From Critique to Creation: Utilizing Generative AI for Transformative Product Design

Claus Hegmann-Napp
University of Hamburg
Tijmen Jansen
University of Hamburg

Cite as:
Hegmann-Napp Claus, Jansen Tijmen (2024), From Critique to Creation: Utilizing Generative AI for Transformative Product Design. *Proceedings of the European Marketing Academy*, 52nd, (119539)

Paper from the 53rd Annual EMAC Conference, Bucharest, Romania, May 28-31, 2024

From Critique to Creation: Utilizing Generative AI for Transformative Product Design

Customer reviews are not only a source of information for customers but also an important source of knowledge for businesses. However, the dissimilarity of these reviews poses a challenge to handle and use the information profitably. In this paper, we propose a novel approach to decoding the intricate language of customer reviews by utilizing Generative AI in form of Large Language Models (LLMs). We use a pre-existing dataset, classify relevant reviews containing beneficial information and use these to train a LLM. The model synthesizes the linguistic patterns in designated customer reviews, extracts valuable insights and generates innovative ideas. The results indicate that LLM training on a set of review data is feasible and the employment of Generative AI tools leads to promising ideas.

Keywords: *LLM, Customer Reviews, Transformative Product Design*
Track: *Innovation Management & New Product Development*

1 Introduction

Today's consumer markets rapidly growing, and new products are introduced on a daily basis. Especially in online distribution channels, there are different various decision-making aids such as customer reviews to help customers decide what they need. These reviews are not only a source of information for customers but also contain valuable insights for companies and brands. However, the sheer amount of review data and the dissimilarity pose a challenge to handle and use the information profitably (Flory et al. 2017; Zhou et al. 2018). Liu et al. (2020) inspect the processing of customer data in small and medium-sized enterprises and highlight the difficulties this presents without using modern, Generative AI-based solutions.

As businesses strive to stay afloat in the ever-evolving marketplace, leveraging Large Language Models (LLMs) has become a strategic necessity. With advanced natural language processing capabilities, LLMs offer a novel perspective for decoding the intricate language of customer reviews, uncovering hidden patterns, and extracting valuable attributes (Samha et al. 2014).

Lee and Bradlow (2011) present a method for extracting product attributes and brand positions from customer reviews, suggesting that reviews can reveal important attributes not captured by traditional methods and can serve as a filter to highlight meaningful attributes for product design and marketing strategies such as potential product innovation. Roelen-Blasberg et al. (2023) propose a new approach for inferring product attribute evaluations from online product reviews, indicating that automated methods can extract relevant attributes and their importance by analyzing patterns in designated customer reviews.

For companies, a potential subsequent benefit from such analysis could be the development of potential product innovations (Zhao et al. 2021; Tian and Yang 2023). Previous research suggests using LLMs to maximize innovation potential. Tian and Yang (2023) investigate the influence of online customer reviews on iterative innovation of software products, finding that both the sentiment and quantity of reviews positively affect innovation, although product complexity may limit the utility of reviews for high-complexity products. Zhao et al. (2021) highlight the significant influence of online customer reviews and deep learning on product innovation within the mobile application sector.

Since the majority of the presented publications are conducted in relation to software products, this study introduces a novel approach to investigate the analytical capabilities of a LLM to generate innovative ideas on non-technical product groups and its potential implementation as a supporting tool in product innovation and development. We want to show how AI's significant improvement in the ability to generate textual outputs can provide benefitial methods in the field of product innovation. In order to evaluate this topic, we leverage online available customer review data to investigate (1) the LLM's capability to extract pertinent information and (2) the perceived utility of the generated innovations among potential users of the selected product.

To provide answers to these questions, we use existing research to firstly identify relevant reviews which contain valuable information. In terms of developing innovation ideas, we follow the 'inside the box' approach of Boyd and Goldenberg (2019) and examine what level of creativity delivers the best results. The authors conclude that well perceived innovations do not necessarily need to be as creative and exceptional as possible but rather an adaption of already known features. The work by Toubia and Netzer (2017) delves into this subject through the utilization of big data, demonstrating that the optimal balance between novelty and familiarity enhances perceived creativity and liking of an innovative idea. For the assessment of our results, we conduct a survey about the generated innovations.

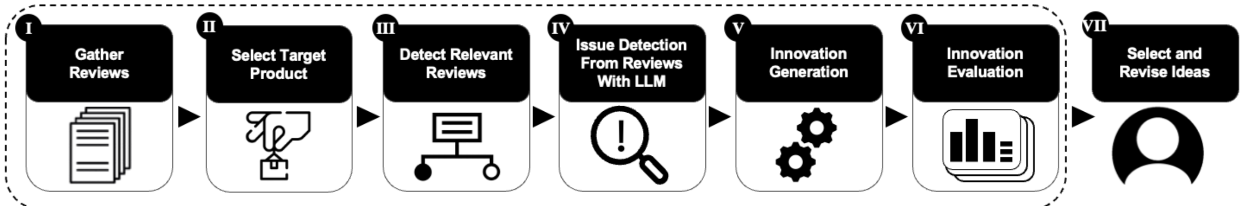
2 Using LLMs to Analyze Reviews and Generate Innovative Ideas

To investigate our research questions, we gradually create a model that analyzes reviews, identifies user problems and then develops possible innovations. The general approach can be seen in figure 1. Step I to VI cover the extent to which we focused on in this study, while Step VII could be the link to a product development department as a supporting tool.

2.1 Gather Reviews & Select Target Product (Step I & II)

We utilized a publicly available review dataset of 230 million reviews created by the University of California, San Diego (link: `https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/`). Next to the customer reviews, the rich dataset contains additional information for each individual product, including product description and the retail price. In order to cover a wide range of users and, as mentioned above, not to examine purely technical products, we limit the dataset to the category 'Office Products'. An advantage is the common daily use of such products, which ensures that mainly widely used products are included.

Figure 1: An Integrated Approach to Generating Innovative Ideas



After manually exploring the dataset, we randomly select a subcategory to demonstrate our approach. We chose the *'Desk Accessories'* and *'Workspace Organizers'* subcategory, comprising 9,243 products and 328,300 reviews. After reviewing the data, we selected the portable lap desk with the highest number of reviews, totaling 2,427.

2.2 Detect Relevant Reviews (Step III)

In the third step, we identify the relevant reviews in the remaining dataset. The idea is that we distinguish helpful reviews which contain insightful information from reviews that lack insights (e.g., 'I like this product'). Prior research on review attributes enhancing helpfulness is analyzed to narrow down the number of reviews and to exclude distracting and non-helpful reviews. Hong et al. (2017) conduct a meta-analysis of existing studies which examined the determinants of online review helpfulness. One significant positive factor is a lower review age. Considering that, we solely include reviews not older than five years.

Another aspect we take into account is high readability and longer text length of reviews, which both have a positive influence on the helpfulness of reviews (Korfiatis et al. 2012). The authors further identify the existence of extreme positions in a review enhances its helpfulness. In respect to these findings, we employ the Transformer-based classifier *SetFit* from the Hugging Face library for the purpose of discerning useful reviews. We take a subset of 100 reviews, manually assigning labels (1 = helpful, 0 = not helpful) and leveraging this labeled subset to the classifier. Following the fine-tuning of the model's hyperparameters, an optimal accuracy of 90.47% is achieved.

The concluding phase of the third step involves the prediction of labels for the selected dataset, culminating in the generation of the final dataset employed in further generation of innovations. This dataset ultimately includes 1,186 reviews.

2.3 Issue Detection From Reviews and Innovation Generation With LLM (Step IV & V)

The fourth step of our approach is to identify the main issues mentioned in the reviews and serves as preparation for later innovation generation. It is intended to analyze the dataset developed in prior steps. The core of our methodology involves the strategic integration of a LLM with our pre-existing review dataset. We decided to employ OpenAI's *GPT-4-Turbo* in our research. Through meticulous training iterations, the model synthesizes the linguistic patterns in the user reviews, extracting valuable insights and discerning recurrent themes.

As a consequence of this training phase, we extract the quintessence of consumer sentiment and prompt the model to extract the five most prevalent issues highlighted in the reviews. This compact analysis of user concerns not only serves as a testament to the LLM's ability to comprehend customer feedback, but also as the general starting point for our subsequent exploration into the generation of potential innovative solutions. Based on these extracted insights, we now start generating innovations using the LLM. In a pipeline-setup, we develop a prompt template to provide to the LLM. This template includes the product description from the original dataset and the previously extracted main issues derived from all reviews. To further specify the level of detail for the prompt, we deploy OpenAI's *GPT-4 with Vision (GPT-4V)* and provide it with the image of our chosen product. We add the resulting image description to the prompt template. At last, we prompt the LLM to come up with 15 innovations based on the designated information.

2.4 Innovation Evaluation (Step VI)

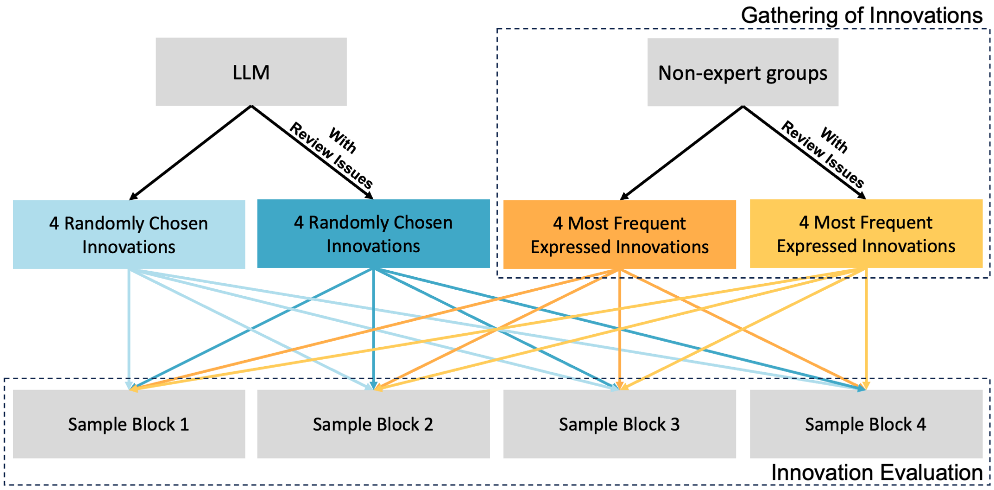
With the LLM generating innovative ideas based on the identified relevant information, it is essential to proceed with evaluation of the innovations. In a potential market situation, a product developer could play a part in that. Our approach is trained exclusively on customer input without any human interpretation. This implies the potential for more reliable outcomes and consistently improved innovation performance.

Nonetheless, the evaluation should consider potential users in the process. In an analysis of innovation-decision making models, Parthasarathy et al. (1995) emphasize the involvement of users and "imitators" in evaluation processes to optimize further implementation of innovative ideas in product development. This recommendation stems from the limitations identified in existing innovation-decision models that primarily focus on a sequential decision-making process.

To test our proposed research questions, we conduct a two-part study to demonstrate the capabilities of the pretrained LLM and compare this to non-expert groups. The general flow of the study is shown in figure 2. In the first part, the Gathering of Innovations, we gather innovative ideas from non-expert innovators. By doing so, we can compare a variety of innovations to our benchmark, which are the LLM-based innovations. In the second part, the Innovation Evaluation, we examine how the LLM-generated ideas perform against the non-expert ideas gathered in the first part of the study.

We supply participants with the same basic scope of information we used for our LLM in *Step V*. The difference is that instead of the image description, they are presented with the actual image of the product. We ask the non-expert can come up with up innovative ideas for the given product. In light of Boyd and Goldenberg (2019)'s previously mentioned 'inside the box' approach, we differentiate between two groups where one group is only presented with the product description, without the main issues derived from the reviews. In that way, we obtain one set of ideas influenced by the main issues and one set of ideas whose degree of creativity is solely up to the respondent. Additionally, we form a subset comprising

Figure 2: Customer Evaluation Process



15 LLM-generated innovations that were created without incorporating the pertinent issues derived from the reviews. In total, that leaves us with four sets of innovations.

As preparation for the evaluative second part of the study, we then select the four most frequent expressed innovations from the user groups and randomly select four innovations each from the LLM-generated subsets. After narrowing it down to 16 total ideas, four ideas from each set, four random blocks are created which each consist of four innovative ideas, one from each group. These blocks are later randomly presented to newly gathered respondents, who evaluate the individual innovations on the basis of selected questions.

3 Empirical Analysis

To test our proposed approach, we conduct the two-part study mentioned to demonstrate the capabilities. The findings are displayed in two parts. First, we analyze the extent of the first part of the study and further describe the questions employed to evaluate the innovations in the second part of the study. Second, the results of the study are examined and analyzed.

3.1 Method of the Study

Before we could ultimately evaluate the LLM's performance, we gathered innovative ideas from 50 non-expert innovators via Prolific during the first part of the study (25 females, 25 males, $\text{age}_{\text{range}} = 19\text{–}61$). The group which is disclosed with the main review issues summarized by the LLM includes 26 non-experts. The other group without knowledge of the review contents includes 24 non-experts. We ask both groups to come up with as many innovative ideas as they want. In total, the group with disclosure of review summary comes up with 93 ideas, the comparison group without contributes 74 ideas. We manually group these ideas and use the four most mentioned ideas for the study. After identifying the most common ideas for each group of users, we create the four blocks of innovations (see fig. 2) and use them for further evaluation.

Consequently, we employ Prolific to recruit 65 participants (36 females, 29 males, $\text{age}_{\text{range}}$ = 20-70) to evaluate the innovative ideas. Participants were randomly assigned a pre-created block of ideas to work with and provide responses to a set of questions regarding the innovative ideas. To prevent biased answers, the origin of the innovative ideas was concealed, so the participants were unaware of the use of a LLM.

The evaluative questions used are a combination of Likert scale questions and one initial ranking question. This setup therefore includes a combination of single-criterion and multicriteria items, which allows for a multifaceted evaluation (Riedl et al. 2013).

After presenting the product description, we ask the respondents to state their initial purchase intent ("I can see myself buying this product in the next 12 months"). Consequently, we present the participants with the main issues derived from the reviews and the four ideas. We ask the participants to rank the ideas from 1 (best idea) to 4 (worst idea). Lastly, the respondents are asked to provide responses to four statements for the four product ideas on a Likert scale ranging from 1 to 7 regarding Liking ("I like this idea"), the perception of innovativeness of the idea ("I perceive this idea as innovative"), purchase intent ("When the product includes this idea, I can see myself buying it in the next 12 months"), and help to solve potential issues ("This idea can solve the potential issues mentioned before"). The questionnaire resulted in a total of 1,040 ratings on 16 innovative ideas.

3.2 Results

Evaluating the ranking of the ideas from each group in Table 1, we see significant differences when we compare our model to the other groups. Note that the best idea is indicated by 1 and the worst idea is indicated by 4, meaning a lower scoring indicates a higher ranking. This indicates that the ideas generated by our model are considered better ideas than all other groups. We do not find significant differences between the other groups.

Table 1: Average Ranking of Innovative Ideas

Group	**Descriptives**		**Mann-Whitney U**
	M	SD	Z
LLM with Reviews	2.18	1.08	
LLM without Reviews	2.52	1.14	1818*
Non-experts with Reviews	2.71	1.11	1600**
Non-expters without Reviews	2.59	1.11	1730*

$^*p \leq .05$; $^{**}p \leq .01$.

Table 2 shows the average ratings for each group and question. Note that the groups and results were summarized across all blocks. We can see that the ideas generated by the LLM, which is aware of all reviews on the product and the product description, performs significantly better than all other groups on the liking of the ideas. We do not find significant differences between the other groups.

Table 2: Average Ratings of Innovative Ideas

Question	Group	Descriptives		Welch's t-test
		M	SD	t
Liking	LLM with Reviews	5.50	1.10	
	LLM without Reviews	4.80	1.70	2.79**
	Non-experts with Reviews	4.80	1.48	3.07**
	Non-experts without Reviews	4.98	1.65	2.11*
Innovativeness	LLM with Reviews	4.26	1.44	
	LLM without Reviews	4.48	1.49	-0.89
	Non-experts with Reviews	4.23	1.47	0.12
	Non-experts without Reviews	4.36	1.59	-0.40
PI[1] Difference	LLM with Reviews	0.45	1.52	
	LLM without Reviews	-0.26	1.94	2.35*
	Non-experts with Reviews	0.05	1.70	1.46†
	Non-experts without Reviews	0.21	1.65	0.88
Tackle Issue	LLM with Reviews	5.30	1.34	
	LLM without Reviews	3.45	1.82	6.64**
	Non-experts with Reviews	4.73	1.44	2.38**
	Non-experts without Reviews	4.42	1.67	3.33**

†$p \leq .10$; *$p \leq .05$; **$p \leq .01$; [1]purchase intent

Regarding the perception of innovativeness of the generated ideas, we see that there are no significant differences between any group. A reason for this can be that respondents see innovative ideas which fit the mentioned issues, not as innovative. There are no significant differences between the other groups. If we look at the difference in the purchase intent of the respondents, which is calculated by subtracting the purchase intent of the product if it includes the innovative idea from the initial purchase intent of the original product, we see that our model outperforms the LLM which is not aware of the reviews and the non-experts which are aware of the main issues significantly. No other significant differences between the groups are found. Lastly, we see that our model is better in addressing the issues mentioned in the reviews than all other groups. Interestingly, we find significant differences between the LLM which is not aware of the reviews and non-experts which are aware of the main issues ($p < .01$) and between the LLM without reviews and non-experts without reviews ($p < .01$). Both non-expert groups outperform the LLM without reviews significantly. A potential reason for this is that the LMM which is not taking account the customer reviews does not have the knowledge and experience of non-experts and therefore can come up with ideas which are prone to be subjectively more creative but solve the issues less effectively.

4 Conclusion and Outlook

Product-related insights in the form of customer reviews provide a rich source of information for companies to work with. Automated analysis powered by LLM's could pose as a valuable tool in future marketing questions. In our study, we evaluate LLM's performance regarding its (1) analytical abilities and (2) capability to turn the insights into useful innovative ideas. Our results indicate that training a LLM on a set of review data is feasible and leads to promising innovative ideas. Its performance in relation to comparison groups demonstrates its ability to extract relevant information and transform them into well-perceived suggestions. This is underlined by the unbiased answers we got by not disclosing the use of AI in our study. Providing this structure as a more sophisticated tool could help companies effectively identify customer needs and tackle relevant issues occurring in the use of its products.

However, there are obvious limitations to this research as of now. While we did show that the use with non-technical products or application and its reviews is also feasible, the limitation to only one final product does not yet allow a full conclusion on its potential. Furthermore, by integrating random participants, the results may be less conclusive.

Further research needs to be done by working with actual product development teams to get practical evidence if the created model could successfully work with any set of reviews. Another aspect which has not been considered yet is the feasibility of the innovative ideas with actual production requirements.

To get answers to these questions, we suspect extensions to other domains are possible, but further training and testing would be needed to assess potential across various fields.

References

Boyd, D. and Goldenberg, J. (2019). *Inside the Box: Warum die besten Innovationen im Geschäftsleben direkt vor Ihren Füßen liegen*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Flory, L., Osei-Bryson, K.-M., and Thomas, M. (2017). A new web personalization decision-support artifact for utility-sensitive customer review analysis. *Decision Support Systems*, 94:85–96.

Hong, H., Xu, D., Wang, G. A., and Fan, W. (2017). Understanding the determinants of online review helpfulness: A meta-analytic investigation. *Decision Support Systems*, 102:1–11.

Korfiatis, N., García-Bariocanal, E., and Sánchez-Alonso, S. (2012). Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. *Electronic Commerce Research and Applications*, 11(3):205–217.

Lee, T. Y. and Bradlow, E. T. (2011). Automated marketing research using online customer reviews. *Journal of Marketing Research*, 48(5):881–894.

Liu, Y., Soroka, A., Han, L., Jian, J., and Tang, M. (2020). Cloud-based big data analytics for customer insight-driven design innovation in smes. *International Journal of Information Management*, 51:102034.

Parthasarathy, M., Rittenburg, T. L., and Dwayne Ball, A. (1995). A re-evaluation of the product innovation-decision process: the implications for product management. *Journal of Product Brand Management*, 4(4):35–47.

Riedl, C., Blohm, I., Leimeister, J. M., and Krcmar, H. (2013). The effect of rating scales on decision quality and user attitudes in online innovation communities. *International Journal of Electronic Commerce*, 17(3):7–36.

Roelen-Blasberg, T., Habel, J., and Klarmann, M. (2023). Automated inference of product attributes and their importance from user-generated content: Can we replace traditional market research? *International Journal of Research in Marketing*, 40(1):164–188.

Samha, A. K., Li, Y., and Zhang, J. (2014). Aspect-based opinion extraction from customer reviews. In *Computer Science Information Technology (CS IT)*, pages 149–160.

Tian, P. and Yang, Q. (2023). The impact of online customer reviews on product iterative innovation. *European Journal of Innovation Management*.

Toubia, O. and Netzer, O. (2017). Idea generation, creativity, and prototypicality. *Marketing Science*, 36(1):1–20.

Zhao, H., Yang, Q., and Liu, Z. (2021). Impact of online customer reviews and deep learning on product innovation empirical study on mobile applications. *Business Process Management Journal*, 27(6):1912–1925.

Zhou, S., Qiao, Z., Du, Q., Wang, G. A., Fan, W., and Yan, X. (2018). Measuring customer agility from online reviews using big data text analytics. *Journal of Management Information Systems*, 35(2):510–539.