Bye-bye Bias: What to Consider When Training Generative AI Models on Subjective Marketing Metrics

THE AUTHORS

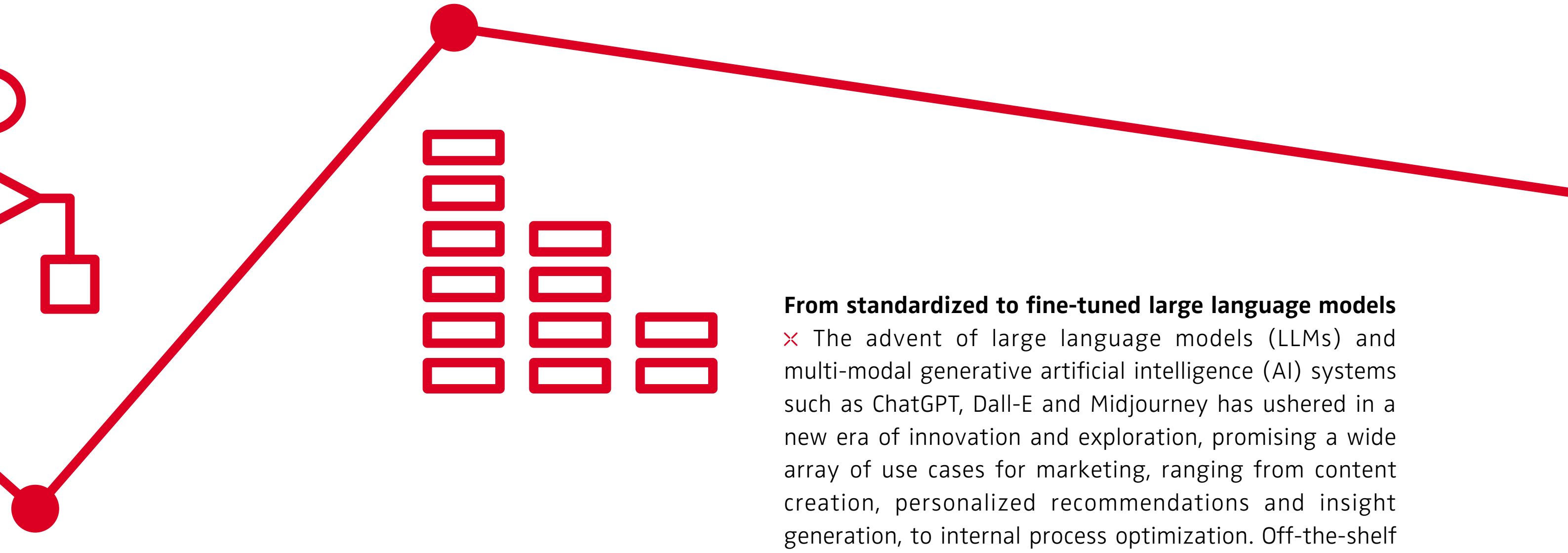
Christina Schamp
Full Professor of Marketing, Institute of Digital Marketing and Behavioral Insights, Vienna University of Economics and Business

Jochen Hartmann
Assistant Professor of Digital Marketing, TUM School of Management, GenAI Lab, Technical University of Munich

Dennis Herhausen
Associate Professor of Marketing, School of Business and Economics, Vrije Universiteit Amsterdam

KEYWORDS

Large Language Models (LLMs), Generative AI, Data Quality, Bias, Marketing Metrics



From standardized to fine-tuned large language models

The advent of large language models (LLMs) and multi-modal generative artificial intelligence (AI) systems such as ChatGPT, Dall-E and Midjourney has ushered in a new era of innovation and exploration, promising a wide array of use cases for marketing, ranging from content creation, personalized recommendations and insight generation, to internal process optimization. Off-the-shelf LLMs often exhibit remarkable "zero-shot" capabilities for these use cases, allowing them to generate content or make predictions without explicit training on specific tasks or the specific brand context. However, these models are trained on vast data sets scraped from the Internet, such as Common Crawl or LAION, that contain little information on the types of perceptual measures marketing is often interested in, such as perceived brand image and consumer engagement, or that lack the specific brand context. For example, some services offer direct ad creation hand in hand with search optimization. These ads are designed and optimized for high average click-through rates but do not provide brand differentiation and, at worst, can undermine brand image. Unlocking the true power of GenAI requires fine-tuning models to the specific brand context, which is often reflected

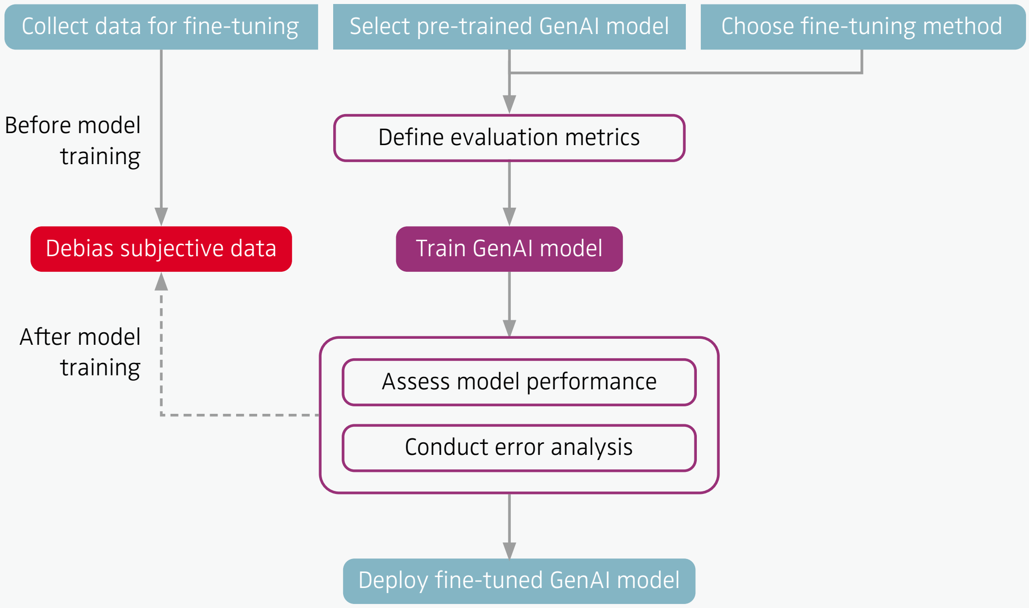
Potential biases in training data should be assessed and addressed before GenAI models are trained.

BOX 1

How companies can fine-tune GenAI models for specific marketing applications

The overall process of fine-tuning a GenAI model is comparable to any other supervised machine learning pipeline and is illustrated in Figure 1. The refined model emerging from such a high-level, multi-step pipeline is a function of three main inputs: the training data, a pre-trained GenAI model such as Llama 2 and the fine-tuning method. It goes without saying that the quality of the used data is key. Data quality can be assessed both before and after model training. Assessments of the model and its training data ex post are commonly conducted in computer and social sciences to understand the levers for improving model performance. Companies should, however, leverage insights from behavioral sciences and address potential biases already in the pre-model training stage. For example, training data can be debiased by selecting representative human coders and reasonable coding scales and by designing a frictionless "data annotation journey." All these factors will substantially contribute to a higher performance of the refined GenAI model, once deployed in a real-world setting.

FIGURE 1 > Process for fine-tuning GenAI models


Collect data for fine-tuning
Select pre-trained GenAI model
Choose fine-tuning method
Before model training
Define evaluation metrics
Debias subjective data
Train GenAI model
After model training
Assess model performance
Conduct error analysis
Deploy fine-tuned GenAI model

The performance of a fine-tuned pre-trained, open-source model such as Llama 2 is often competitive if not better than that of a general-purpose, closed-source model such as GPT-3.5. However, the performance benefits depend on the application context. Hence, quantifying the performance difference between a customized model and a baseline model is an important step before deploying a GenAI model. Clearly, additional considerations, such as data privacy, might inhibit companies from sharing their training data with commercial providers, making an offline fine-tuning process on local servers even more attractive.

FIGURE 2 > Checklist for ensuring data quality and preventing biases in training data

How to prevent

Sampling Bias

- Use random or representative sampling to ensure the representativeness of ratings.
- Collect control questions on relevant (demographic or customer segment) characteristics to check for equal distribution of variables in training and real life.
- Check for non-response bias and probable oversampling of certain subgroups with lower participation likelihood.
- Balance out dataset by up-weighing underrepresented relevant subgroups.

Measurement Bias

- Formulate short, clear and precise questions.
- Use clear instructions and introductory comments to define variables of interest/ psychological concepts.
- Rely on established scales that might break down difficult concepts into several items/ sub-questions and check the internal reliability of the questions.
- Check binary versus ordinary rating scales for more precise answers.

Social Desirability Bias

- Ensure and emphasize confidentiality in introductory remarks, incentivizing honest responses.
- Formulate neutral questions that do not imply any social norm.
- Reduce sensitivity of questions, e.g., by including trade-offs.
- Check responses with previous answers or existing data.

Response Bias

- Provide a frictionless survey environment to minimize survey fatigue.
- Limit overall survey time and set number of questions per repeated stimuli accordingly.
- Randomize the order of presented stimuli/data to be labeled to avoid order effects.
- Inspect and likely delete responses of bad quality using speed-clicking, straightlining or monotonous response patterns and inconsistencies as likely candidates.

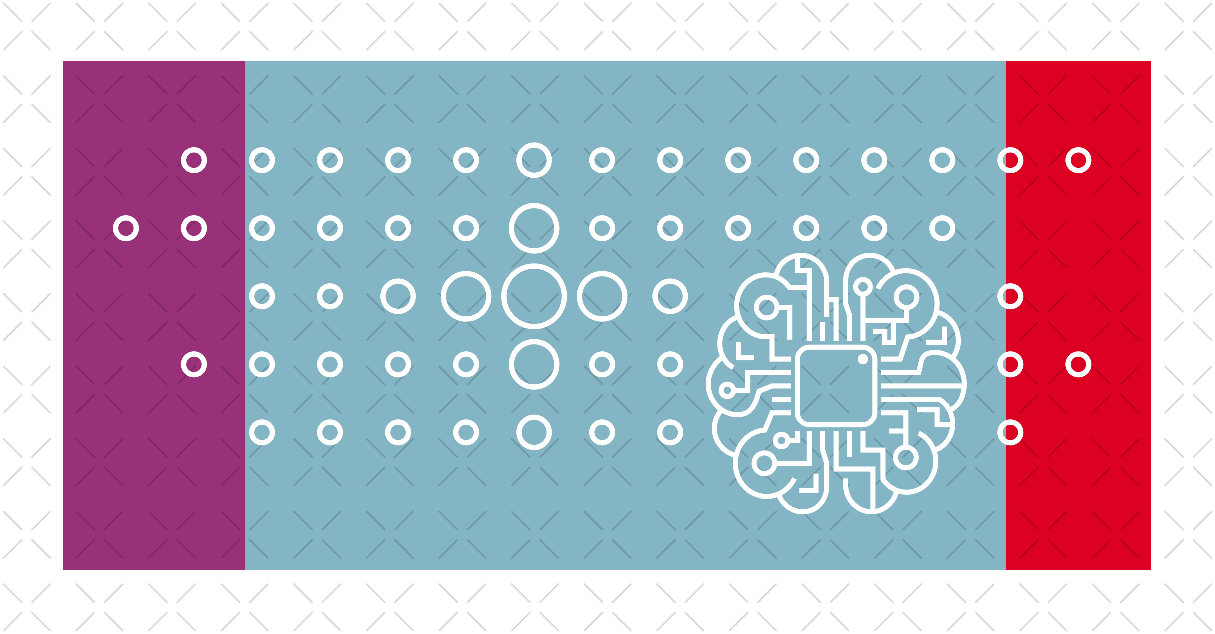
in rather subjective marketing metrics. Unlike objective metrics that can be classified as "right or wrong" by a viewer (e.g., is there a person in the ad?), subjective metrics capture the perceptions, opinions, feelings or beliefs related to the ad, which might differ across viewers. Box 1 and Figure 1 describe the technical process of fine-tuning GenAI models. The performance of these models critically depends on the quality of the underlying training data. Therefore, we focus on how to achieve top-quality training data to fine-tune GenAI models. This will be the differentiating factor in the race for the best AI applications in marketing.

Objective versus subjective training tasks × For objective training data, the underlying task is to detect elements or classify objects as either right or wrong. Typically, there is a clear ground truth: Is there a cat on an image? Is there a logo on a social media post? Is an email spam? In contrast, subjective training data capture perceptions, opinions,

» *Objective training data are already prone to error, but ensuring high-quality subjective training data on marketing metrics is even more challenging.* «

feelings and beliefs that might differ across different customer or population segments. These data might include the emotions elicited by a generated ad, whether an ad is funny or which arguments are most convincing during a customer interaction. Objective training data are already prone to error, but ensuring high-quality subjective training data on marketing metrics is even more challenging. For objective training data, a relatively low number of human coders is sufficient to achieve a low variability of answers and high intercoder reliability for labels. Conversely, subjective training tasks will, by definition, result in labels with higher variance. These labels are often measured on ordinal or even metric scales rather than categorical classes, at best from a larger representative sample of respondents. In addition, consumer perceptions might depend on the geographical or cultural context, and change over time, making it necessary to rerun the training procedures and update the algorithms on a regular basis.

Subjective training data allow for a more "customized" trained algorithm that is fine-tuned to the specific brand, product or customer context. To ensure their quality, marketers need to combine competences from different fields: behavioral insights based on survey research and machine learning (ML) based on objective data. Infusing traditional insights into AI models teaches GenAI valuable lessons from traditional marketing in how to attain differentiation, win the hearts and minds of consumers and improve bottom-line effectiveness.



Biases in training data and how to overcome them

× Ensuring reliable, high-quality training data, especially for subjective marketing metrics, has received surprisingly little attention. Currently, the quality of the underlying data is primarily evaluated ex post – after model training – by identifying systematic errors of the ML model. While quantifying errors is straightforward, understanding their reasons based on ML classifications is not. Further complicating matters, training data biases can be systematic and not detectable ex post.

We therefore recommend not only evaluating the quality of underlying AI training data after a ML model has been applied, but also accounting for the most prevalent pitfalls and biases in classic market research such as surveys or experiments in a pre-training stage (see Figure 1). This includes measuring and improving data quality ex ante – before feeding them into the training model – and providing recommendations that make the training more efficient. Figure 2 summarizes the most relevant biases that might occur in training data and provides recommendations on how to check for and avoid them for subjective marketing metrics tasks.

- **Sampling Bias** × This bias occurs when the humans used for training the algorithm or data labeling differ from the context in which the algorithm will be applied. This leads to systematic errors in model predictions. For subjective training tasks, sampling bias arises when the input data are not representative of the relevant population so that the distribution of the sampled population differs from the "true" underlying distribution in the relevant population. For instance, an algorithm that intends to measure to what extent an image reflects a brand's personality should reflect the perceptions of all customer segments for the relevant market. Different customer segments might have different consumer perceptions, markets might differ due to cultural differences and consumer perceptions might change over time. If training data fail to be sampled correctly, the algorithm will consequently fail to make good predictions and generalize to different contexts.

- **Measurement Bias** × Measurement bias relates to how the labeling occurs. If the chosen questions or labels are imperfect proxies for the real variables of interest, the outcomes of the model will result in misleading predictions. Even for objective tasks, measurement bias might occur in terms of label bias where labeled data systematically deviate from the underlying truth in the population. In reality, suggested labels might fail to precisely capture meaningful differences between classes, or cultural and individual differences might cause systematic deviations. For instance, generated texts for an authentic advertising claim might fail to meet the complex human perceptions of authenticity because the training data are based on only one single question and hence render an imprecise measurement. One way to address measurement bias is to collect multiple conceptually related measures to triangulate the underlying labeling intentions of respondents. Another way is to assess the variance between respondents from coherent target groups.

- **Social Desirability Bias** × Any training data capturing human perceptions, opinions or historical data are prone to social biases. These biases occur when available data reflect existing biases, norms or prejudices in the sampled population, resulting in unwanted outcomes of the AI model. For example, numerous biases have emerged where algorithms trained on past data discriminate against females or Black people in the context of banking, hiring or jurisdiction because the training data already reflected biases. One established mitigation method is to exclude protected attributes such as race, gender or ethnicity as input from the model to ensure fairness and equality. However, effects of discrimination might still prevail, as protected attributes might correlate with non-protected attributes of the model. To help understand and avoid such biases requires an in-depth investigation of the correlation matrix of the underlying training data as well as an expert discussion of potential consequences of the use of the algorithm in the real-world context.
  Relatedly, training data can include social (un)desirability bias. In contrast to responses that might inherently reflect prejudices and inequality, respondents often label and answer in a way that conforms to cultural norms. Thus, if respondents are aware of certain social expectations, they may label the data to accommodate these expectations. This is most likely to occur in AI models that attempt to predict consumer orientations towards sustainable, moral or healthy behavior. As a consequence, a marketing campaign and related generated content might assume exaggerated consumer preferences for influencers representing minorities or organic and sustainable products.

- **Response Bias** × While measurement bias relates to the questions and response options for the labeling, response

> *Biased training data, which in turn can bias the outputs of a GenAI model, should be a core concern in the development of GenAI models.*

bias relates to the labeling process itself. Compared to traditional surveys, labeling training data is often more repetitive and monotonous, focusing on a small set of questions for repeated varying stimuli. Whereas objective tasks can already be tiring and burdensome for the human coders, the more complex measurement for subjective tasks multiplies the likelihood of this bias. Thus, coders will be prone to response style biases that occur in overly lengthy or complex questionnaires. These include acquiescence – the tendency to agree with questions regardless of their content – and disacquiescence, where coders tend to disagree with questions or careless and arbitrary responses.

The generation of AI outcomes therefore depends on the quantity and sequence of coder tasks. Response biases can severely harm the efficiency and performance of the model and are of particular concern when the model is trained on only a few responses for each video, image or text, as GenAI requires sufficient variance at the content level.

More effectiveness and cost-efficiency of models with debiased data × Generative AI has the potential to transform marketing. True competitive advantage can be achieved when fine-tuning standard GenAI models to brand-specific tasks that can capture subjective marketing metrics. A crucial requirement, however, is to ensure top-quality training data. Biased training data, which in turn can bias the outputs of a GenAI model, should be a core concern in the development of GenAI models. A best-in-class model can be accomplished by assessing and addressing potential biases even before the model training, thereby complementing the current practice of error analysis in the post-training stage. This will make the AI model training not only more effective but also more cost-efficient. We recommend setting up interdisciplinary research teams that have both technical and market research skills and using software platforms to ensure a cutting-edge, frictionless data annotation journey. These measures help to combine all relevant perspectives and enable the development of successful GenAI use cases that have competitive advantages over standard applications. ×

FURTHER READING

Feuerriegel, S., Hartmann, J., Janiesch, C. et al. (2023). Generative AI. Business & Information Systems Engineering, 66, 111–126. https://doi.org/10.1007/s12599-023-00834-7

Hartmann, J., Heitmann, M., Schamp, C., & Netzer, O. (2021). The power of brand selfies. Journal of Marketing Research, 58(6), 1159–1177.

Van Giffen, B., Herhausen, D., & Fahse, T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. Journal of Business Research, 144, 93–106.