

SI Talk

Können LLMs Probanden ersetzen?



Building Competence. Crossing Borders.

Prof. Dr. Steffen Mueller, Mai 2024

steffen.mueller@zhaw.ch

Einsatzgebiete von LLMs

- Large Language Models (LLMs) finden verschiedene Einsatzbereiche im Marketing – z.B. im Content Marketing durch Content Generation, im Customer Experience Management durch Hyperpersonalisierung – und auch in der Marktforschung
- Während der Fokus in der Marktforschung meist auf der Textanalyse liegt, gibt es auch erste Studien, die sich damit beschäftigen, ob LLMs Probanden ersetzen können
- Man spricht dann auch von «Silicon Samples» – «... that seek to mimic human respondents to describe, explain, and predict human behavior»
- Z.B. ermöglicht das Startup «Synthetic Users», Interviews mit Personas zu führen
- Es stellt sich jedoch die Frage, ob «Silicon Samples» valide Antworten liefern – «... since an LLM's working principle involves computing the most probable next text element in a sequence; this process differs considerably from a human participant's feelings and reasoning abilities»

Beispiel: Synthetic Users

The screenshot displays the Synthetic Users web application interface. At the top, there is a navigation bar with four interview types: Dynamic Interview (Provide the goal), Custom Interview (You are in control), Problem Exploration Interview (selected), and Solution Feedback Interview. On the right of the navigation bar, there is a control for the number of interviews (set to 3) and a 'Generate SUs' button. The main content area is titled 'Your study' and is divided into three columns: Audiences, Problems, and Solutions. The Audiences column contains one entry: 'Product manager in an early stage company'. The Problems column contains two entries: 'Not having time to schedule user interviews' and 'Not having budget to run user interviews'. The Solutions column contains one entry: 'Synthetic users powered by GPT-3 that provide feedback on product or features'. Each entry has a checkbox, a delete icon, and a more options icon. There are '+ Add' buttons at the bottom of each column. On the left side, there is a dark sidebar with navigation options: Home, All History, Recent studies, Send Feedback, Account Settings, and Workspace Settings. At the bottom left of the sidebar, the user's profile is shown as 'personal'. On the right side, there is a vertical 'Insights Panel' and a chat window from 'Hugo from Synthetic Users' with a message: 'Hey there 🤖 Welcome to Synthetic Users 🎉 Take a look around! If you...'. A notification icon with a red '1' is visible in the bottom right corner.

Erste Ergebnisse zu «Silicon Samples»

- Obwohl es eine sehr neue Anwendung ist, haben erste wissenschaftliche Studien «Silicon Samples» mit «Human Samples» verglichen
- Positiv schnitten Vergleiche zu Persönlichkeitsmerkmalen (Caron & Srivastava 2022), zu Framing-Effekten (Chen et al. 2023) und zu politischen Präferenzen (Argyle et al. 2023) ab
- Negativ schnitten Vergleiche zum Endowment-Effekt, zum Mental Accounting und zur Sunk-Cost-Fallacy ab (Chen et al. 2023)
- Auch von 14 Studien aus Many Labs 2 konnten nur etwa 1/3 repliziert werden (Park et al. 2023)
- Aber: Die Ergebnisse hängen stark von den Trainingsdaten ab – diese stellen oft keine geeignete Grundgesamtheit dar
- Und: GPT 3.5 und höher zeigen Chain-of-Thought-Reasoning, während frühere Versionen eher System-1-Denken aufweisen

Beispiel: System-1-Denken vs. Chain-of-Thought-Reasoning

Together, a potato and a camera cost \$1.40. The potato costs \$1 more than the camera.
How much does the camera cost?"

Aktuelle Studie von Brand, Israeli & Ngwe 2023

Nachfragekurve

Prompt

“A customer is randomly selected while shopping for laptops. Their annual income is $\$income$.

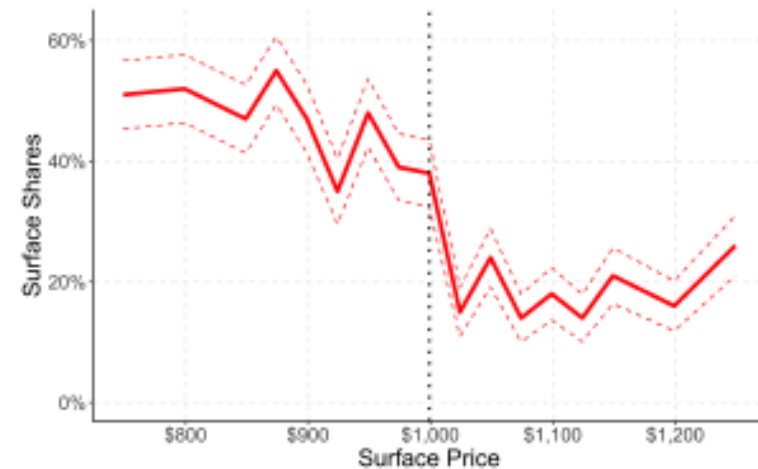
While shopping, the customer has three options:

- Surface Laptop 3, Price: $\$surfacePrice$, Processor: Intel Core i5, RAM: 8GB, Screen Size: 13.5in, SD: 128GB
- Macbook Air (2019), Price: \$999, Processor: Intel Core i5, RAM: 8GB, Screen Size: 13.3in, SD: 128GB

They also have the option not to purchase a laptop. The customer is asked, after they finish shopping: Which laptop, if any, did you purchase?

Customer: ”

Ergebnis



→ Wie anzunehmen, hat die Nachfragekurve eine negative Steigung

Aktuelle Studie von Brand, Israeli & Ngwe 2023

Nachfragekurve

Prompt

“A customer is randomly selected while shopping for laptops. Their annual income is $\$income$.

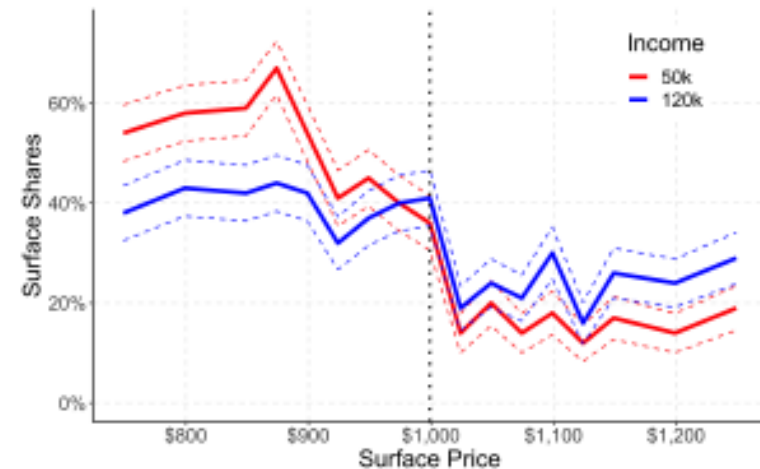
While shopping, the customer has three options:

- Surface Laptop 3, Price: $\$surfacePrice$, Processor: Intel Core i5, RAM: 8GB, Screen Size: 13.5in, SD: 128GB
- Macbook Air (2019), Price: \$999, Processor: Intel Core i5, RAM: 8GB, Screen Size: 13.3in, SD: 128GB

They also have the option not to purchase a laptop. The customer is asked, after they finish shopping: Which laptop, if any, did you purchase?

Customer: ”

Ergebnis



➔ Wie anzunehmen, wirkt sich das Einkommen auf die Nachfragekurve aus

Aktuelle Studie von Brand, Israeli & Ngwe 2023

Nachfragekurve

Prompt

"A customer is randomly selected while shopping in the supermarket. Their annual income is $\$income$.

While shopping, the customer passes by the toothpaste aisle and sees two options:

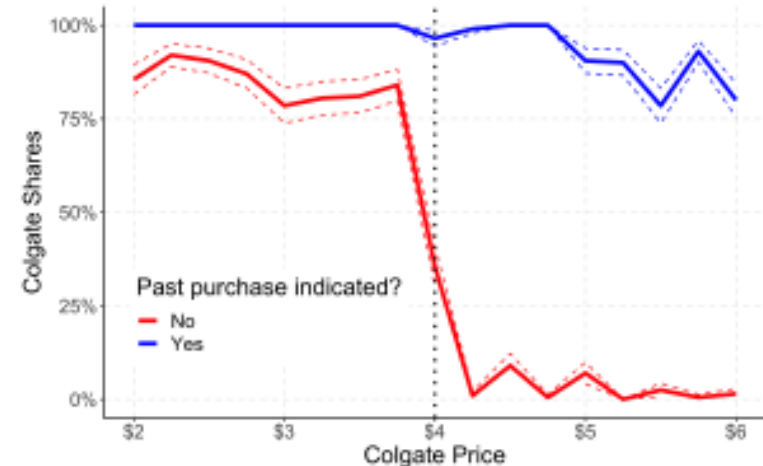
- Colgate whitening toothpaste with fluoride, price $\$colgatePrice$.
- Crest whitening toothpaste with fluoride, price $\$4$.

They also have the option not to purchase toothpaste. The customer says that last time they shopped for toothpaste they purchased the Colgate whitening toothpaste.

The customer is asked, after they finish shopping: which toothpaste, if any, did you purchase this time?

Customer: "

Ergebnis



→ Wie anzunehmen, wirkt sich der letzte Einkauf auf die Nachfragekurve aus

Aktuelle Studie von Brand, Israeli & Ngwe 2023

Zahlungsbereitschaft

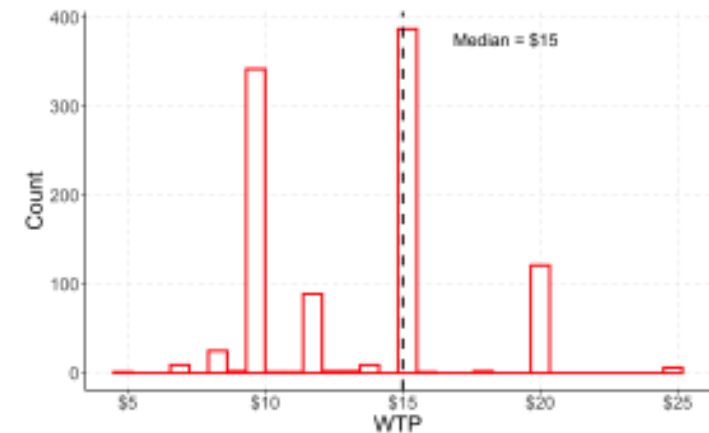
Prompt

“A customer is randomly selected while sitting at a restaurant. Their annual income is $\$income$.

The customer is asked: What is the maximum price you would be willing to pay for one glass of wine? please give a single price as your answer.

Customer: \$”

Ergebnis



Aktuelle Studie von Brand, Israeli & Ngwe 2023

Conjoint Measurement

Prompt

“A customer is randomly selected while shopping in the supermarket. Their annual income is $\$income$.

While shopping, the customer passes by the toothpaste aisle and sees two options:

- Colgate whitening toothpaste $colgateFluoride$ fluoride, price $\$colgatePrice$.
- Crest whitening toothpaste $crestFluoride$ fluoride, price $\$crestPrice$.

They also have the option not to purchase toothpaste. The customer is asked, after they finish shopping: Did you purchase any toothpaste? If so, which one?

Customer: ”

Ergebnis

Table 1: Toothpaste Conjoint: Choice Determinants

	$I(Choice=Colgate)$ (1)	$I(Choice=Crest)$ (2)
Crest price	0.120*** (0.004)	-0.081*** (0.005)
Colgate price	-0.080*** (0.004)	0.134*** (0.005)
$I(Crest\ fluoride)$	-0.199*** (0.007)	0.307*** (0.008)
$I(Colgate\ fluoride)$	0.251*** (0.007)	-0.210*** (0.008)
Constant	0.122*** (0.014)	0.142*** (0.015)
Observations	10,800	10,800

Significance level: 10% (*); 5% (**); 1% (***).

→ n=10'800, 35 Minuten, Kosten nahe Null

Was tun?

- «Silicon Samples» sind für Pretests und für Vorstudien besonders geeignet – z.B. lassen sich auch Stimuli mit DALL-E generieren

Beispiel: Pretests

You

Is the following survey item, which respondents should answer on a scale from 1 (I fully disagree) to 7 (I fully agree) appropriate? "I'm satisfied with the products and services of the company"

ChatGPT

The survey item "I'm satisfied with the products and services of the company," with a response scale from 1 (I fully disagree) to 7 (I fully agree), is a typical and generally appropriate way to measure overall satisfaction. However, there are a few points to consider for its optimal use:

1. **Combining Products and Services:** The item combines both 'products' and 'services' into one statement. This could be problematic if a respondent has different satisfaction levels for the products versus the services. If it's important to distinguish between these aspects, consider using separate items for each.
2. **Scale Clarity:** The 7-point scale provides a good range of options, allowing for nuanced responses. Ensure that the meaning of each point on the scale is clear to respondents.
3. **Subjective Nature:** The item is based on personal opinion, which is suitable for gauging satisfaction. Be aware that responses will be subjective and influenced by individual expectations and experiences.
4. **Breadth of the Question:** This item is quite broad, encompassing all products and services of a company. While useful for a general sense of satisfaction, it doesn't provide insight into specific areas of strength or improvement.
5. **Neutral Option:** The scale includes a midpoint (4), which typically serves as a neutral option. This allows respondents who neither agree nor disagree to express their stance.
6. **Consistency with Other Items:** If your survey includes similar items, maintain consistency in the scale and phrasing to ensure comparability of responses.

In summary, while the item is generally appropriate for measuring overall satisfaction, the combination of products and services into a single item might mask differing perceptions of each. Consider your specific research goals when deciding whether to use this item as is or to modify it for greater specificity.

You

Is the following survey item, which respondents should answer on a scale from 1 (I fully disagree) to 7 (I fully agree) appropriate? "I use this service very often"

ChatGPT

The survey item "I use this service very often," with a response scale from 1 (I fully disagree) to 7 (I fully agree), is generally appropriate but could benefit from some refinement for clarity and accuracy. Here are a few considerations:

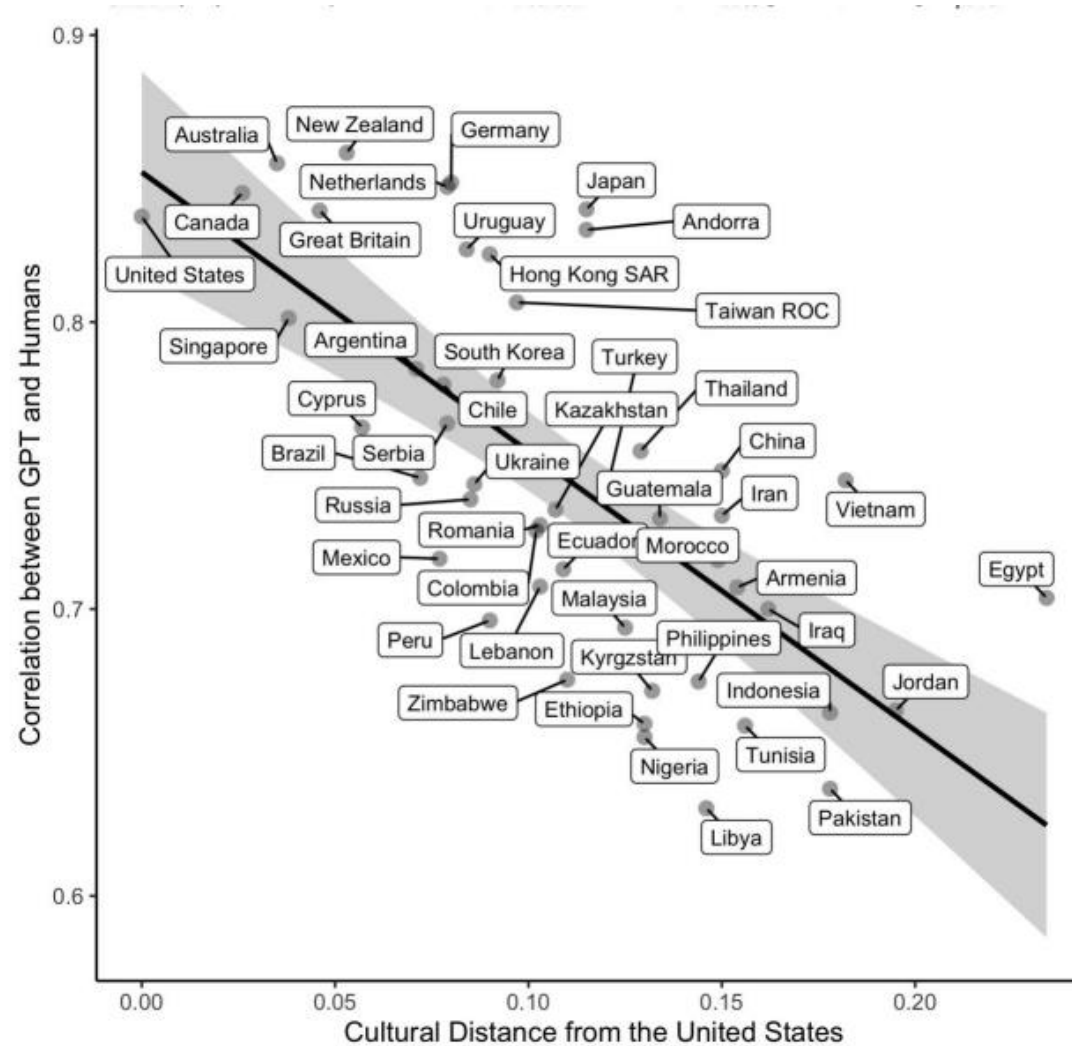
1. **Vagueness of "Very Often":** The term "very often" is subjective and can vary greatly in interpretation from one person to another. What one person considers "very often" might be considered less frequent by someone else. It would be more effective to define what "very often" means in this context or to use a more specific frequency scale.
2. **Frequency Scale Alternative:** Instead of an agree-disagree scale, a frequency scale might be more suitable for this question. For example, you could rephrase the question to "How often do you use this service?" with a scale like:
 - 1 = Never
 - 2 = Rarely
 - 3 = Occasionally
 - 4 = Sometimes
 - 5 = Often
 - 6 = Very Often
 - 7 = Always
3. **Contextual Relevance:** Make sure that the frequency of using the service is relevant to your survey's goals. Understanding the frequency of use can be important for gauging customer engagement, satisfaction, or identifying potential areas for improvement.
4. **Consistency with Survey Design:** If you're using a Likert scale for other items in the survey (e.g., measuring agreement or satisfaction), introducing a frequency scale for this item might disrupt the consistency. However, if this item's purpose is specifically to gauge usage frequency, the inconsistency might be justified.

In summary, while the original item is not inherently inappropriate, rephrasing it to be more specific and potentially using a frequency-based response scale could yield more precise and useful data.

Was tun? (Fortsetzung)

- Für Hauptstudien empfiehlt es sich ...
 - zu prüfen, ob die Trainingsdaten Antworten auf die Forschungsfrage liefern können – z.B. sind allgemeine Gründe für Unzufriedenheit bei Hotels aus Review-Daten verfügbar, aber eher nicht zu einem spezifischen Hotel
 - zu prüfen, ob die Grundgesamtheit in den Trainingsdaten abgedeckt ist – z.B. kann über den Prompt spezifischer eingeschränkt werden («act as»)
 - ggf. zusätzliche Trainingsdaten bereitzustellen
 - Prompts zu testen und anzupassen
 - Primär- oder Sekundärdaten als Vergleich heranzuziehen und Mittelwerte und Varianzen zu vergleichen
 - alle Inputs und Prompts zu dokumentieren

Beispiel: Bias in den Trainingsdaten



White
Educated
Industrialized
Rich
Democratic

Aktuelle Ergänzung des Sawtooth A&I-Summit – Ipsos & Microsoft

- “What did they find? For well-known product categories in markets (e.g., countries) well-represented by the information available on the Internet, AI respondents directionally give similar predictions as human respondents ... usually. Sometimes, it doesn't get the preference order right for levels within attributes and concepts within simulated markets. But for less-known product categories and markets not well represented by information on the web, AI is going to be lost.”
- “Mean preferences and mean estimates for well-known products and well-covered markets seem reasonable in most cases. But the variance of the estimates representing the diverse tastes of respondents is vastly understated. The prompts Ipsos and Microsoft teams gave AI to differentiate among customer groups didn't lead to nearly the degree of differentiation in tastes observed in real human respondents. Furthermore, the temperature setting, which is meant to add more random variability to AI's responses was largely inconsequential.”

Zentrale Quellen

- Atari et al. (2023): Which Humans?, verfügbar unter <https://coevolution.fas.harvard.edu/publications/which-humans>
- Brand, Israeli & Ngwe (2023): Using GPT for Market Research, verfügbar unter https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4395751
- Sarstedt et al. (2023): Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines, verfügbar unter <https://doi.org/10.1002/mar.21982>

Kontakt

Prof. Dr. Steffen Mueller

Leiter Fachstelle Behavioral Insights & Pricing
ZHAW School of Management and Law
Theaterstrasse 17
8400 Winterthur
Schweiz

steffen.mueller@zhaw.ch
+41 58 934 7924



LinkedIn 