Comparing the Ideation Quality of Humans With Generative Artificial Intelligence

Jan Joosten
Nuremberg Institute of Technology
Volker Bilgram
Nuremberg Institute of Technology
Alexander Hahn
Nuremberg Institute of Technology
Dirk Totzek
University of Passau

Acknowledgements:
The authors thank the collaborating company for their willingness to participate in the ideation phase of this study. The authors also thank the participating executive for evaluating the ideas in a real-world scenario.

Cite as:
Joosten Jan, Bilgram Volker, Hahn Alexander, Totzek Dirk (2024), Comparing the Ideation Quality of Humans With Generative Artificial Intelligence. *Proceedings of the European Marketing Academy*, 52nd, (119705)

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

Comparing the Ideation Quality of Humans With Generative Artificial Intelligence

Abstract:

Traditionally, ideating new product innovations is primarily the responsibility of marketers, engineers, and designers. However, a rapidly growing interest lies in leveraging generative artificial intelligence (AI) to brainstorm new product and service ideas. This study conducts a comparative analysis of ideas generated by human professionals and an AI system. The results of a blind expert evaluation show that AI-generated ideas score significantly higher in novelty and customer benefit, while their feasibility scores are similar to those of human ideas. Overall, AI-generated ideas comprise the majority of the top-performing ideas, while human-generated ideas scored lower than expected. The executive's emotional and cognitive reactions were controlled for potential biases and showed no differences between the idea groups. These findings suggest that, under certain circumstances, companies can benefit from integrating generative AI into their traditional idea-generation processes.

Keywords: AI-augmented innovation, idea generation, innovation

Track: Innovation Management & New Product Development

1. Introduction

The innovation landscape increasingly incorporates large language models (LLMs) with conversational interfaces as individuals and organizations leverage AI systems as problem-solving agents (Bouschery et al., 2023; Bilgram & Laarmann, 2023). Professionals in marketing, engineering, and design traditionally lead innovative idea generation (Amabile & Pratt, 2016). While valuable, this method may limit idea diversity and overlook blind spots, demanding substantial time and effort for human creativity (Ritala et al., 2021).

Generative AI models like OpenAI's ChatGPT and Google's Bard are ushering in a paradigm shift in idea generation, extending beyond analytical tasks to include creative content creation in text, image, video, and audio formats (Füller et al., 2022). Given the intentional manipulation of thinking patterns in innovative techniques, the probabilistic nature of LLMs, such as GPT-3.5, proves conducive to content generation without the need for fact-checking (Brem, 2019; Bilgram & Laarmann, 2023).

This study addresses the intersection of generative AI and innovation, examining how product ideas from human professionals compare to those generated by ChatGPT in terms of novelty, customer benefit, feasibility, and overall quality. 95 Ideas from both sources underwent blind evaluation by an experienced innovation expert. Findings indicate that AI-generated ideas excel in novelty and customer benefit, with comparable feasibility scores. The composite quality score of AI-generated ideas is significantly higher, showcasing AI's potential to expand idea diversity and enhance innovation strategies. The study emphasizes the importance of harmonious human-AI collaboration in innovation and provides a blueprint for evaluating AI's role in ideation tasks.

2. Background and Related Work

2.1 Company Innovation and Creativity

Companies historically relied on human creativity for product innovation (Amabile, 1988). The multifaceted innovation process, especially in idea generation, is crucial for bridging the gap between problem domains and desired solutions (Amabile, 1988). Professionals play a key role in this phase (von Hippel, 1988). Scholars exploring factors influencing creativity and innovation have proposed various models (Amabile, 1988; Amabile & Pratt, 2016) and frameworks like defined-phase workshops to enhance creativity and innovation (Brem, 2019).

Innovation sources expanded to users and consumers (von Hippel, 1986; von Hippel, 1988), further accelerated by external contributors through crowdsourcing, crowdfunding, and the lead user method (Acar, 2019; Brem et al., 2018; Brem et al., 2019; Füller, 2010). A pivotal study (Poetz & Schreier, 2012) found both professional and user-generated ideas to be innovative, with user-generated ideas being notably more novel and customer-beneficial, though often less feasible. Similar research supported these findings, indicating higher market success and revenues for user-generated product ideas (Nishikawa et al., 2013).

The advent of generative AI introduces a new source of innovation, prompting inquiry into how these AI systems contribute to ideation compared to human professionals.

2.1 Generative Artificial Intelligence and Creativity

Amidst increasing interest and substantial developments in generative AI, researchers have begun evaluating the creative potential of AI systems. In 2021, innovation managers perceived AI as relatively less important for idea generation than domains like data analytics. Their stance was that humans would continue outperforming AI in creativity over the next 5-10 years (Füller et al., 2022).

In 2023, Bouschery et al. examined knowledge extraction and idea generation using the GPT-3 algorithm. Their focus on integrating AI directly into human ideation sessions revealed limitations and suggested potential benefits of hybrid intelligence, where AI systems and human professionals collaborate in synergy (Bouschery et al., 2023). In contrast, this study explores independent innovation by the LLM.

LLMs, exemplified by newer iterations like GPT-3.5 and GPT-4, mark a significant milestone in generative AI, showcasing exceptional capabilities. Proficiency in tasks such as passing academic and professional exams, including the Uniform Bar Exam and SAT Math, has been demonstrated (OpenAI, 2023). The evolving capabilities of LLMs prompt intriguing questions about their creative potential in innovative ideation compared to human professionals.

3. Study Method

3.1 Overview

This section outlines the methodology employed to address the primary research question: How do new product ideas generated by a company's professionals compare to those generated by ChatGPT in terms of novelty, customer benefit, and feasibility? A cross-group blind evaluation was conducted involving an executive innovation expert who assessed each idea's novelty, customer benefit, and feasibility in a recorded session, presented in randomized order. Typeform facilitated the data recording, and iMotions Online was utilized for unconscious bias analysis. Section 3.2 will detail the ideas for assessment, while Section 3.3 will introduce the executive responsible for the evaluation. The analysis strategy is explained in Section 3.4. GDPR-compliant informed consent was obtained before any data collection.

Human-generated ideas were obtained from a company meeting specific criteria: 1) focused on innovation within their core business segment; 2) utilized internal professionals for idea generation; 3) agreed to provide their human-generated ideas for comparison with AI-generated ideas; and 4) committed to a blind evaluation by their executive innovation expert, following the three dimensions of idea quality assessment (Poetz & Schreier, 2012). The selected company: a European supplier of specialized packaging solutions operating in the food, aroma protection, pharmaceutical, and cosmetics industries, emphasizes sustainability in its innovation efforts. The company conducts design thinking workshops for employees to generate ideas, which are refined through a standardized innovation process.

3.2 Idea Generation

The company initiated an internal idea-generation process to maintain its competitive edge. Simultaneously, the authors embarked on an AI-driven idea-generation approach. Both groups were given identical task descriptions that included the goal, some company background for context, and formal guidelines of the expected outcomes, like idea length.

The company's professionals engaged in idea generation through an online brainwriting platform. This ideation session generated 65 raw ideas. 43 ideas qualified for the evaluation phase after careful review by the authors based on predefined criteria. For example, only valid ideas were included, and ironic or off-topic comments were excluded. The authors similarly reviewed the AI-generated ideas, resulting in 52 ideas selected for evaluation. The authors edited the ideas, removing clear indicators on whether the source was human or AI, to enable a blind evaluation process. For example, all ideas were adapted to refer to the company in 3rd person, and spelling errors were corrected.

3.3 Idea Evaluation

95 Ideas (43 from humans and 52 from ChatGPT) underwent evaluation by the managing director of the company's innovation unit – a pivotal decision-maker in the new product development process, possessing extensive technical and market knowledge. To minimize bias, evaluations were conducted blindly, with the idea order randomized.

The executive underwent comprehensive training on the evaluation criteria, aligning with established research procedures (Poetz & Schreier, 2012). The assessment focused on three key variables: 1) *Novelty*, gauging distinctiveness compared to existing market norms; 2) *Customer benefit*, evaluating the idea's efficacy in addressing underlying problems; and 3) *Feasibility*, assessing the ease of transforming the idea into a commercial product. A 5-point Likert scale

(1 – very low; 5 – very high) measured all three variables. The composite *quality* dimension integrated scores of novelty, customer benefit, and feasibility multiplied by each other, replicating Poetz and Schreier's formula (2012) for a holistic assessment.

Additionally, three binary variables were calculated to compare the best ideas. "Top ideas" were those rated four to five, while "other ideas" scored less than or equal to three. This classification differentiated top-performing ideas within each quality dimension from the remaining ones.

3.4 Idea Evaluation Setup

The evaluator viewed the ideas in a randomized order through the Typeform platform, which provided a seamless interface for reading and evaluating each idea while ensuring unawareness of the idea's source. The executive's evaluations were recorded via webcam and screen recordings to facilitate an ex-post analysis of emotions and attention using facial coding (affective computing) and eye tracking. These measures aim to detect cognitive or emotional biases during the evaluation process. Advancements in computing capabilities now enable the detection of human attention and emotions through cognitive algorithms (Richardson, 2020). Affective computing technologies encompass analysis of facial expressions, utilizing the Facial Action Coding System (FACS) based on the emotion model by Ekman and Friesen (1978) and El Bolock et al. (2020). This study employs the AFFECTIVA algorithm for facial coding and the iMotions WebET 3.0 algorithm for webcam eye tracking (iMotions, 2023).

3.5 User Sample

Professionals (Group 1): Seven company professionals from top to lower management levels who regularly participated in similar design thinking workshops. Five professionals were part of the company's research and development department, of which four were application engineers. The remaining two professionals were from the company's marketing department.

Generative AI Tool (Group 2): Generative AI was harnessed as the idea-generation tool. Specifically, ChatGPT (March 2023 default version with GPT-3.5) was used to generate ideas in April 2023. The AI system represents a competent generative AI technology, offering a unique perspective in comparing idea quality.

4. Analysis and Results

4.1 Evaluation Findings

First, the Mann-Whitney U Test results in Table 1 show that human ideas scored significantly lower in novelty ($M = 2.65$) than those generated by ChatGPT ($M = 3.42$; $U = 740$, $p = 0.004$). Second, human ideas scored significantly lower on customer benefit ($M = 3.07$)

than ideas generated by ChatGPT (M = 3.60; U = 741, p = 0.003). Third, there is no significant difference between the human ideas' rated feasibility (M = 2.49) and that of ChatGPT ideas (M = 2.33; U = 1250, p = 0.292). The relatively low means in comparison to novelty and customer benefit indicate that idea realization is a bottleneck for both groups.

	Human Ideas (n = 43)		**ChatGPT Ideas (n = 52)**		**U**	**(p-value)***
Dimension	**Mean**	**(SD)**	**Mean**	**(SD)**		
Novelty	2.65	(1.21)	3.42	(1.24)	740.5	(**0.004**)
Customer Benefit	3.07	(0.94)	3.60	(0.80)	741	(**0.003**)
Feasibility	2.49	(0.91)	2.33	(0.88)	1250.5	(0.292)
Quality**	18.95	(10.75)	26.75	(13.18)	734.5	(**0.004**)

* Mann-Whitney U Tests were conducted because the dependent variables have no normal distribution.
** Three-way dimension (Novelty x Customer Benefit x Feasibility)

Table 1. Average Novelty, Customer Benefit, Feasibility, and Quality Scores

Finally, the composite measure of quality shows that human ideas' overall ratings (M = 18.95) are significantly lower than ChatGPT ideas' overall ratings (M = 26.75; U = 734, p = 0.004). In a typical corporate innovation funnel, the advanced stages of the innovation process focus on the highest-performing ideas. They will undergo subsequent in-depth evaluation and potential testing. Therefore, in the next step of the analysis, ideas are split into two groups: "Top ideas," which includes all ideas scoring four or five on the 5-point Likert scale per respective dimension, and "other ideas" (see Tables 2-4).

First, 36 of the 95 ideas are considered highly novel compared to existing market norms. Of these 38%, ChatGPT generated 25, while company professionals only contributed 11. Table 2 portrays the results of a chi-squared test (χ^2(1) = 4.15, p = 0.042), showing that significantly more ChatGPT ideas and fewer human ideas were rated four or five than expected in terms of novelty (p = 0.042).

	Novelty			
	Human ideas		**ChatGPT ideas**	
	Observed	(Expected)	Observed	(Expected)
Top*	11	(16.29)	25	(19.71)
Other	32	(26.71)	27	(32.29)
Chi-2 (p-value)	4.15 (**0.042**)			

*Top ideas scored four or five per respective dimension.

Table 2. Top Ideas vs. The Rest in Terms of Novelty

Second, 40 of the 95 ideas are categorized as highly customer beneficial. Of these 42%, more were again generated by ChatGPT, with 29 ideas compared to 11. Table 3 portrays the results of a chi-squared test (χ^2(1) = 7.6, p = 0.006), showing that ChatGPT scored significantly higher than expected. In contrast, human ideas scored worse than expected in terms of customer benefit (p = 0.006).

	Customer Benefit			
	Human ideas		ChatGPT ideas	
	Observed	(Expected)	Observed	(Expected)
Top*	11	(18.11)	29	(21.89)
Other	32	(24.89)	23	(30.11)
Chi-2 (p-value)	7.6 (**0.006**)			

*Top ideas scored four or five per respective dimension.

Table 3. Top Ideas vs. The Rest in Terms of Customer Benefit

Third, 11 of the 95 ideas – only 12% of all – qualified as feasible, which is very low compared to the other two dimensions. Humans are responsible for six of these ideas and ChatGPT for five, with no significant differences between the observed and expected frequencies ($\chi^2(1) = 0.11$, $p = 0.737$) in Table 4.

	Feasibility			
	Human ideas		ChatGPT ideas	
	Observed	(Expected)	Observed	(Expected)
Top*	6	(4.98)	5	(6.02)
Other	37	(38.02)	47	(45.98)
Chi-2 (p-value)	0.11 (0.737)			

*Top ideas scored four or five per respective dimension.

Table 4. Top Ideas vs. The Rest in Terms of Feasibility

Finally, it is worth emphasizing that no idea achieved a “top” rating in all three dimensions, which aligns with the observed negative correlation of the feasibility dimension with novelty and customer benefit.

5. Discussion

5.1. Summary

This research delves into whether professionals or an AI system, ChatGPT, can generate superior ideas for innovative products regarding novelty, customer benefit, and feasibility. The study reveals that ChatGPT excels in producing ideas with greater novelty and customer benefit, with both groups scoring similarly in feasibility. This emphasizes the need for integrating generative AI into innovation processes, where AI efficiently generates ideas that humans can refine iteratively to align with the company’s capabilities and strategic focus.

While the study establishes that AI-generated ideas may surpass human-generated ones in overall quality, it highlights AI's time and cost advantages. The AI ideation process, guided by a simple prompting strategy, took only 30 minutes, and 240 minutes were spent screening the

50 raw ideas. In contrast, traditional group ideation sessions with managers can take half a day and cost between 5,000 and 20,000 Euros.

Findings suggest a hybrid ideation approach, integrating human domain knowledge with AI capabilities. Human professionals, well-versed in the company's nuances, can focus on feasibility aspects. However, specialized LLMs that are fine-tuned for specific requirements may generate more feasible ideas. The study challenges the existing literature on the potential of generative AI and marks a shift in its perceived scope and applicability in creative domains.

Despite AI's ability to generate intriguing ideas, the bottleneck lies in realization, which is evident in lower-than-anticipated feasibility scores. This proves professionals' irreplaceable role in assessing possibilities within resource boundaries. A recommended strategy involves collaborating with generative AI for diverse ideas, followed by human professionals refining top-performing concepts in a secondary review stage to foster groundbreaking innovations.

This study's theoretical contributions include comparing professional and AI-generated ideas in terms of novelty, customer benefit, and feasibility. Building on previous research, it analyzes the strengths and weaknesses of both idea-generator groups, highlighting the importance of human creativity and judgment. The study advocates for collaborative human-AI approaches in innovation processes, offering a framework for evaluating AI's potential in ideation tasks and serving as a blueprint for future experimentation in this field.

5.3 Practical Implications

This study offers insights for optimizing innovation strategies and processes in businesses. It quantitatively evaluates professional and AI-generated ideas, assessing their novelty, customer benefit, and feasibility. This data-driven evaluation enables informed decision-making about future idea selection. The findings guide integrating AI capabilities within innovation, identifying scenarios where human creativity and judgment remain indispensable. Balanced human-AI collaboration is crucial, and workflows can be developed to foster a dynamic creative environment. Organizations can consider generative AI tools like ChatGPT for more efficient and diverse idea generation. Continuous learning and adaptation of AI integration strategies are essential, as is addressing ethical considerations for responsible and ethical AI use. Resource allocation for AI implementation, training, and maintenance may be necessary for innovation strategies.

6. Limitations and Future Research

This study has identified several potential internal and external factors that may impact the validity of the research findings. As the study is based on the analysis of correlational data, it is essential to refrain from making causal inferences about the observed relationships. Employing multiple evaluators is recommended to enhance reliability and validity, fostering consistent ratings, interrater reliability, and reduced score variance, thereby strengthening future research. While the focus is on assessing professional versus AI-generated ideas, previous findings indicate that user-generated ideas can rival professional ones (Poetz & Schreier, 2012), promoting research to compare user-generated, professional-generated, and AI-generated ideas for a deeper understanding of ideation dynamics.

Recognizing AI's capacity to rapidly generate numerous ideas with minimal time and resource investment, further exploration into AI's role in subsequent creative process stages like clustering, challenging, enriching, and combining ideas is needed. Additionally, investigating tasks beyond idea generation, such as text-to-code generators facilitating prototype development, presents avenues for innovation and efficiency (Bilgram & Laarmann, 2023).

Future research could delve into the cost dimension of AI-assisted ideation, comparing costs associated with conventional formats like workshops or design sprints with AI-based processes. Time and financial investment could serve as relevant comparative dimensions. The authors propose exploring human ideation formats from a social perspective, emphasizing potential "soft" benefits like group collaboration to counter the "not-invented-here syndrome" and foster innovation momentum (Antons & Piller, 2015). Subsequent research may investigate how AI-generated ideas, without human involvement, perform in later innovation phases.

References

Acar, O. (2019). Motivations and solution appropriateness in crowdsourcing challenges for innovation. *Research Policy, 48*(8), 103716. doi:10.1016/j.respol.2018.11.010

Amabile, T. (1988). A model of creativity and innovation in organizations. *Research in Organizational Behavior, 10*(2), 123-167.

Amabile, T., & Pratt, M. (2016). The dynamic componential model of creativity and innovation in organizations: Making progress, making meaning. *Research in Organizational Behavior, 36*, 157-183. doi:10.1016/j.riob.2016.10.001

Antons, D., & Piller, F. (2015). Opening the Black Box of "Not Invented Here": Attitudes, Decision Biases, and Behavioral Consequences. *Academy of Management Perspectives, 29*(2), 193-217. doi:10.5465/amp.2013.0091

Bilgram, V., & Laarmann, F. (2023). Accelerating Innovation with Generative AI: AI-augmented Digital Prototyping and Innovation Methods. *IEEE Engineering Management Review, 51*(2), 1-5. doi:10.1109/EMR.2023.3272799

Bouschery, S., Blazevic, V., & Piller, F. (2023). Augmenting human innovation teams with artificial intelligence: Exploring transformer-based language models. *Journal of Product Innovation Management, 40*(2), 139-153. doi:10.1111/jpim.12656

Brem, A. (2019). Creativity on Demand: How to Plan and Execute Successful Innovation Workshops. *IEEE Engineering Management Review, 47*(1), 94-98. doi:10.1109/EMR.2019.2896557

Brem, A., Bilgram, V., & Gutstein, A. (2018). Involving Lead Users in Innovation: A Structured Summary of Research on the Lead User Method. *International Journal of Innovation and Technology Management, 15*(3), 1850022. doi:10.1142/S0219877018500220

Brem, A., Bilgram, V., & Marchuk, A. (2019). How crowdfunding platforms change the nature of user innovation – from problem solving to entrepreneurship. *Technological Forecasting and Social Change, 144*, 348-360. doi:10.1016/j.techfore.2017.11.020

Ekman, P., & Friesen, W. V. (1978). *PsycTESTS Dataset.* doi:10.1037/t27734-000

El Bolock, A., Abdelrahman, Y., & Abdennadher, S. (2020). *Character Computing.* Cham: Springer International Publishing. doi:10.1007/978-3-030-15954-2

Füller, J. (2010). Refining Virtual Co-Creation from a Consumer Perspective. *California Management Review, 52*(2), 98-122. doi:10.1525/cmr.2010.52.2.98

Füller, J., Hutter, K., Wahl, J., Bilgram, V., & Tekic, Z. (2022). How AI revolutionizes innovation management – Perceptions and implementation preferences of AI-based innovators. *Technological Forecasting and Social Change, 178*, 121598. doi:https://linkinghub.elsevier.com/retrieve/pii/S0040162522001305

iMotions. (2023). *iMotions WebET 3.0 Webcam Based Eye Tracking - Whitepaper v3*. Retrieved September 13, 2023, from https://go.imotions.com/WebET3Whitepaper

Nishikawa, H., Schreier, M., & Ogawa, S. (2013). User-generated versus designer-generated products: A performance assessment at Muji. *International Journal of Research in Marketing, 30*(2), 160-167. doi:10.1016/j.ijresmar.2012.09.002

OpenAI. (2023, September 13). *GPT-4 Technical Report.* Retrieved September 13, 2023, from https://cdn.openai.com/papers/gpt-4.pdf

Poetz, M., & Schreier, M. (2012). The Value of Crowdsourcing: Can Users Really Compete with Professionals in Generating New Product Ideas? *Journal of Product Innovation Management, 29*(2), 245-256. doi:10.1111/j.1540-5885.2011.00893.x

Richardson, S. (2020). Affective computing in the modern workplace. *Business Information Review, 37*(2), 78-85. doi:10.1177/0266382120930866

Ritala, P., Baiyere, A., Kraus, S., & Hughes, M. (2021). Digital strategy implementation: The role of individual entrepreneurial orientation and relational capital. *Technological Forecasting and Social Change, 171*, 120961. doi:10.1016/j.techfore.2021.120961

von Hippel, E. (1986). Lead Users: A Source of Novel Product Concepts. *Management Science, 32*(7), 791-805. doi:10.1287/mnsc.32.7.791

von Hippel, E. (1988). *The sources of innovation.* New York: Oxford University Press.