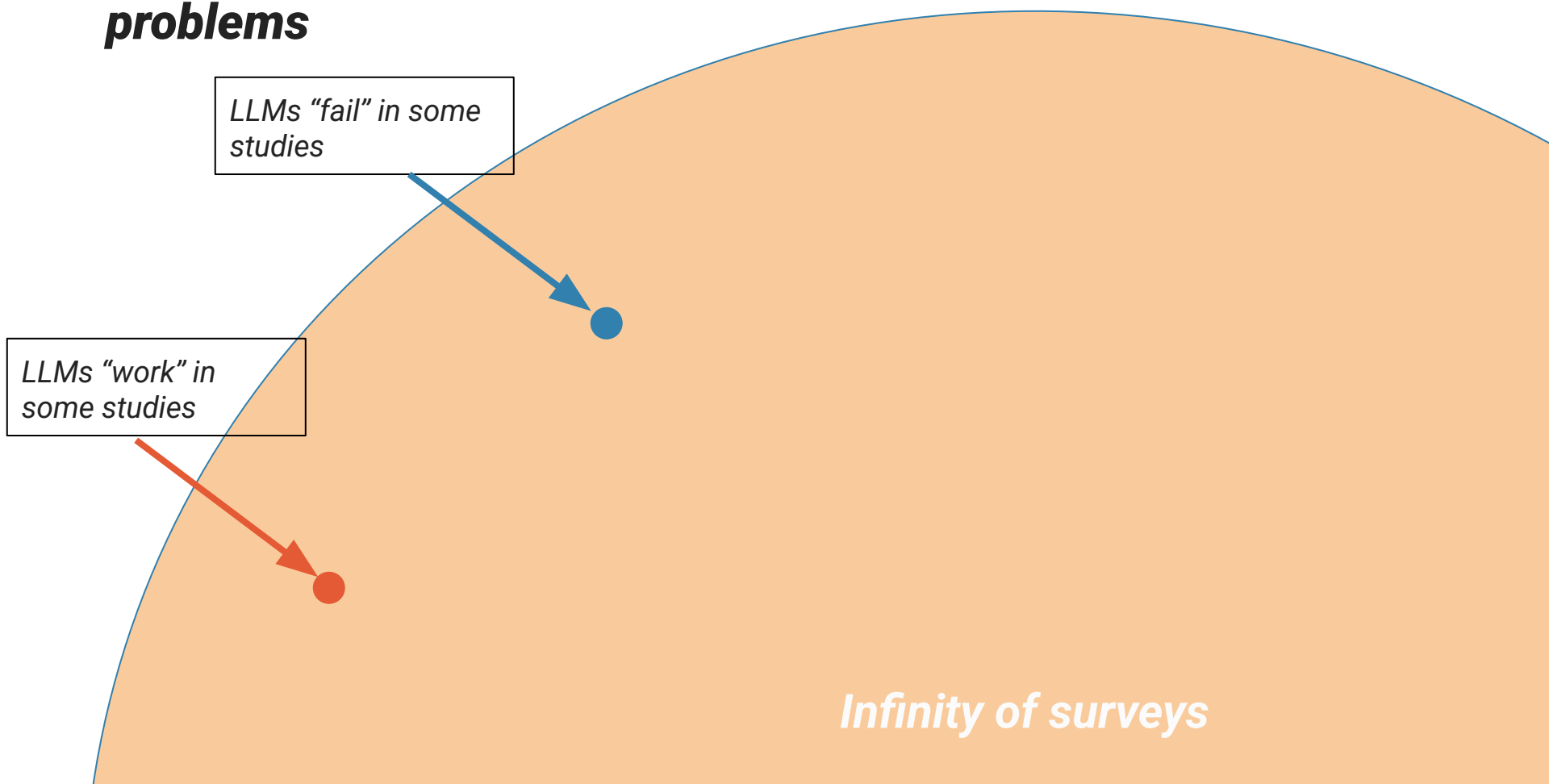


The **"difference" is N=1**, is **uninterpretable**, and **does not generalize** to other samples or problems



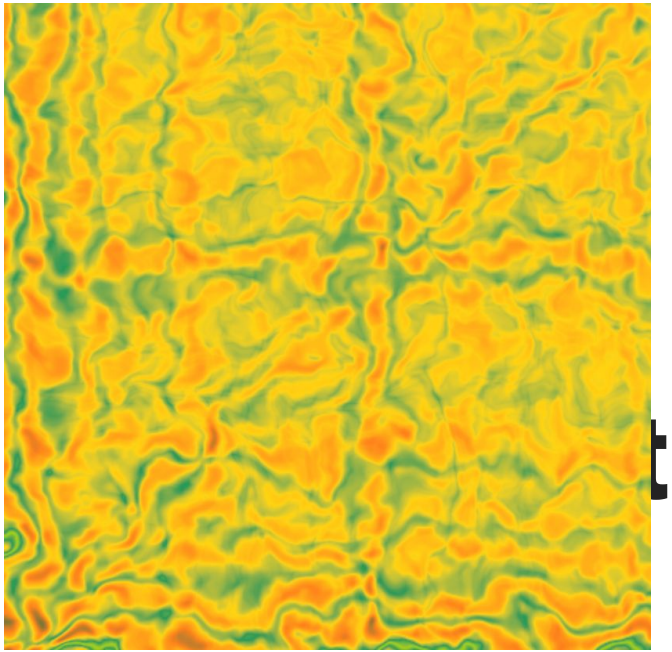
"Equivalence" is also N=1, uninterpretable, and **doesn't generalize** to other samples / problems

- **There is no way to characterize or sample the space of problems**



Realistically, we have no idea of the structure of "survey space" in an LLM.

Failure [white] and Success [green] might be completely unpredictable for any particular problem

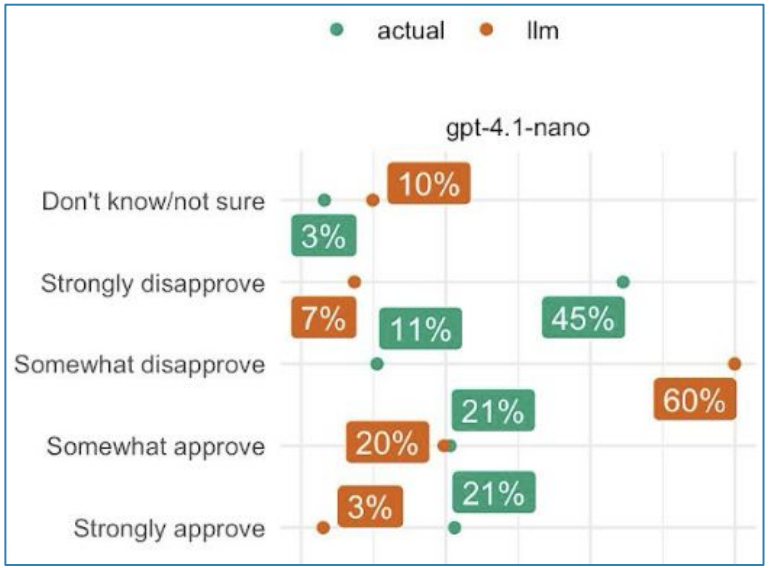


03

It misunderstands the point of surveys

A digression into survey “error”

The “error”



We can compute the MAE for such data (example, 1 model):

[(10-3) + (45-7) + (60-11) + (21-20) + (21-3)] / 5

== 113 / 5

== **22.6 point difference in the average rating**

(shockingly high for polling data!)

Morris & Verasight Data Team (2025), *Your Polls on ChatGPT*. At https://report.verasight.io/synthetic-sampling

Let’s turn that around for a moment

You might think, “*We only need directional evidence!*” So …

Let’s turn that around for a moment

You might think, “*We only need directional evidence!*” So …

Q: Do you care whether an answer is off by 22 points or not?

Let’s turn that around for a moment

Q: Do you care whether an answer is off by 22 points or not?
⇒ A: No, we only need a direction!

Let's turn that around for a moment

Q: Do you care whether an answer is off by 22 points or not?
⇒ A: No, we only need a direction!

Suppose I say:

Our new product will interest 1B people!
… but if we're off by 22% of the the world population, ***it might be 0.***

Let's turn that around for a moment

Q: Do you care whether an answer is off by 22 points or not?
⇒ A: No, we only need a direction!

Suppose I say:

Our new product will interest 1B people!
… but if we're off by 22% of the the world population, ***it might be 0.***

— OR —

Our new product interests 0 people!
… but ***it might interest 1.8 Billion*** *people.*

Let's turn that around for a moment

Q: Do you care whether an answer is off by 22 points or not?
⇒ A: No, we only need a direction!

Suppose I say:

Our new product will interest 1B people!
… but if we're off by 22% of the the world population, ***it might be 0.***

— OR —

Our new product interests 0 people!
… but ***it might interest 1.8 Billion*** *people.*

Not actionable & kills credibility!

So …

Q: Do you care whether an answer is off by 22 points or not?
A: No, we only need a direction!

⇒ If you don’t care about an error of 22 points in a survey, then:

So …

Q: Do you care whether an answer is off by 22 points or not?
A: No, we only need a direction!

⇒ If you don't care about an error of 22 points in a survey, then:

- Your question is **insufficiently focused** *AND / OR*
- You **don't have a need** for research *AND / OR*
- You can **answer it through small-scale research** …
  which will give a better, more nuanced, and deeper answer anyway!

Yet all of this still relies on a misleading concept:

“a latent true score”

Three Truths about Survey Responses

- Respondents take surveys **for their own reasons** *(*they have *motivated responses)*
- **Responses differ profoundly** based on circumstances, method, and wording
- Respondents **do not owe us** any “true” response as *we* conceive it

Truth #1 : Surveys are Motivated Responses

Respondents take surveys because they …

- .. are **interested** in the topic
- … **want to help** the research or its sponsor
- … are **incentivized** with pocket money, points, or the like
- … are survey **researchers** and like to see other surveys
- … want to **complain** about a product or topic
- … are **bored**
- … wish to **sabotage** the research (maybe for some ethical reason)
- … want to **earn money** to take care of themselves or their families

All of those are legitimate! *Yet none is part of any "true score"*

Truth #2: Responses vary tremendously by wording etc

(... I won't belabor that point because we all know it ...)

However, I will add this corollary:

The mapping of **[circumstance + wording + motive] ⇒ [response]**

... ***cannot be presumed to be identical for LLMs and Humans***

Truth #3: Respondents do not owe us a “true response”

- First of all, if there is no “true” response because of motivated responses, then how could anyone owe us a “true” response?

Truth #3: Respondents do not owe us a "true response"

- First of all, if there is no "true" response because of motivated responses, then how could anyone owe us a "true" response?

- Second, if

  [motivation + circumstance] ⇒ [response]

  then ***a respondent's*** "true response" doesn't map to ***our*** "true response"

- *BTW, taking surveys for LOLZ is perfectly OK as far as I'm concerned. People are people! My job is to motivate them …*

The opposite of “true score” is not a “false score” !

“True scores” concept

- In people’s heads
- Unchanging value plus error
- Measurable on a scale
- Exact

The opposite of “true score” is not a “false score” !

“True scores” concept	Actual responses:
In people’s heads	Expressed on survey
Unchanging value plus error	Vary with context
Measurable on a scale	Imperfectly mapped to scale
Exact	Fuzzy, flexible

Useful assumptions to model tendencies in data (e.g., factors)

"True scores" concept	Actual responses
In people's heads	Expressed on survey
Unchanging value plus error	Vary with context
Measurable on a scale	Scalar value is imperfect
Exact	Fuzzy, flexible

Useful assumptions to model tendencies in data (e.g., factors)

"True scores" concept

- In people's heads
- Unchanging value plus error
- Measurable on a scale
- Exact

But don't confuse that with reality!

(they align, but only imperfectly)

Actual responses

- Expressed on survey
- Vary with context
- Imperfectly mapped to scale
- Fuzzy, flexible

Bottom Line: “True scores” are useful statistical metaphors

- The concept of a latent true score is ***helpful*** for things like:

  Reliability assessment
  Factor analysis
  Item response theory
  Change modeling

- We shouldn’t confound ***helpfulness*** with any deep ***truth***

Remember George Box: “all models are wrong but some are useful”

So, are surveys useless?

⇒ Design Surveys *for* Motivated Communication

Motivated communication is not new in the human experience!

⇒ Design Surveys *for* Motivated Communication

Motivated communication is not new in the human experience!

Our job when writing surveys is to design surveys that:

Answer a business decision
by engaging relevant respondents
in real time
using questionnaire methods & wording
that align well with respondent motivations
and maximize the odds of a useful answer to our business

⇒ Design Surveys *for* Motivated Communication

Motivated communication is not new in the human experience!

Our job when writing surveys is to design surveys that:

Answer a business decision
By engaging relevant respondents
In real time
Using questionnaire methods & wording
That align well with respondent motivations
And maximize the odds of a useful answer to our business

Now let's take a look at how those map to LLM synthetic data …

Motivated Design vs. LLMs

Our job when writing surveys is to design surveys that:

Answer a business decision ⇐ ***Extraneous to LLMs***
By learning from people
In real time
Using questionnaire methods & wording
That align well with respondent motivations
And maximize the odds of a useful answer
to our business

Motivated Design vs. LLMs

Our job when writing surveys is to design surveys that:

Answer a business decision ⇐ ***Extraneous to LLMs***
By learning from people ⇐ ***Not what LLMs do***
In real time
Using questionnaire methods & wording
That align well with respondent motivations
And maximize the odds of a useful answer
to our business

Motivated Design vs. LLMs

Our job when writing surveys is to design surveys that:

Answer a business decision	⇐ ***Extraneous to LLMs***
By learning from people	⇐ ***Not what LLMs do***
In real time	⇐ ***Not possible for LLMs***
Using questionnaire methods & wording	
That align well with respondent motivations	
And maximize the odds of a useful answer to our business	

Motivated Design vs. LLMs

Our job when writing surveys is to design surveys that:

Answer a business decision	⇐ ***Extraneous to LLMs***
By learning from people	⇐ ***Not what LLMs do***
In real time	⇐ ***Not possible for LLMs***
Using questionnaire methods & wording	⇐ ***LLMs != Human perception***
That align well with respondent motivations	
And maximize the odds of a useful answer to our business	

Motivated Design vs. LLMs

Our job when writing surveys is to design surveys that:

Answer a business decision	⇐ ***Extraneous to LLMs***
By learning from people	⇐ ***Not what LLMs do***
In real time	⇐ ***Not possible for LLMs***
Using questionnaire methods & wording	⇐ ***LLMs != Human perception***
That align well with respondent motivations	⇐ ***LLMs != Human motives***
And maximize the odds of a useful answer to our business	

Motivated Design vs. LLMs

Our job when writing surveys is to design surveys that:

Answer a business decision	⇐ ***Extraneous to LLMs***
By learning from people	⇐ ***Not what LLMs do***
In real time	⇐ ***Not possible for LLMs***
Using questionnaire methods & wording	⇐ ***LLMs != Human perception***
That align well with respondent motivations	⇐ ***LLMs != Human motives***
And maximize the odds of a useful answer to our business	⇐ ***Extraneous to LLMs***

Motivated Design vs. LLMs

Our job when writing surveys is to design surveys that:

Answer a business decision	⇐ ***Extraneous to LLMs***
By learning from people	⇐ ***Not what LLMs do***
In real time	⇐ ***Not possible for LLMs***
Using questionnaire methods & wording	⇐ ***LLMs != Human perception***
That align well with respondent motivations	⇐ ***LLMs != Human motives***
And maximize the odds of a useful answer to our business	⇐ ***Extraneous to LLMs***

Takeaway: there is no survey answer that LLMs can generate for us

Wait, what was that about *random sampling* of people?

Wait, what was that about *random sampling* of people?

Our goal is to learn from people

Not to meet a statistical definition of "random"
(such assumptions are only helpful simplifications)

Wait, what was that about random sampling of people?

Our goal is to learn from people

Not to meet a statistical definition of "random"!
(such assumptions are only helpful simplifications)

The question we need to ask for research is:

Have we reached a large enough and diverse set of people to feel comfortable about what we learned?

LLMs don't reach people and are biased … so they never meet a reasonable bar to learn from people.

Wait, what was that about random sampling of people?

Our goal is to learn from people

Not to meet a statistical definition of "random"!

The question we need to ask for research is:
Have we reached a large enough and diverse set of people to feel comfortable about what we learned?

LLMs don't reach people and are biased … so they never meet a reasonable bar to learn from people.

Conversely, a "random sample" is not necessarily a good sample. We always use judgment.

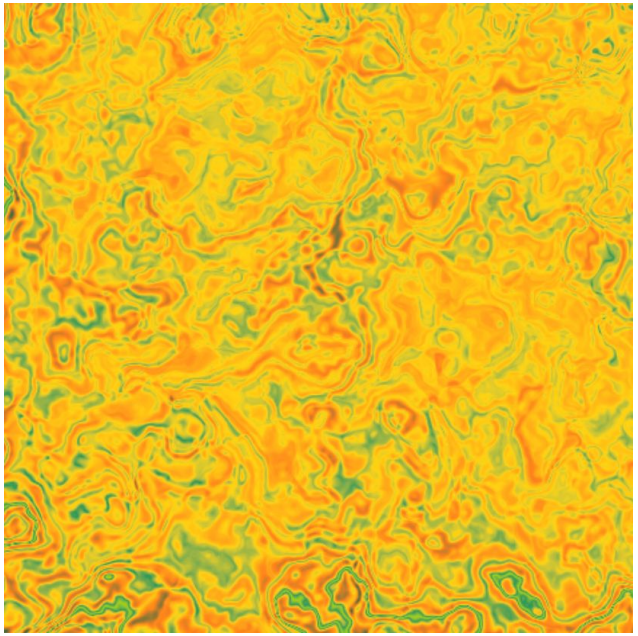
Design Surveys *FOR* Motivated Communication

⇒ *Instead of hoping AI will magically work, rely on what we know:*

Survey science + Pre-test + Iterate + Pair qual & quant

04

“Rebuttals” & Counter-rebuttals



Some claims and rebuttals

Claim	*Rebuttal*
Synthetic data can **accelerate research** by getting quick answers or more sample	
Use it to get **better priors** for Bayesian models	
Don’t use the data but use it to **pre-test** your survey	
Don’t use the data but use to **preview** potential results for discussion	
Technology **always gets better**	
Other colleagues are using it and **I’ll be left behind**	
I’m **ordered to use it** by my job	

Some claims and rebuttals

Claim	*Rebuttal*
Synthetic data can **accelerate research** by getting quick answers or more sample	**Do you care if the answers are correct?**
Use it to get **better priors** for Bayesian models	
Don’t use the data but use it to **pre-test** your survey	
Don’t use the data but use to **preview** potential results for discussion	
Technology **always gets better**	
Other colleagues are using it and **I’ll be left behind**	
I’m **ordered to use it** by my job	

Some claims and rebuttals

Claim	***Rebuttal***
Synthetic data can **accelerate research** by getting quick answers or more sample	Do you care if the answers are correct?
Use it to get **better priors** for Bayesian models	**This confuses *information* (priors) with *math*.**
Don't use the data but use it to **pre-test** your survey	
Don't use the data but use to **preview** potential results for discussion	
Technology **always gets better**	
Other colleagues are using it and **I'll be left behind**	
I'm **ordered to use it** by my job	

Some claims and rebuttals

Claim	***Rebuttal***
Synthetic data can **accelerate research** by getting quick answers or more sample	Do you care if the answers are correct?
Use it to get **better priors** for Bayesian models	This confuses *information* (priors) with *math*.
Don't use the data but use it to **pre-test** your survey	**It's better to pre-test using random responses because they are unbiased.**
Don't use the data but use to **preview** potential results for discussion	
Technology **always gets better**	
Other colleagues are using it and **I'll be left behind**	
I'm **ordered to use it** by my job	

Some claims and rebuttals

Claim	***Rebuttal***
Synthetic data can **accelerate research** by getting quick answers or more sample	Do you care if the answers are correct?
Use it to get **better priors** for Bayesian models	This confuses *information* (priors) with *math*.
Don't use the data but use it to **pre-test** your survey	It's better to pre-test using random responses; they are unbiased.
Don't use the data but use to **preview** potential results for discussion	**It's better to construct potential results with specific knowledge of the business**
Technology **always gets better**	
Other colleagues are using it and **I'll be left behind**	
I'm **ordered to use it** by my job	

Some claims and rebuttals

Claim	***Rebuttal***
Synthetic data can **accelerate research** by getting quick answers or more sample	Do you care if the answers are correct?
Use it to get **better priors** for Bayesian models	This confuses *information* (priors) with *math*.
Don't use the data but use it to **pre-test** your survey	It's better to pre-test using random responses; they are unbiased.
Don't use the data but use to **preview** potential results for discussion	It's better to construct potential results with specific knowledge of the business
Technology **always gets better**	**Does it? Or is that a form of survivorship bias?**
Other colleagues are using it and **I'll be left behind**	
I'm **ordered to use it** by my job	

Some claims and rebuttals

Claim	***Rebuttal***
Synthetic data can **accelerate research** by getting quick answers or more sample	Do you care if the answers are correct?
Use it to get **better priors** for Bayesian models	This confuses *information* (priors) with *math*.
Don't use the data but use it to **pre-test** your survey	It's better to pre-test using random responses; they are unbiased.
Don't use the data but use to **preview** potential results for discussion	It's better to construct potential results with specific knowledge of the business
Technology **always gets better**	Does it? Or is that a form of survivorship bias?
Other colleagues are using it and **I'll be left behind**	**Left behind in what way? Learn something useful!**
I'm **ordered to use it** by my job	

Some claims and rebuttals

Claim	***Rebuttal***
Synthetic data can **accelerate research** by getting quick answers or more sample	Do you care if the answers are correct?
Use it to get **better priors** for Bayesian models	This confuses *information* (priors) with *math*.
Don't use the data but use it to **pre-test** your survey	It's better to pre-test using random responses; they are unbiased.
Don't use the data but use to **preview** potential results for discussion	It's better to construct potential results with specific knowledge of the business
Technology **always gets better**	Does it? Or is that a form of survivorship bias?
Other colleagues are using it and **I'll be left behind**	Left behind in what way? Learn something useful!
I'm **ordered to use it** by my job	***Maybe it's time for a new job?***

Conclusion, Q&A
… and AMA

Key Points: if you remember nothing else

- Synthetic data has **no theoretical reason to work**
- Assume that **it is presumptively wrong** for interesting & actionable questions

Key Points: if you remember nothing else

- Synthetic data has **no theoretical reason to work**
- Assume that **it is presumptively wrong** for interesting & actionable questions

- Aligning with that, empirical results often show **vast differences** **from humans**

Key Points: if you remember nothing else

- Synthetic data has **no theoretical reason to work**
- Assume that **it is presumptively wrong** for interesting & actionable questions

- Aligning with that, empirical results often show **vast differences from humans**
- Even when empirical results show similarities:
  - There is **no expectation that any result will generalize**
  - Statistical comparisons of LLMs vs humans are **generally invalid**
  - **because LLMs do not sample from any definable population**

Key Points: if you remember nothing else

- Synthetic data has **no theoretical reason to work**
- Assume that **it is presumptively wrong** for interesting & actionable questions

- Aligning with that, empirical results often show **vast differences from humans**
- Even when empirical results show similarities:
  - There is **no expectation that any result will generalize**
  - Statistical comparisons of LLMs vs humans are **generally invalid**
  - **LLMs do not sample from any definable population**

- And it all rests on a **flawed & simplistic conception** of what surveys do

Key Points: if you remember nothing else

- Synthetic data has **no theoretical reason to work**
- Assume that **it is presumptively wrong** for interesting & actionable questions

- Aligning with that, empirical results often show **vast differences from humans**
- Even when empirical results show similarities:
  - There is **no expectation that any result will generalize**
  - Statistical comparisons of LLMs vs humans are **generally invalid**
  - **LLMs do not sample from any definable population**

- And it all rests on a **flawed & simplistic conception** of what surveys do

Instead …

⇒ **Learn from people** in **real time** to answer **crucial questions**

The whole talk in one slide

Generalizes to nothing

Compile 10B digital docs	Collect some human data
Build an LLM model	Use data as “digital twins”
Hire people to train it	Write API code for the LLM
Create a way to query it	Identify your question
Build 100s of data centers	Ask the LLM for each “twin”

The whole talk in one slide

Generalizes to nothing

- Compile 10B digital docs
- Build an LLM model
- Hire people to train it
- Create a way to query it
- Build 100s of data centers

- Collect some human data
- Use data as “digital twins”
- Write API code for the LLM
- Identify your question
- Ask the LLM for each “twin”

Generalizes to people

- Identify your question
- Ask people

References

More from smart people *(n.b., caveat re statistics!)*

Morris & Verasight Data Team (2025). Your Polls on ChatGPT. *[The empirical work I cited here; a very thorough example, using the world's probably most-polled topic.]*

Leff (2025). The Limits of Synthetic Data for Consumer & Public Opinion Research. *[Examines brand perception and whether adding "twin" attributes improves results from LLMs.]*

Samolyev (2024). Synthetic Respondents are the Homeopathy of Market Research. *[Includes R code you can use to test synthetic respondents for your own problem.]*

Paxton & Yang (2024). Do LLMs simulate human attitudes about technology products? *[Compares LLMs and people on sentiment ratings of products.]*

More from me

Chapman (2025). Synthetic Survey Data? It's Not Data. *[An only somewhat overlapping discussion of the same topic as this talk; has additional theory-of-science discussion.]*

Chapman (2025). Four areas of (UXR) thinking about AI / LLMs. *[Why focusing on* utility *— like this talk — misses crucial issues of aesthetics, ethics, and social power.]*

Chapman (2024). Surveys and the True Score Mistake. *[Why surveys are about learning from people in motivated situations, not about measuring any "true score".]*

Chapman (2024). "Research" Concerns for LLM Applications. [*Additional philosophy-of-research reasons that too many studies using LLMs are uninformative.]*

Join the Quant UX Con mailing list:

https://quantuxcon.org

Thank you! & AMA

chris@quantuxa.org

2)

```
# data comparison for synthetic survey talk
# chris chapman, september 2025
# based on Morris & Verasight, Your Polls on ChatGPT,
https://report.verasight.io/synthetic-sampling

rating.scale <- c("Strongly disapprove", "Disapprove", "Unsure", "Approve",
                  "Strongly approve")
human.prop   <- c(0.446, 0.108, 0.030, 0.208, 0.208)   # from paper above
#                    ^^ adjusted 3rd decimals for rounding error
llm.prop     <- c(0.070, 0.600, 0.100, 0.200, 0.030)

# create counts table
morris.prop <- matrix(c(human.prop, llm.prop), ncol=2, byrow=FALSE)
colnames(morris.prop) <- c("Human", "LLM")
rownames(morris.prop) <- rating.scale
morris.prop <- as.table(morris.prop)
N <- 1500
round(morris.prop * N, 2)

# chisquare test
chisq.test(morris.prop * N)

# plot
library(tinyplot)
# get proportions for plotting
morris.df <- as.data.frame(morris.prop)
names(morris.df) <- c("Rating", "Source", "Freq")

tinytheme("clean2")
tinyplot(Freq ~ Rating, facet = ~Source, data = morris.df,
        type = "barplot", beside = TRUE,
         flip = TRUE, facet.args = list(ncol = 2), yaxl = "percent")

tinytheme("clean2", palette.qualitative = "Blue-Red")
tinyplot(Freq ~ Source | Rating, data = morris.df, type = "barplot",
         center = TRUE, flip = TRUE, facet.args = list(ncol = 1),
        yaxl = "percent")
```

```
# fibonacci sequence
# based on M van Tilborg,
https://stackoverflow.com/questions/48566319/fibonacci-sequence-in-r
fib <- function(x = 0, y = 1, stopval = 5000000) {
  s <- x + y
  c(s, if (s + y < stopval) fib(y, s))
}

fib()

# make fake data set that would purport to be "similar" for Morris data
morris.fake <- morris.df
set.seed(98250)
morris.fake[6:10, "Freq"]   <- jitter(morris.fake[1:5, "Freq"],
factor=3)
tinytheme("clean2")
tinyplot(Freq ~ Rating, facet = ~Source, data = morris.fake,
         type = "barplot", beside = TRUE,
         flip = TRUE, facet.args = list(ncol = 2), yaxl = "percent",
         main="(with fake LLM results)")


# LLM “structure” slide for unpredictable neighboring areas
library(aRtsy)
set.seed(98199)
canvas_chladni(colors = colorPalette("nature")[c(2, 3)], warp=0.3,
        waves=c(5, 15, 5)*20)
```

Extras

Questions for LLM Synthetic Data Suppliers' Studies

1. **Did they pre-register the study? Will they pre-register one for you?** You pick question(s) and they commit to an expected level of accuracy *before* any prompts to their model.

2. **Has their model been *trained* on the same data they are trying to predict?** (Test contamination. See this blog post for a likely example of contamination.)

3. **Was their *model* designed specifically to *target the test* they gave it?** (See previous.)

4. **What kinds of problems will their model generalize to?** What will it *not* generalize to? How do they know those boundaries? Can they define them crisply?

5. **When they update the model, will it change these results?** Why or why not? How much?

6. **Is their model based on any particular, falsifiable theory?** Or is it instead a variation of a generic LLM? If it is specific, is that theory ad hoc or is it supported by strong theory?

7. **What test does it *fail* but are unpublished?** (File drawer problem.) What do those imply?

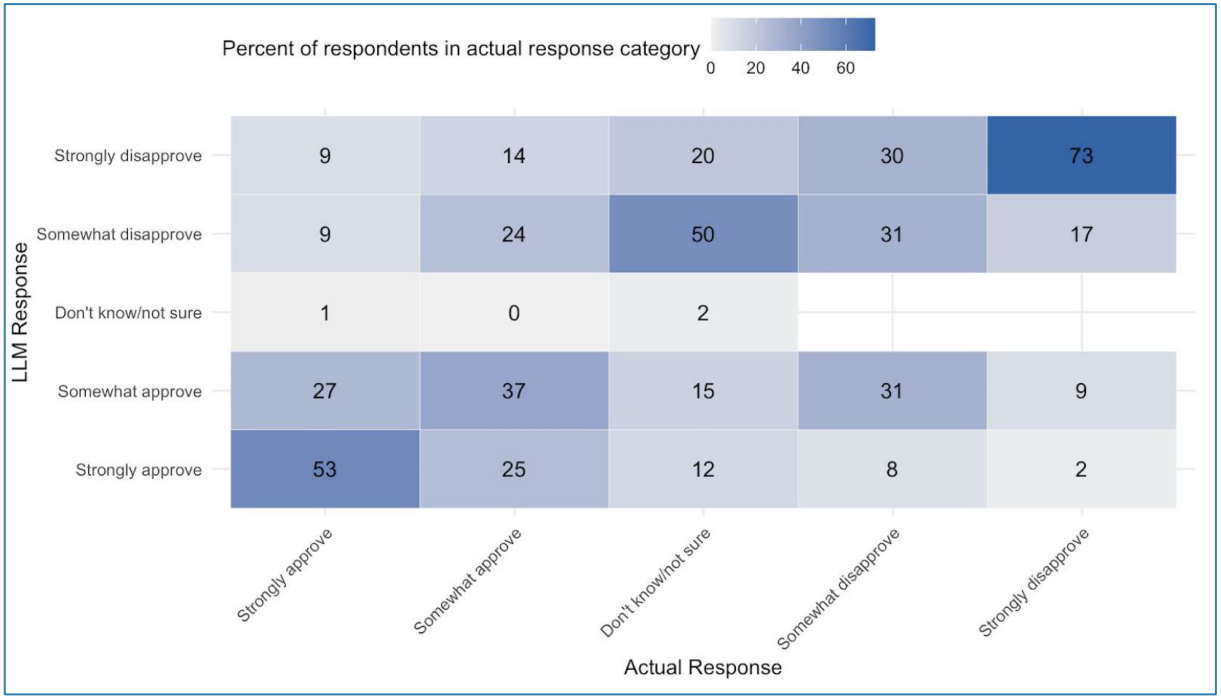
R code bonus

The progressively deteriorating data art that I used on the section slides was generated using the R **aRtsy** package:

```
library(aRtsy)

set.seed(98199); artwork <- canvas_chladni(colors = colorPalette("mixer4"), waves=15, warp=0.01)
saveCanvas(artwork, filename = "~/Downloads/wave-art1.png")

set.seed(98199); artwork <- canvas_chladni(colors = colorPalette("mixer4"), waves=15, warp=0.04)
saveCanvas(artwork, filename = "~/Downloads/wave-art2.png")

set.seed(98199); artwork <- canvas_chladni(colors = colorPalette("mixer4"), waves=15, warp=0.08)
saveCanvas(artwork, filename = "~/Downloads/wave-art3.png")

set.seed(98199); artwork <- canvas_chladni(colors = colorPalette("mixer4"), waves=15, warp=0.25)
saveCanvas(artwork, filename = "~/Downloads/wave-art4.png")
```

Extra: individual level comparison for Morris data

Better than group-level stats is to examine the confusion matrix at an individual level. ***Did the LLM "twins" agree with individuals' actual ratings?***



Morris & Verasight Data Team (2025), *Your Polls on ChatGPT*. At https://report.verasight.io/synthetic-sampling

Extra: individual level comparison for Morris data

Better than group-level stats is to examine the confusion matrix at an individual level. ***Did the LLM "twins" agree with individuals' actual ratings?***

```
> confusionMatrix(morris.conf)
Confusion Matrix and Statistics

                    Strongly disapprove Disapprove Unsure Approve Strongly approve
Strongly disapprove                  73         30     20      14                9
Disapprove                           17         31     50      24                9
Unsure                                0          0      2       0                1
Approve                               9         31     15      37               27
Strongly approve                      2          8     12      25               53

Overall Statistics

               Accuracy : 0.3928
                 95% CI : (0.3497, 0.4372)
    No Information Rate : 0.2024
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2404
```

Overall: only 39% accuracy, low Kappa

R code for confusion matrix

```
# data comparison for synthetic survey talk
# chris chapman, september 2025
# based on Morris & Verasight, Your Polls on ChatGPT, https://report.verasight.io/synthetic-sampling

## analyze confusion matrix
morris.conf <- matrix(c(  rev(c(9, 14, 20, 30, 73)),
                          rev(c(9, 24, 50, 31, 17)),
                          rev(c(1,  0,  2,  0,  0)),
                          rev(c(27, 37, 15, 31,  9)),
                          rev(c(53, 25, 12,  8,  2))), ncol=5, byrow=TRUE)

colnames(morris.conf) <- rating.scale
rownames(morris.conf) <- rating.scale
morris.conf <- as.table(morris.conf)
morris.conf

library(caret)
confusionMatrix(morris.conf)
```

tabled below here, misc

And, again with the math

Imagine we have a **CI** of +/- 22.6 points, with median 0.2 (⅕) binomial **proportion**? *What does that tell us about implicit sample size? **

CI = 1.96 x SE = 0.226

SE = sqrt(p * (1-p)/N) = CI / 1.96 = 0.226 / 1.96

*** this is a straw argument under the* disproven *assumption of random sampling by an LLM! But if you want to investigate, this is one way you might start. For example, you might simulate using gamma distributions for various parameters.*

And, again with the math

Imagine we have a **CI** of +/- 22.6 points, with median 0.2 (⅕) binomial **proportion**? *What does that tell us about implicit sample size?*

CI	=	1.96 x SE	=	**0.226**		
SE	=	sqrt(p * (1-p)/N)	=	CI / 1.96	=	**0.226** / 1.96
	=	**sqrt**(0.2 * 0.8 / N)			=	0.115
		(0.2 * 0.8 / **N**)	=	0.115 **^ 2**	=	0.0133
		(0.2 * 0.8)	=	**N** * 0.0133		
N	=	(0.2 * 0.8) / 0.0133	=	12	⇐	***implicit N=12 for the LLM ?***