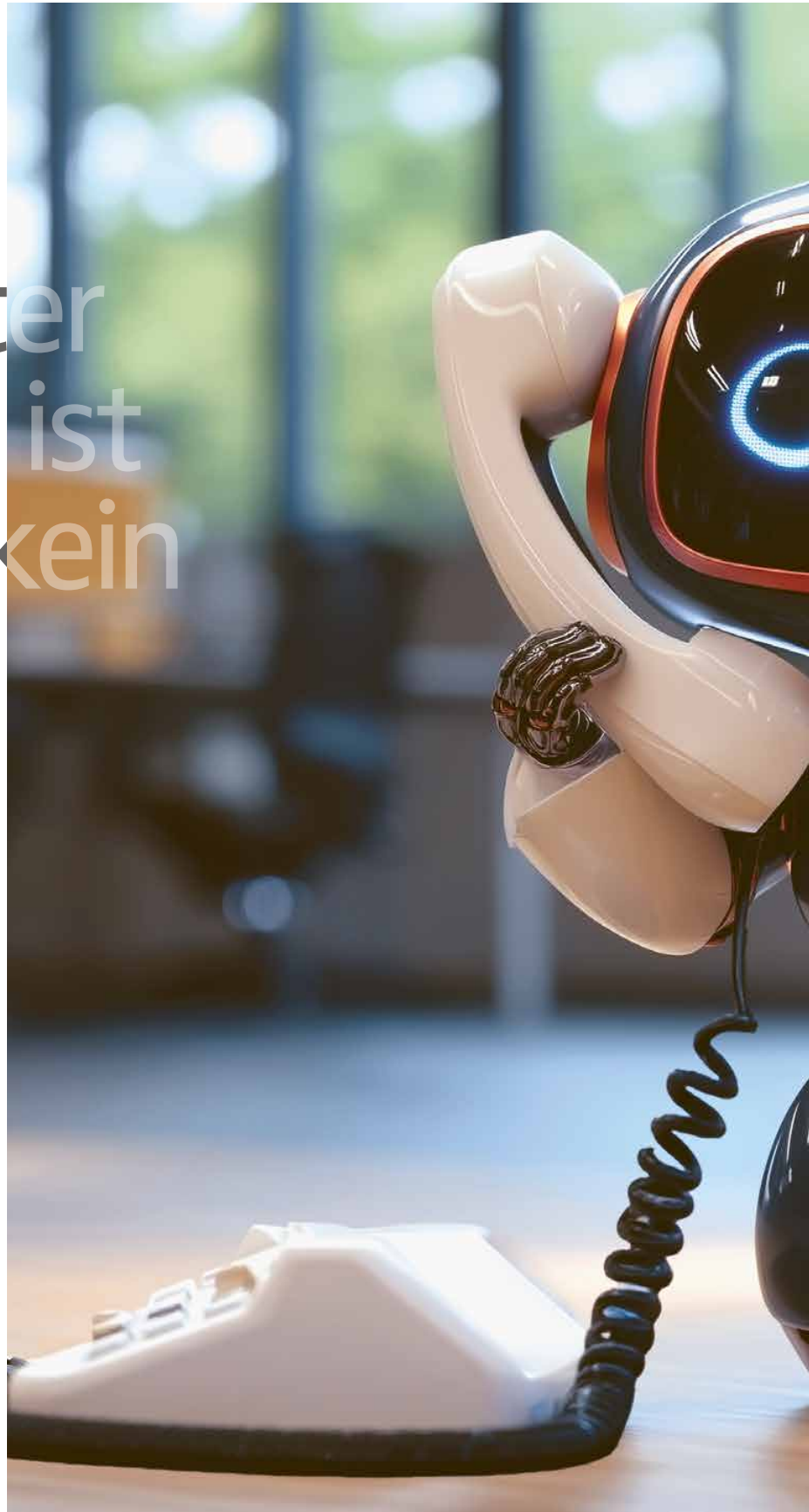




Ihr nächster Befragter ist vielleicht kein Mensch



*Leitlinien zum
Einsatz von
Silicon Samples
in der Markt-
forschung*



Midjourney AI prompt by GCO

**AUTOR:INNEN****Marko Sarstedt**

Professor für Marketing
 Institut für Marketing
 Ludwig-Maximilians-Universität München
 und Faculty of Economics and Business
 Administration
 Babeş-Bolyai University
 Cluj-Napoca, Rumänien

Susanne J. Adler

Postdoktorandin
 Institut für Marketing
 Ludwig-Maximilians-Universität München

Lea Rau

Wissenschaftliche Mitarbeiterin
 Institut für Marketing
 Ludwig-Maximilians-Universität München

Bernd Schmitt

Professor of International Business
 Columbia Business School
 Columbia University, New York

KEYWORDS

Generative künstliche Intelligenz
 Große Sprachmodelle (LLMs)
 Silicon Samples

Stellen Sie sich ein Marktforschungspanel vor, das nie verspätet Ergebnisse liefert, bei dem niemand mitten in der Befragung aussteigt und das sich per Mausklick sofort auf Tausende von Teilnehmenden erweitern lässt. Das ist keine Utopie, sondern die sich abzeichnende Realität sogenannter „Silicon Samples“ – synthetischer Befragter, die von großen Sprachmodellen (Large Language Models, LLMs) erzeugt werden. LLMs sind eine Form der generativen künstlichen Intelligenz und können mehr als nur Texte produzieren: Sie werden zunehmend eingesetzt, um menschliche Antworten in Interviews, Fokusgruppen oder experimentellen Szenarien zu simulieren. Für die Marktforschung eröffnet sich hierdurch eine große Chance. Anstatt sich ausschließlich auf menschliche Teilnehmende zu verlassen, könnten Daten auf Abruf, in großem Umfang und zu geringen Kosten durch LLMs generiert werden. Natürlich wirft dies auch kritische Fragen auf. Wie sehr ähneln Silicon Samples tatsächlich den Antworten menschlicher Probanden? In welchen Forschungskontexten sind sie zuverlässig und wo stoßen sie an Grenzen? In diesem Beitrag greifen wir diese Fragen auf. Wir skizzieren praxisnahe Anwendungsfälle, schlagen Richtlinien für Silicon Sam- ▶

Wichtige Fragen zur Erzeugung von Silicon Samples

ABBILDUNG 1



GettyImagey/KALA STUDIO

pling vor und beschreiben Einschränkungen, die es zu berücksichtigen gilt.

Silicon Samples: Was sie sind und wie man sie erzeugt

Silicon Samples bezeichnen synthetisch generierte Datensätze, die darauf abzielen, menschliches Antwortverhalten zu simulieren, und die zur Beschreibung, Erklärung und Prognose menschlichen Verhaltens herangezogen werden können. In der Regel sind Silicon Samples dabei auf vordefinierte Zielgruppen und demografische Segmente zugeschnitten. Die Generierung von Silicon Samples ist allerdings nicht trivial. Sie erfordert mehrere Prozessschritte, wie beispielsweise die Auswahl eines geeigneten Modells und dessen Anpassung an den jeweiligen Kontext, den Einsatz spezifischer Prompting-Techniken, geeignete Sampling-Verfahren sowie Validierungsmaßnahmen (siehe Abbildung 1).

Anwendungsfälle und Potenziale von Silicon Sampling

Durch ihre Fähigkeit, menschliche Antworten zu imitieren, können Silicon Samples entlang des gesamten Marktforschungsprozesses eingesetzt und in bestehende Workflows integriert werden. Die akademische Forschung zu diesem Thema entwickelt sich derzeit rasant weiter. Auf Basis der bis-

herigen Erkenntnisse lassen sich folgende Anwendungsfelder identifizieren:

Pretests von Stimuli und Fragebogenelementen
LLMs können in Pretests und Pilotstudien als beratende Instanz fungieren. Sie schlagen alternative Formulierungen vor und unterstützen die Überarbeitung von Forschungsmaterialien, etwa durch das Erkennen mehrdeutiger Fragen oder unklarer Antwortkategorien. Multimodale LLMs können zudem visuelle Stimuli evaluieren, indem sie deren Gestaltungselemente analysieren und Verbesserungsvorschläge machen. Durch gezieltes Prompting lassen sich außerdem unterschiedliche kulturelle und individuelle Perspektiven abbilden, um verschiedene Zielgruppen widerzuspiegeln.

Generierung umfangreicher Daten für quantitative Untersuchungen
Silicon Samples können quantitative Studiendaten ergänzen oder in bestimmten Anwendungsfällen sogar ersetzen. LLMs sind in der Lage, aktuelle Konsumentendaten aus Quellen wie sozialen Medien oder Bewertungsplattformen zu analysieren, was ein kontinuierliches Benchmarking ermöglicht. Auf dieser Grundlage lassen sich beispielsweise Kunden- oder Marken Kennzahlen nahezu in Echtzeit ableiten. Darüber hinaus eignen sich LLMs aufgrund ihrer hohen Geschwindigkeit und Skalierbarkeit dazu, bestehende Datensätze gezielt zu ergänzen.

Generierung synthetischer Personas für qualitative Erkenntnisse

LLMs können zur Entwicklung synthetischer Personas eingesetzt werden, welche an qualitativen Untersuchungen teilnehmen. Auf diese Weise können Forschende fiktive Interviews durchführen, alternative Gesprächsverläufe analysieren und nachträglich neue Fragen in bereits geführte Interviews einbeziehen. Solche qualitativen Datenerhebungen lassen sich auch auf Fokusgruppen ausweiten, in denen synthetische Personas unter der Leitung eines menschlichen Moderators miteinander diskutieren.

Herausforderungen und Grenzen von Silicon Sampling

Trotz dieser Potenziale stellt der Einsatz von Silicon Samples weiterhin eine erhebliche Herausforderung dar, wodurch ihre Eignung für strategische Entscheidungen begrenzt ist. Zwar können LLMs typische Antwortmuster reproduzieren, sie erfassen jedoch nur eingeschränkt die individuellen Erfahrungen, kulturellen Kontexte und situativen Faktoren, die menschliche Entscheidungsprozesse prägen.

Domänenspezifische (Under-)Performance

In einer aktuellen systematischen Analyse von 285 Vergleichen zwischen synthetisch generierten und menschlichen Stichproben zeigt sich, dass 24,9 % der Vergleiche zu ähnlichen Ergebnissen führten, während 65,3 % voneinander abwichen und nur 9,8 % teilweise übereinstimmten. Obwohl LLMs Ergebnisse in Bezug auf einige stabile Konstrukte, wie Persönlichkeitsmerkmale und politische Präferenzen, replizieren konnten, stimmten ihre Ergebnisse oft nicht mit etablierten Effekten der Konsumentenverhaltensforschung überein.

Verzerrungen durch Trainingsdaten

Die Qualität der Ergebnisse hängt unmittelbar von den zugrunde liegenden Trainingsdaten ab – ein klassischer Fall von „Garbage in, garbage out“. LLMs werden überwiegend mit frei zugänglichen, englischsprachigen Textdaten aus Quellen wie Wikipedia und GitHub trainiert. Entsprechend spiegeln ihre Antworten vor allem Perspektiven und Merkmale westlicher Gesellschaften wider und sind für die Forschung mit anderen Populationen nur eingeschränkt geeignet.

Sensitivität gegenüber Prompt Design

LLMs reagieren sensibel auf die Gestaltung der Prompts. Bereits geringe Variationen in Struktur, Antwortreihenfolge, Formulierungen oder der Darstellung von Informationen können die generierten Outputs beeinflussen und zu systematischen Verzerrungen führen.

Fehlende Varianz

Antworten von LLMs weisen häufig eine geringe Varianz auf. Viele Forschungsprojekte sind jedoch gerade auf Varianz angewiesen, um Zusammenhänge zu identifizieren oder Segmente zu bilden. Zwar lässt sich durch gezielte Prompting-Techniken künstliche Variabilität erzeugen, jedoch wird diese

bewusst von den Forschenden definiert. Dadurch werden genuine, unvorhergesehene Effekte, wie sie in menschlichen Stichproben typischerweise auftreten, unterdrückt.

Leitlinien zur Generierung valider Silicon Samples in der Marktforschung

Angesichts der genannten Limitationen sollte der Output von LLMs nicht unreflektiert als Entscheidungsgrundlage herangezogen werden. Die Erzeugung valider Silicon Samples erfordert vielmehr eine sorgfältige Prüfung, einschließlich einer kritischen Bewertung des jeweiligen LLM-Outputs durch Forschende sowie eines systematischen Vergleichs mit von Menschen generierten Daten – sowohl auf Ebene einzelner synthetischer Befragter als auch für das gesamte Silicon Sample (siehe Abbildung 2).

Modellauswahl und Prompt-Design spielen dabei eine zentrale Rolle. LLMs unterscheiden sich deutlich in ihrer Fähigkeit, menschliche Antwortmuster zu replizieren. Modelle, die an nutzerspezifische Anforderungen angepasst sind, übertreffen in der Regel Modelle „von der Stange“. Entsprechend empfehlen wir den Einsatz mehrerer Modelle sowie – sofern möglich – die Einbindung zusätzlicher Kontextinformationen durch Fine-Tuning oder Retrieval-Augmented Generation (RAG). Ebenso entscheidend ist das Prompt-Design für die Validität eines Silicon Samples. Ein Prompt sollte Informationen zum Kontext, zur Aufgabe sowie zum gewünschten Antwortformat und gegebenenfalls Beispieloutputs enthalten und so lange optimiert werden, bis das LLM das gewünschte Ausgabeformat zuverlässig liefert. Ergebnisse verschiedener Prompts sollten anschließend in Bezug auf die Verteilungen einzelner Werte analysiert werden (siehe Abbildung 2). Wie sieht das „durchschnittliche“ Ergebnis der Verteilung, gemessen an Mittelwert, Median oder Modus aus? Wie stark streuen die Ergebnisse in Bezug auf Varianz oder Spannweite? Verteilungen können dann mit einer Referenzverteilung menschlicher Probanden verglichen werden. Je näher die vom LLM erzeugte Verteilung an den menschlichen Referenzwerten liegt, desto besser ist die Qualität des Silicon Samples.

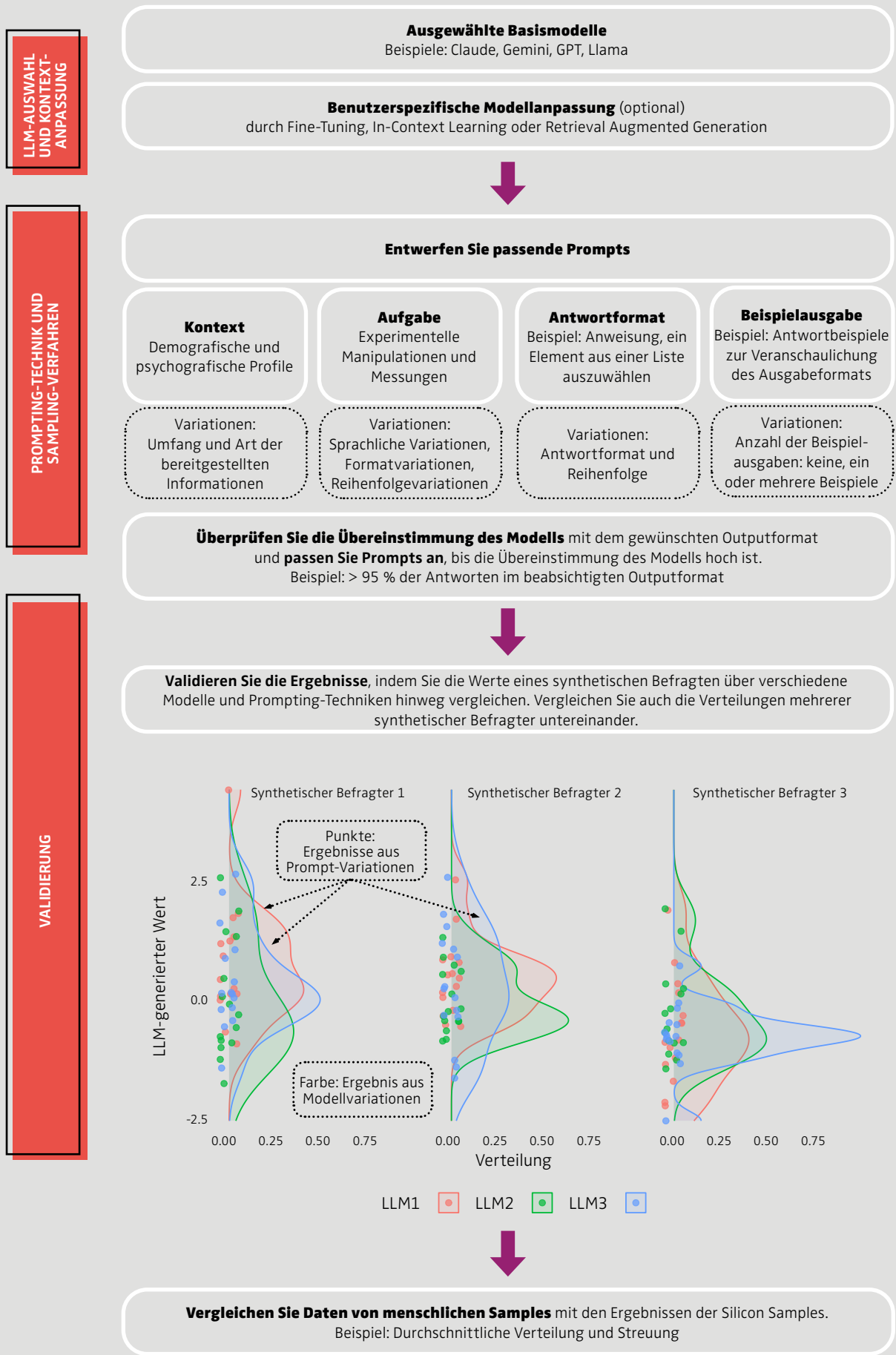
Die bisherige Diskussion verdeutlicht, dass Silicon Samples zwar ein beachtliches Potenzial aufweisen, ihre Generierung und Evaluation jedoch mit erheblichen Herausforderungen verbunden sind. Abbildung 3 fasst die zentralen Aspekte zusammen.

Die Zukunft liegt in spezialisierten Modellen für spezifische Aufgaben

LLMs können menschliches Verhalten zukünftig immer besser abbilden, insbesondere wenn sie multimodale Eingaben wie Bilder, Videos und Sprache integrieren und differenziertere, menschenähnliche Denkfähigkeiten entwickeln – ein Bereich, der in der wissenschaftlichen Literatur zuletzt verstärkt diskutiert wird. Fortschritte beim Prompt-Design und der Kontextintegration (sog. In-Context Learning) werden ebenfalls dazu beitragen, Silicon Samples weiterzuentwickeln. ▶

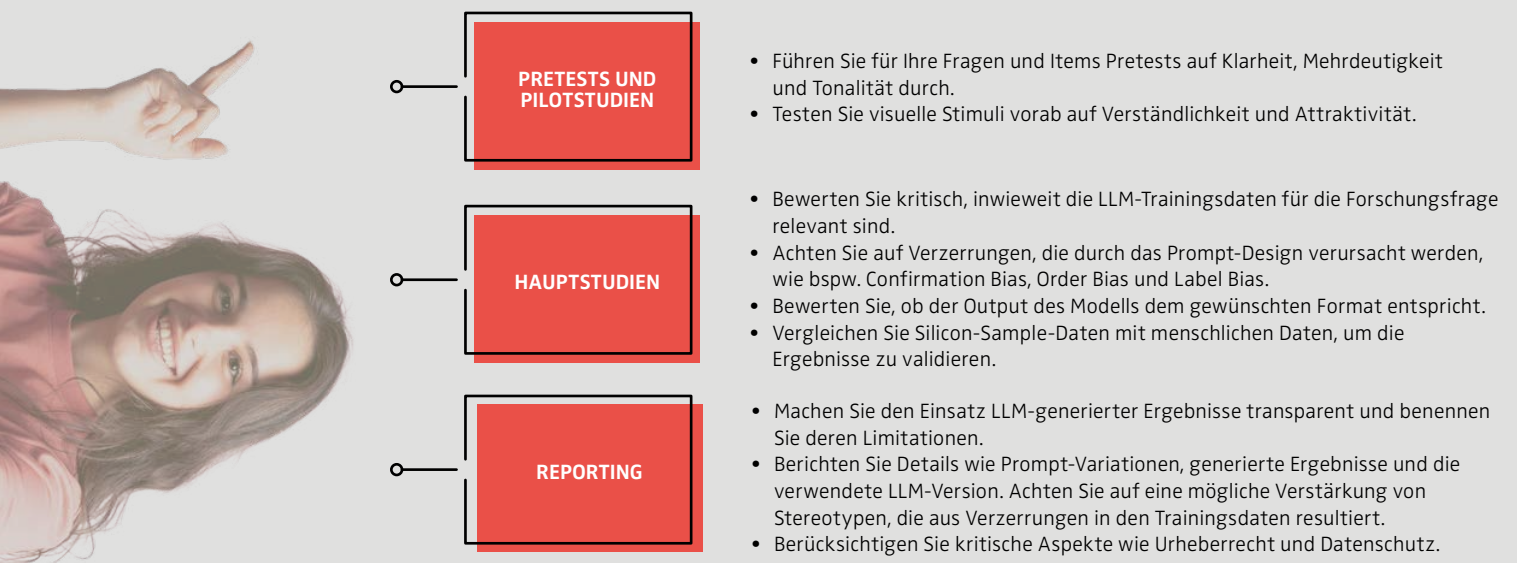
Workflow zur Generierung valider Silicon Samples

ABBILDUNG 2



Leitlinien für Forschungsprojekte, die mit Silicon Samples arbeiten

ABBILDUNG 3



GettyImagey/KALA STUDIO

Im Gegensatz zu menschlichen Daten bilden die Antworten eines LLMs allerdings nicht die vielseitigen Erfahrungen und Persönlichkeitseigenschaften ab, die Menschen ausmachen. Bislang ist unklar, wie der menschliche Geist im Detail funktioniert und wie ähnlich Outputs von LLMs dem menschlichen Denken tatsächlich werden können. Unstrittig ist jedoch, dass LLM-Architekturen nicht auf die gleiche Weise entstanden sind wie der menschliche Geist im Rahmen biologischer Evolutionsprozesse. Obwohl LLMs menschliche Antwortmuster imitieren können, verfügen sie weder über eine körperliche noch über eine sensorische Verankerung. LLMs „erleben“ die Welt daher anders als Menschen und damit fehlt ihnen eine zentrale Grundlage menschlichen Urteilens. Diese Divergenz begrenzt die Fähigkeit von LLMs, menschliche Entscheidungsprozesse, emotionale Reaktionen und kulturelle Vielfalt adäquat abzubilden. Deshalb wird die Validierung von Silicon Samples auch künftig eine zentrale Rolle spielen.

Große, universell einsetzbare Modelle, wie die neuesten Versionen von GPT, Llama oder Claude, werden wohl weiterhin Schwierigkeiten bei spezialisierten Aufgaben im Silicon Sampling haben. Substanzielle Fortschritte – wie etwa der Sprung von GPT-3.5 zu GPT-4o – wären notwendig, um die Leistungsfähigkeit beim Silicon Sampling deutlich zu erhöhen, doch solche Verbesserungen scheinen für kommende Modelle eher unwahrscheinlich, wie die teilweise enttäuschten Reaktionen auf GPT-5 nahelegen. Die Ära exponentiellen Wachstums in der allgemeinen LLM-Performance scheint vorerst vorbei zu sein. Stattdessen ist eher mit einer Fragmentierung, d. h. einer Entwicklung hin zu spezialisierten LLMs, zu rechnen, die anhand kontextspezifischer Daten auf einen begrenzten Aufgabenbereich abgestimmt sind. Im Bereich der Verhaltensforschung hat beispielsweise das Cen-

taur-Modell, ein feinjustiertes Llama-Modell, das Llama-Basis-Modell bei der Prognose menschlicher Antworten übertrifft. Ansätze wie In-Context Learning oder RAG machen Modellanpassungen zudem kostengünstiger und einem breiteren Nutzerkreis zugänglich als aufwendiges Fine-Tuning von Modellgewichten. Wir erwarten daher künftig eine stärkere Fokussierung auf spezialisierte Modelle, die in einer spezifischen Aufgabe besonders gut performen, beispielsweise bei der Nachahmung von Antworten einer bestimmten Zielgruppe. Erste Versionen solch personalisierter Modelle existieren bereits und das Training von LLMs mit großen Mengen persönlicher Daten wird vermutlich noch bessere Silicon Samples ermöglichen. Um dieses Potenzial auszuschöpfen, sind enge Kooperationen zwischen Marktforschung, Psychologie und Informatik notwendig, damit die gewonnenen Erkenntnisse für die Messung realen menschlichen Verhaltens relevant bleiben. ◀

LITERATURHINWEISE

- Brucks, M., & Toubia, O. (2025).** Prompt architecture induces methodological artifacts in large language models. *PLoS One*, 20(4), <https://doi.org/10.1371/journal.pone.0319159>
- Gao, Y., Lee, D., Burtch, G., & Fazelpour, S. (2025).** Take caution in using LLMs as human surrogates. *Proceedings of the National Academy of Sciences*, 122(24), <https://doi.org/10.1073/pnas.2501660122>
- Sarstedt, M., Adler, S. J., Rau, L., & Schmitt, B. (2024).** Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. *Psychology and Marketing*, 41(6), 1254–1270. <https://doi.org/10.1002/mar.21982>
- Toubia, O., Gui, G. Z., Peng, T., Merlau, D. J., Li, A., & Chen, H. (2025).** Twin-2K-500: A data set for building digital twins of over 2,000 people based on their answers to over 500 questions. *Marketing Science*, 44(6), 1446–1455. <https://doi.org/10.1287/mksc.2025.0262>