Synthetic market research: perceptions, concerns and quality standards

Simone Griesser
Institute of Communication and Marketing, Lucerne School of Business
Carina Junker
HSLU - Lucerne University of Applied Sciences and Arts Business
Alicia Aregger
HSLU - Lucerne University of Applied Sciences and Arts Business
Nadja Carletti
HSLU - Lucerne University of Applied Sciences and Arts Business
Nicole Schmid
HSLU - Lucerne University of Applied Sciences and Arts Business
Arta Syla
HSLU - Lucerne University of Applied Sciences and Arts Business

Acknowledgements:
The authors thank Chantal Magnin and Robin Kabul for their kind and valuable support.

Cite as:
Griesser Simone, Junker Carina, Aregger Alicia, Carletti Nadja, Schmid Nicole, Syla Arta (2026), Synthetic market research: perceptions, concerns and quality standards. *Proceedings of the European Marketing Academy*, 55th, (134088)

Paper from the 55th Annual EMAC Conference, Bath, United Kingdom, June 2-5, 2026

Synthetic market research: perceptions, concerns and quality standards

Abstract:

Large Language Models (LLMs) create new possibilities for market research by generating synthetic data that mimics human responses. Little is known about how marketing professionals perceive this innovation. Understanding their perceptions and quality concerns is critical because adoption depends on trust in data quality.

This study combines semi-structured interviews with marketing professionals and synthetic responses generated with Gemini to explore perceptions, usage and quality concerns. Findings show that speed and cost are key benefits while lack of customer interaction, unclear data origin, and risk of incorrect or biased outputs are major drawbacks. Transparency, standardisation, and certification are demanded. While some themes are unique to marketing professionals, others are unique to synthetic respondents. Moreover, answers from synthetic respondents tend to be more thematically similar than answers from marketing professionals. Overall, the results underscore opportunities for synthetic data in market research but emphasize the urgency of certification and trust-building measures.

Keywords: synthetic data, market research, Gen AI

Track: Methods, Modelling & Marketing Analytics

Full data, analyses and Python code available upon request.

1. Introduction

Ever since OpenAI introduced the first Large Language Model (LLM) called ChatGPT in November 2022, LLMs are changing many aspects of marketing. One aspect that has hitherto received limited attention is market research. Specifically, LLMs can imitate human responses to survey questionnaires, interviews, and focus groups, thereby creating synthetic data. According to the Alan Turing Institute, "*synthetic data is data that has been generated using a purpose built mathematical model or algorithm*." (Jordon et al. 2022:5).

Synthetic data is a new and dynamic field of research with numerous opportunities for research and industry alike. Synthetic data generation is quicker and costs less than data collected in a traditional manner. In the case of qualitative synthetic data, i.e. responses to interviews and focus groups, synthetic data are less influenced by interview dynamics such as social desirability effects or interviewee fatigue. However, standards for evaluating the validity and reliability of synthetically created responses are still being developed. Moreover, the opportunities that synthetic data pose for market research can only be harnessed if marketing professionals are willing to use synthetic data.

This study explores how marketing professionals think about synthetic data, perceived advantages, disadvantages, and how they assess the quality of synthetic data with semi-structured interviews. To the best of our knowledge, this study is the first of its kind.

In the spirit of synthetic data, the study draws on a traditional and synthetic sample totalling 14 responses. While synthetic responses to validated survey items (e.g. Serapio-García et al. 2023; Sorokovikova et al. 2024) or other surveys (e.g. Brand, Israeli, and Ngwe 2023; Estevez, Ballestar, and Sainz 2025) were used in few studies, only a handful of studies work with synthetically generated responses to interviews (e.g. Amirova et al. 2024; Arora, Chakraborty, and Nishimura 2025).

The contributions of this ongoing work are twofold. First, it portrays how marketing professionals in Switzerland think about synthetic data. Second, synthetic interview responses are compared with the actual responses. Previous work shows that the content analysis produced similar themes for the synthetic and human sample, but some themes were unique to the synthetic and some unique to the human sample (Arora et al. 2025). Therefore, more research is needed to understand how useful LLMs are to generate synthetic interview responses, especially for languages other than English. The previous study was conducted in

English (Arora et al. 2025). In contrast, the current study draws on interview data in German from Switzerland.

2. Literature review

Existing studies about synthetic data can be classified into five categories: synthetic responses to i) validated survey items, ii) questions with a (ranking) scale as answering option, iii) open questions in surveys requiring a text response, iv) interviews, and v) focus groups.

The first category comprises scales that have been validated by numerous studies such as the personality trait scale NEO-PI-R (Costa Jr 1992) or The International Personality Item Pool Big-Five Factor Markers (IPIP BFFM) (Pettersson and Turkheimer 2010) for example. In several studies synthetic personality data were generated and validated. GPT-4 realistically emulated 400 responses to the IPIP BFFM survey with internal consistency measures ranging between 0.97 and 0.99 and convergent validity scores ranging between 0.90 and 0.94. In fact, the synthetic responses are internally more consistent and yield a higher discriminant validity than the 400 human responses. In another study, 18 different LLMs generated synthetic NEO-PI-R personality scores of variable quality ranging from unacceptable to excellent (Serapio-García et al. 2023). Taking reliability, as well as convergent, discriminant, and criterion validity measures as quality indicators, the LLMs GPT-4o and Flan-PaLM achieved excellent scores across all measures. LLMs from the Mistral and Llama family produced unacceptable or weak results. Both studies benefited from existing datasets to compute quality measures.

In the second category, the responses an LLM generates are also based on a scale, but the items are not validated and comparison datasets do not necessarily exist. GPT-4 mimics the purchase likelihood and attitudes towards refrigerated dog food reasonably well, based on a comparison between mean and standard deviation in the synthetic and human data (Arora et al. 2025). GPT-3.5 Turbo imitated willingness-to-pay (WTP) for two tooth paste brands (Brand et al. 2023). While GPT ranked the importance of the pre-defined product attributes determining WTP akin to the human sample, it overestimated the WTP for the attribute fluoride and the importance of non-mint flavours. Different LLMs indicated beer price, flavour, consumption habits, and advertising perceptions on a 4-point scale. The answers were comparable to 1002 human survey responses collected in Spain (Estevez et al. 2025).

The third category comprises text responses in surveys. The words GPT-3 generated to describe republican or democratic voters were similar to human respondents (Argyle et al. 2023). In another study, GPT-3.5 answered open-ended questions about perceived barriers and enablers of physical activity (Amirova et al. 2024). While the topics found in the synthetic responses are comparable to human responses, the tone and structure of the response were not.

Transcribed responses to questions asked in an interview or focus group fall in the fourth and fifth categories. In their seminal study Arora and colleagues (2025) tasked GPT-4 to replicate interviews with consumers about their attitudes towards Friendsgiving, a more modern alternative to Thanksgiving. This is possible because GPT-4 supports user and system roles. The system emulated the interviewer and the user the interviewee. When GPT-4 is given a description of each of the interviewees in the human sample, it produces thematically similar answers. However, there are themes that are unique to the synthetic responses and themes unique to human responses, highlighting the need for a hybrid approach combining human and synthetic data. Another study aimed to enrich synthetic responses in career interviews with a dialogue system utilising different LLM agents (Hasegawa et al. 2025). The system design is as such that one LLM checks whether the synthetic response includes the necessary information. If it does, the dialogue moves on to the next question. If it does not, a follow-up question is asked to elicit the necessary information from the LLM agent imitating the interviewee. Thus, the dialogue system resembles the conversational flow of semi-structured interviews as well as the learning curve experienced by an interviewer. Results show that the dialogue system elicited responses with more attributes about the interviewees than in the baseline condition.

3. Methodology

This research adopts a mixed-method approach combining a qualitative and computational quantitative methods. In previous work, this produced more insights (Arora et al. 2025).

3.1 Qualitative method

The qualitative method embraces a phenomenological interpretive approach (Larkin et al., 2006) aimed at providing rich accounts of how marketing professionals perceive and define synthetic data, their experiences with (synthetic) market research, and how they assess quality. For this reason, 7 semi-structured interviews with German-speaking marketing professionals in

Switzerland were conducted lasting between 30 to 50 minutes. Interviewees were recruited with heterogeneous purposive sampling according to four criteria: i) job role, ii) industry, iii) use of market research, iv) use of LLMs. Respondents work in distinct roles, some in managerial roles. They work in different industries and have varying levels of experience with market research. However, they all use or even produce market research in their professional capacity. Similarly, all respondents use LLMs but in diverse ways. With this sampling method heterogeneous responses can be collected increasing theoretical saturation (Glaser and Strauss 1998) even when the sample is small. Table I gives an overview of the sample.

	Job Role	**Industry**	**Education**	**Company Size**
A	Marketing Manager leading team of marketers	Automotive	Business Administration & Marketing	Medium
B	Marketing Manager leading team of marketers	Logistics & financial services	Marketing & Opinion Research	Large
C	Marketing Manager	Information Technologies	Business Administration & Marketing	Small
D	Chief Operation Officer	Real estate	CAS in Real Estate Management	Medium
E	Content and Communication Specialist	Civil service organisation	Media Studies & Journalism	Medium to large
F	Head of Distribution Support (includes marketing)	Financial retirement provisions	(interviewee did not disclose information)	Medium
G	Founder & CEO of market research company	Various	MSc Information Technology	Small

Table 1: Description of human sample

Data was gathered between October and November 2025 with semi-structured interviews to account for the interpretive paradigm (Larkin et al., 2006). In line with the phenomenological stance, the interview questions were open and carefully worded to encourage narration. The interviews were recorded and transcribed with NoScribe, an open-source software, and analysed with the Interpretive Phenomenological Analysis (IPA) approach (Larkin et al., 2006). To identify and cluster statements on definitions, perceptions, experiences, and quality concerns the transcribed text has been analysed in two rounds, by manually coding, discussing initial categorisation, and refining the code system with different researchers.

3.2 Computational quantitative method

Synthetic data were generated in November 2025 with Google's Advanced Programming Interface using the LLM Gemini 2.5 flash-light. Pre-tests showed that Gemini 2.5 flash-light performed as well as Gemini 2.5 pro but required less computing time and cost. Gemini received system and user prompts (Arora et al. 2025). The system prompt emulated the

interviewer. It contained context and interviewee persona descriptions in German. In line with previous work (Arora et al. 2025) and for comparability reasons, interviewee personas are based on table 1. Persona descriptions were supplemented with how the interviewees use market research in their professional capacity and a market research familiarity rating given by the interviewer on a scale ranging from 1 (little market research familiarity) to 5 (comprehensive familiarity). For each persona description an interview transcript was generated. The user prompt emulated the interviewee answering the same questions the marketing professionals were asked. Temperature was set to 0.3 to reflect the nature of qualitative research. No interviewee responds in the exact same manner when asked the same question multiple times but is likely to answer with the same key themes for the same topic. Responses were analysed with topic modelling, a computational method that automatically identifies semantic clusters in large amounts of text. The topic model was computed with BERTopic (Grootendorst 2022) and the multilingual language sentence embedding model "paraphrase-multilingual-MiniLM-L12-v2" (Reimers and Gurevych 2019).

4. Results

First, the themes that emerged from interviews with marketing professionals are presented in the sections 4.1 to 4.4 followed by a comparison by the topics that emerged in the human and the synthetic sample.

4.1 How Marketing Professionals Define Synthetic Data

Respondents define synthetic data as data being produced from Artificial Intelligence (AI). Interviewee G highlights the artificial nature of synthetic data "*It's simply that behind it all, there's some kind of machine or some kind of technical device that did something.*". Interviewee A specifies that synthetic data is based on previous data. In contrast, interviewee F is unsure how to define synthetic data. When asked how synthetic data can be differentiated from human data, interviewee A states that it is impossible. Interviewees B and C state that it is impossible for quantitative data, but possible for qualitative data. Interviewee B adds that it is difficult to know whether the data comes from human or synthetic respondents unless it is made transparent in the methodology section. Interviewee E thinks that s/he can differentiate between real and synthetic data.

4.2 How Marketing Professionals Perceive Synthetic Data

Speed is a key advantage of synthetic data according to all interviewees. Synthetic data enables to "*get results at the touch of a button*" (interviewee B). Another advantage is the lower cost (interviewee A, C, and G). Particularly small and medium size enterprises (SMEs) can benefit from synthetic data as they have less market research resources than large companies.

Respondents mention various disadvantages of synthetic data. With synthetic data, there is no interaction with former, existing or potential customers (interviewee A). Important opportunities for customer relationship management are lost. This aspect is particularly important for customer-centric companies (interviewee B). Another disadvantage is traceability, i.e. where the data comes from. Interviewee C states "*Data can always come from Asia, America, or the Middle East*.". Interviewee D points out another aspect "*Because AI gets data from somewhere, and if someone writes something on the Internet that is not true but gets a lot of clicks, then AI assumes that it is true*." – a view echoed by interviewee F and G. Interviewee G notes that LLMs generate plausible but possibly inaccurate data. The mean tends to be correct, but the data distribution is wrong.

Another concern is the uncritical use of synthetic data. Synthetic data tends to meet the expectations of its audience and exclude surprising or irritating facts according to interviewee B. In a similar vein, interviewee A raises the concern of financial and reputational loss due to overlooking opportunities when basing decisions solely on synthetic data.

4.3 Experiences with (Synthetic) Market Research

None of the respondents have used synthetic data. While four of the seven respondents use market research, interviewee A uses forecasting models rather than market research studies. Interviewee C uses market research to optimize B2B Social Media campaigns. Interviewee D determines target groups for shopping malls in which his or her employer lets retail spaces with market research and optimises the (brand) positioning. Interviewee E uses market research to know how other municipalities communicate. Interviewee F leverages market research for competitor insight. Interviewees B and G produce market research.

4.4 Quality Assessments of Synthetic Data

A major quality concern is data origin. Interviewees A, C, and E ask for transparency about how the data was collected to be able to assess its quality. Moreover, interviewee C states that

LLM usage depends on socio-economic background. The user base influences an LLM's learning base, thus subsequently influencing the synthetic data LLMs generate. Similarly, for interviewee B, sample representativeness is key in determining data quality. This includes considerations such as whether the sample is a convenience sample or whether a self-selection bias is present.

According to interviewee B, the sample cannot be determined for synthetic data. Thus, additional quality criteria are needed. Interviewees A and B propose data heterogeneity as a quality criterion. Interviewee B calls this "*bandwidth of the data*" but also mentions split half reliability and comparability to human responses as quality criteria. Interviewee B remarks that even though plausibility is used as a quality criterion, it should not be used. Interviewee C draws on a similar line but has a different viewpoint when stating that s/he uses intuition as a quality criterion, particularly when findings are contradictory.

The reputation of the institute or authors who generate and analyse synthetic data is another important quality criterion for interviewees A, C, and G. Interviewee C adds that publications using synthetic data should be peer reviewed and goes on to say that if others are using synthetic data successfully, then trust in synthetic data increases. In contrast, interviewees E and F trust human data more than synthetic data because the latter is artificially generated. They view AI critically in general.

All respondents welcome the notion of a synthetic data certification with interviewee E pointing out that it will increase acceptance of synthetic data. Interviewee F mentions the importance of who issues the certificate and the reputation of the issuing body. The certification should be standardised (interviewee A) and include fairness and transparency criteria (interviewees A & E). According to interviewee B it should be comprehensive and include relevant criteria. This would make it trustworthy and relevant. Interviewee C points out that the degree of hallucination should form part of the certification. According to interviewee D the certification should require individuals working with synthetic data to state which LLM was used as well as the training data of the LLM.

4.5 Comparing Topics Emerging from Responses from Marketing Professionals and Gemini

There is considerable topic overlap in responses from marketing professionals and interviewee personas according to table 2. However, there are additional topics in the responses from interviewee personas that are not present in the responses from marketing professionals

and vice versa. The diverging responses are marked in italics in table 2. This finding is in line with previous work (Arora et al. 2025).

	Marketing professionals	**Interviewee Personas**
Advantages	Speed Cost *Benefit for SMEs*	Speed Cost *Avoid data protection issues* *Get data when it is otherwise hard or impossible to obtain, e.g. extremely sensitive data, or extremely costly, e.g. data on how municipalities communicate* *Risk management* *Scalability; different scenarios can be tested*
Dis-advantages	*No customer interaction* Data origin Correctness of data *Mean is correct but wrong distribution* Financial & reputational loss due to inadequate decision making Uncritical use of synthetic data	Quality unclear Perceived risk associated data of unknown quality *Models generating synthetic data need to be validated regularly*
Quality	Concerns are: • Transparency: data origin / collection process • Data does not represent required sample Quality criteria are: • comparability to human responses • statistical tests, e.g. half reliability • *reputation of the institute or authors generating synthetic data*	Concerns are: • Data origin / transparency in data collection process Quality criteria are: • Validation metrics • *Data protection* • Reproducibility
Certification	• Increases acceptance of synthetic data • Reputation of the issuing body is important to avoid partiality and increase accptance • Should be standardised • Detail data generation process including training data of LLM • Include fairness, transparency aspects, and degree of hallucination	• Detail transparency of data generation • Include fairness, and ecological aspects • Issuing body must be independent to be accepted • *Is expensive for (market research) companies* • Danger of White-washing • Difficult in fast-moving technological space; *needs to be up-dated regularly* • *Need to be adaptable to developments* • *Helps to select trustworthy data provider*

Table 2: Topic comparison between synthetic and human responses

A visual inspection of the data revealed that interviewee personas gave thematically more similar responses than marketing professionals. There is less variation between respondents in the synthetic sample compared to the human sample.

5. Conclusion and Limitations

Synthetic data provides ample opportunities but urgently needs standardisation and a framework to increase trust and adoption of synthetic data. An official certification would further support this cause. This ongoing work provides preliminary insights into a new and dynamic research field. More research is needed with a larger sample analysing between subject variation in synthetic samples, comparing topics between synthetic and human

respondents on an individual level with topic modelling, key message extraction or sentence similarity.

References

Amirova, Aliya, Theodora Fteropoulli, Nafiso Ahmed, Martin R. Cowie, and Joel Z. Leibo. 2024. 'Framework-Based Qualitative Analysis of Free Responses of Large Language Models: Algorithmic Fidelity'. *Plos One* 19(3):e0300024.

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. 'Out of One, Many: Using Language Models to Simulate Human Samples'. *Political Analysis* 31(3):337–51.

Arora, Neeraj, Ishita Chakraborty, and Yohei Nishimura. 2025. 'AI–Human Hybrids for Marketing Research: Leveraging Large Language Models (LLMs) as Collaborators'. *Journal of Marketing* 89(2):43–70. doi:10.1177/00222429241276529.

Brand, James, Ayelet Israeli, and Donald Ngwe. 2023. 'Using LLMs for Market Research'. *Harvard Business School Marketing Unit Working Paper* (23–062).

Costa Jr, T. P. 1992. 'The NEO-PI-R Professional Manual: Revised NEO Five-Factor Inventory.(NEO-FFI)'. *Psychological Assessment Resources*.

Estevez, Macarena, María Teresa Ballestar, and Jorge Sainz. 2025. 'Market Research and Knowledge Using Generative AI: The Power of Large Language Models'. *Journal of Innovation & Knowledge* 10(5):100796.

Glaser, Barney G., and Anselm L. Strauss. 1998. 'Grounded Theory'. *Strategien Qualitativer Forschung. Bern: Huber* 4.

Grootendorst, M. 2022. 'BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure'. *arXiv Preprint arXiv:2203.05794*.

Hasegawa, Ryo, Yijie Hua, Takehito Utsuro, Ekai Hashimoto, Mikio Nakano, and Shun Shiramatsu. 2025. 'A Dialogue System for Semi-Structured Interviews by LLMs and Its Evaluation on Persona Information Collection'. Pp. 39–59 in *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*.

Jordon, James, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N. Cohen, and Adrian Weller. 2022. 'Synthetic Data--What, Why and How?' *arXiv Preprint arXiv:2205.03257*.

Pettersson, Erik, and Eric Turkheimer. 2010. 'Item Selection, Evaluation, and Simple Structure in Personality Data'. *Journal of Research in Personality* 44(4):407–20. doi:10.1016/j.jrp.2010.03.002.

Reimers, Nils, and Iryna Gurevych. 2019. 'Sentence-Bert: Sentence Embeddings Using Siamese Bert-Networks'. *arXiv Preprint arXiv:1908.10084*.

Serapio-García, Gregory, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. 'Personality Traits in Large Language Models'.

Sorokovikova, Aleksandra, Sharwin Rezagholi, Natalia Fedorova, and Ivan P. Yamshchikov. 2024. 'LLMs Simulate Big5 Personality Traits: Further Evidence'. Pp. 83–87 in, *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*. St. Julians, Malta: Association for Computational Linguistics.