Proceedings of the Design Society, Volume 5: ICED25
https://doi.org/10.1017/pds.2025.10191

Exploring the use of LLMs to evaluate design creativity

Jiazhen Zhang✉, **Ji Han and Saeema Ahmed-Kristensen**

University of Exeter, United Kingdom

✉ j.zhang15@exeter.ac.uk

ABSTRACT: Creativity is a fundamental aspect of design that can bring us novel and useful products. However, measuring creativity in design can always be challenging as there is a lack of standardized quantification methods and the inherent limitations of mathematical modelling. Previous approaches often rely on human experts to assess design creativity. Still, humans can be subjective and biased in their evaluation procedures. Recent advancements in AI have inspired us to integrate LLMs as evaluators in engineering design. In this study, we utilize LLMs to assess the novelty and usefulness of design ideas. We developed an evaluation procedure and tested it using design samples. Experimental results demonstrate that the proposed method enhances creativity evaluation capabilities across various LLMs and improves the alignment between LLM and human expert assessments.

KEYWORDS: creativity, evaluation, designengineering, LLM, AIfordesign

1. Introduction

The design process is described as consisting of divergent and convergent thinking, this is widely recognised in the double diamond for both problem framing and idea generation (The Design Council 2005). A number of design-specific tools and methods, including computational ones, have been developed to support idea generation, such as the WordTree (Linsey et al., 2012), the 77 design heuristics (Yilmaz et al., 2016), the three-driven combinational creativity approach (Han et al., 2017), B-Link (Shi et al., 2017), the Retriever (Han et al., 2018b), Idea Inspire (Siddharth & Chakrabarti, 2018), and Pro-Explora (Obieke et al., 2023). To evaluate and select the ideas or concepts generated, the consensual assessment technique (CAT) is considered the gold standard (Amabile, 1983). Other often-used evaluation techniques involve generic ones, such as SWOT – Strength, Weakness, Opportunity and Threat and PMI – Positive, Negative, Interesting (De Bono, 2009), and design-specific ones, such as the Creative Product Semantic Scale (CPSS) (O'Quin & Besemer, 1989), the SAPPhIRE model of causality for novelty alone (Chakrabarti et al., 2005; Sarkar & Chakrabarti, 2011), and metric-based approaches (Horn & Salvendy, 2009; Shah et al., 2003) or typically involve expert assessments using novelty and usefulness as the key criteria, or the type of reasoning used (Cramer-Petersen & Ahmed-Kristensen, 2015; Keshwani et al., 2017). In comparison with idea generation, much fewer computational tools or approaches exist for supporting evaluation activities, particularly assessing the creative aspects of ideas. This is because design is usually human-oriented and can be challenging to quantify or evaluate with mathematical modelling tools. However, with the use of artificial intelligence, this situation might be changed, and AI would participate to simplify and automate the design evaluation procedures.

The recent development in generative AI, particularly Large Language Models (LLMs), has provided advanced capabilities in natural language generation, semantic understanding, reasoning, and instruction following (Jiang et al., 2024). LLMs provide an opportunity to support various design activities. Well-known LLMs include GPT-3/4 from Open AI, Gemini from Google, Llama from Meta, and Claude from Anthropic. A growing number of LLMs-based approaches and tools have been developed to support creative design activities. For example, Zhu et al. (2023) employed GPT-3 to retrieve and map biological analogies to generate bio-inspired designs in text forms. Chen, Cai, et al. (2024) used GPT-3 to encode

biological knowledge in a triple form to facilitate knowledge retrieval and mapping for supporting divergent thinking. Chen, Xiao, et al. (2024) utilised the reasoning capabilities of GPT-4 to support the interpretation of combinational designs, identifying the components that constitute the designs. Siddharth and Luo (2024) demonstrated the use of LLMs to synthesise design knowledge and generate design responses based on a knowledge base created by adopting engineering design facts. Most of these current studies focus on the use LLMs to support idea generation activities, while assessing design creativity remains underexplored. To better support designers in creative tasks, there is a need to understand if LLMs could be used for design creativity evaluation tasks.

The aim of this paper is to explore the use of LLMs for evaluating design creativity and comparing them with human evaluation results to provide useful insights on employing AI for supporting design evaluation activities. A set of prompts for evaluating novelty and usefulness is proposed. The following section reviews related studies on design creativity evaluation, including the use of computational means. Section 3 describes the methodology for developing the design creativity evaluation prompts. Section 4 presents the study exploring LLMs for evaluating design creativity and the comparison results against human evaluations. The discussion and conclusions of the paper are then presented in Section 5 and 6, respectively.

2. Related works

Creativity is "the process by which something so judged (to be creative) is produced" (Amabile, 1983), which plays a significant role in the early stages of design. It is considered a prerequisite for innovative breakthrough products (Taura & Nagai, 2017) and benefits business performance in the long term (Sarkar & Chakrabarti, 2011). Creativity evaluation is a crucial process for selecting and refining creative design ideas, which lays the foundation for product innovation. It is deemed to be a complex process, where employing experts is preferred (Cropley & Kaufman, 2019; Han et al., 2017). To support designers in creativity evaluation, several design-specific evaluation methods and approaches are proposed. As indicated in the preceding, the consensual assessment technique (CAT) (Amabile, 1983) is the gold standard in creativity evaluation, which has been widely adopted in design research. It measures the creativity of an idea or product by employing experts in the domain in question using a Likert-type scale. Creative Product Semantic Scale (CPSS) (O'Quin & Besemer, 1989) is another often-used design creativity evaluation method. It involves evaluating three dimensions: resolution (usefulness), novelty, and elaboration and synthesis, while each dimension contains a list of evaluation items on a 7-point bipolar rating scale. It is time-consuming to evaluate all 55 items in CPSS, and thereby design researchers have adapted CPSS to focus on certain aspects, such as novelty and usefulness (Keshwani et al 2017, Chulvi et al., 2012).

Several different human judgement-based criteria have been proposed for creativity evaluation, of which novelty and usefulness are the most often used measures, these are typically metric-based evaluation methods. Shah et al. (2003) proposed the use of novelty, quality, quantity and variety to measure creative idea generation effectiveness. Novelty (or originality/uniqueness) refers to newness, quality (or usefulness) refers to feasibility, quantity (or fluency) indicates the number of ideas generated, and variety indicates the number of idea categories generated. Other similar effectiveness measures of creative idea generation involve originality, flexibility, fluency and elaborations (Plucker & Makel, 2010), as well as novelty, feasibility, quantity and variety (Lopez et al., 2011). In addition to measuring creativity effectiveness, researchers also proposed several sets of metrics for measuring design creativity of ideas and concepts. For instance, novelty, usefulness and cohesiveness (Chiu & Shu, 2012); novelty, usefulness, aesthetics, and complexity (Lee et al., 2015); Originality, functionality, and aesthetics (Christensen & Ball, 2016); Novelty and quality (Srinivasan et al., 2018); uniqueness and usefulness (Starkey et al., 2019).

There is a debate on the role of surprise in design creativity (Becattini et al., 2015), and some researchers have argued that surprise should be added to reflect the unexpectedness of the design (Acar et al., 2017; Gero et al., 2019; Grace et al., 2015). Whereas, others considered surprise as a nuance of novelty (Chiu & Shu, 2012; Zheng & Miller, 2020), and a recent empirical design study has shown that surprise and novelty measure the same construct (Han et al., 2021). In this study, novelty and usefulness are considered as the two key elements for evaluating design creativity.

Relying on human evaluation is slow and expensive, and an alternative approach to evaluate design creativity is to use computational methods, in particular with the advancements of AI and the possibility to generate many more solutions. Several studies (e.g. (Han et al., 2020) and (Luo et al., 2021)) have measured design novelty through assessing the semantic distances within a concept by leveraging existing knowledge bases, of which a larger distance value represents a higher novelty degree. Wang et al. (2023) explored computational evaluation metrics for measuring design images, such as sketches, containing aspects of combinational creativity (Han et al., 2018a). Generated Image Quality Assessment (GIQA) metric has shown to have a high level of agreement with human evaluations, but such evaluation is biased due to the quality of the image. Song et al. (2023) proposed an attention-enhanced multimodal (both textual and pictorial design representations) learning model to evaluate a design concept on five metrics: drawing quality, uniqueness, elegance, usefulness, and creativity. This machine-learning approach is not as capable as humans in evaluating complex designs. Although these computational methods have automated design creativity evaluation to a certain extent, they have not been widely adopted in practice due to their limitations. Therefore, a more reliable and applicable computational approach for evaluating design creativity is needed.

The SAPPhIRE model is a causality model that explains natural and engineered systems. It describes the functionality of a product by using a set of elementary constructs involving State change, Action, Parts, Phenomenon, Input, organs, and Effect (Chakrabarti et al., 2005). With the deployments of the SAPPhIRE model, Sarkar and Chakrabarti (2011) proposed a method to evaluate novelty by analysing product functions, structures, and constructs. This approach first evaluates a design sample by the innovations in its functionality, followed by a novelty assessment of the sample structure. A more detailed evaluation of novelty in sample constructs will be given by the SAPPhlRE method, which can categorize novelty into three levels: low, medium, and high. They also introduced a technique for usefulness evaluations. It first defines a level of importance and assesses the usefulness with factors including the popularity rate, use frequency, and usage duration.

This method offers an efficient approach to evaluating design creativity by assessing both the novelty and usefulness of designs. However, its application relies heavily on the expertise of human evaluators, making it challenging to use without adequate training or a thorough understanding of its underlying constructs. In addition, the evaluation procedure can still be subjective, as humans can have a different understanding of creativity leading to varying results. Many abstract concepts in engineering design, such as assessing the quality of a product, the novelty or usefulness of a design idea, are not easily quantified through mathematical models. Thus evaluating creativity objectively in a design project still remains a challenge.

3. Methodology

An LLM is trained on a vast textual database encompassing diverse languages and cultures, making it a universal tool for natural language processing (NLP) tasks. Training data can contain expertise and knowledge from multiple domains, including engineering design. However, LLMs are also known as a ‘black box’. The interpretability of LLMs is still in the early stages of research. The output from LLMs can be challenging to predict as these are easily impacted by the input prompts. Therefore, to simulate human experts for creativity evaluation tasks with LLMs, a standardised evaluation procedure needs to be developed so that the evaluation results between human experts and LLMs can align.

To tackle the aforementioned challenges, this work integrates LLMs into the creativity evaluation process, aiming to reduce subjectivity in human assessments and deliver more objective and consistent evaluations. Each LLM can use data from different sources for training. LLMs can also be biased and make mistakes in their responses. Therefore, to avoid biases from one single model, this work brings multiple LLMs to collaborate in the creativity evaluation tasks. Specifically, we target two key aspects of design creativity: novelty and usefulness. For the novelty, inspired by the SAPPhIRE model for creativity evaluation, we developed our own evaluation procedures and presented them in the prompt so that LLMs can follow them to assess creativity in text-based design ideas. Here, we describe the evaluation procedures.

3.1. Novelty evaluation procedures

To ensure LLMs can reach the alignments with the human experts, we first present a definition of novelty as: "Novelty refers to how unique, original, or different a product or design is compared to what already exists in the market or design space."

During the evaluation phase, we focus on assessing the functionality and structure innovations of a design idea and use the following four questions for LLMs to evaluate the level of novelty.

Q1. Find and compare the function (i.e. how this works) of this product with the functions of other products. Does this function exist in any other product?

Q2. Is the new function applied to the entire product?

Q3. Find and compare the structure of this product with the structure of other products. Is the structure the same?

Q4. Is the new structure applied to the entire product?

To reduce subjectivity in evaluation outcomes, LLMs should respond with only 'YES' or 'NO' to each question. The steps and the structure of the evaluation procedures are shown in Figure 1.

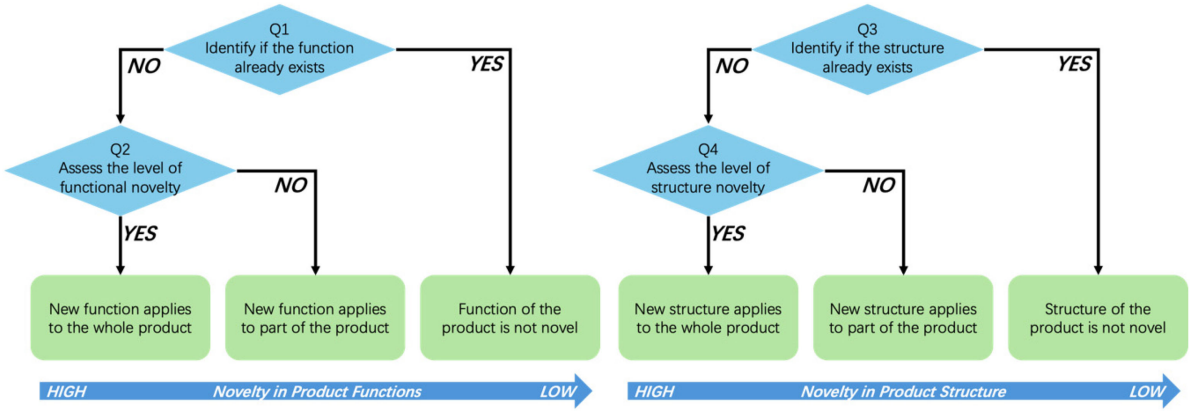

Figure 1. Structure of the evaluation procedure for novelty

The novelty is evaluated for both the functionality and for the structure (structure of the product) through two questions each. We do this by comparing against existing products, suppose the model finds that the function (Q1) or structure (Q3) of the design idea is not present in existing products. In that case, it then moves to the second and fourth questions to identify whether this innovation applies to the entire product. Otherwise, it can skip these two questions and consider the product not novel in function or structure.

Table 1. Rule table for novelty evaluation results

Q1	Q2	Q3	Q4	Result
YES	SKIP	YES	SKIP	0
YES	SKIP	NO	YES	2
YES	SKIP	NO	NO	1
NO	NO	NO	YES	5
NO	NO	NO	NO	4
NO	NO	YES	SKIP	1
NO	YES	YES	SKIP	3
NO	YES	NO	NO	5
NO	YES	NO	YES	6

After the evaluation process, LLMs can view the results from each question and use the rule table provided to return the evaluation result. As shown in Table 1, to quantify the novelty of an idea through numerical results, we use a 7-point scale for creativity evaluation. LLMs can review the questions and return a novelty score for each design idea sample.

3.2. Usefulness evaluation procedures

Same as the novelty in design, usefulness is also a critical aspect of the creativity evaluation process to ensure that design ideas have actual value and real-world applicability. To assess the usefulness of design ideas using LLMs, we first define the usefulness in our prompt as "Usefulness refers to how well the design performs its intended function and how practical or beneficial it is to the user".

Inspired by the evaluation procedure of novelty, we use a similar way to measure the level of usefulness in design ideas with the following four questions:

Q1. Evaluate the effectiveness of this product. Does the product fulfil the task requirements? Can it effectively solve the problem it is designed to address?

Q2. Does it use a more efficient way to solve the problem (compared to similar products)?

Q3. Evaluate the feasibility of this product. Is the design feasible (e.g. technically feasible)?

Q4. Is it easier to use or does it take less cost to maintain (compared to similar products)?

In this evaluation process, effectiveness and feasibility are identified as the two primary factors for assessing the usefulness of design ideas. Questions one and two focus on evaluating the effectiveness of a product design. While some design ideas may be novel and capable of fulfilling task requirements, they might also introduce unnecessary steps to fulfil the task, making the product less effective or difficult to use. On the other hand, some design ideas can satisfy the design requirements, but they can contain security vulnerabilities or high costs in manufacturing and maintenance. These kinds of design ideas can also be less useful in real-world applications. Therefore, this evaluation procedure uses questions three and four to assess the feasibility of design ideas.

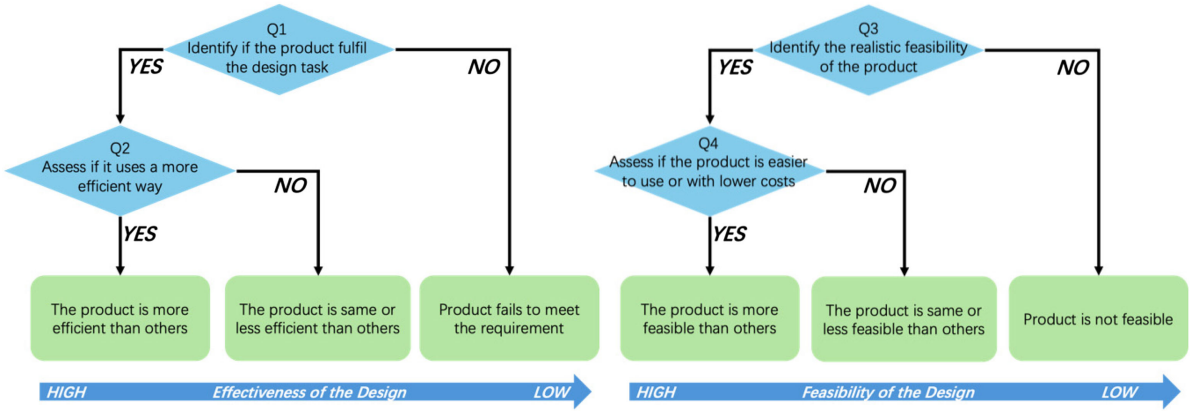

Figure 2. Structure of the evaluation procedure for usefulness

LLMs utilize these four questions to evaluate the usefulness of design ideas, following the steps and structure outlined in Figure 2. The evaluation results are recorded on a 7-point scale. LLMs assign a numerical score that determines the usefulness level of design ideas by analysing responses to questions and referencing the predefined rule table, as presented in Table 2.

Table 2. Rule table for usefulness evaluation results

Q1	Q2	Q3	Q4	Result
NO	SKIP	NO	SKIP	0
NO	SKIP	YES	NO	1
NO	SKIP	YES	YES	3
YES	NO	NO	SKIP	2
YES	NO	YES	NO	4
YES	NO	YES	YES	5
YES	YES	NO	SKIP	3
YES	YES	YES	NO	5
YES	YES	YES	YES	6

Since LLMs may lack a consistent understanding of numerical results, we also standardize the evaluation scores by providing a detailed description to each numerical value for LLMs to reference:

0 (No): The idea offers no novelty and is not useful.
1 (Very Poor): The idea shows minimal novelty or usefulness.
2 (Poor): The idea provides some novelty or usefulness but is limited and not impactful.
3 (Fair): The idea introduces moderate novelty or usefulness.
4 (Good): The idea presents considerable novelty or usefulness; it demonstrates apparent innovation or practical application.
5 (Very Good): The idea shows significant novelty or usefulness with significant new insights or efficient solutions.
6 (Excellent): The idea is highly novel or useful. It offers groundbreaking novelty and transformative practical value..

4. Experiment and results

4.1. Experiment settings

To evaluate the effectiveness of LLMs for creativity assessment, we conducted a creative design session to gather idea samples for engineering design tasks. This session involved multiple human participants with various backgrounds. The design task in this session was framed as: "Design as many novel ideas as possible for holding hot liquids for drinking." Participants are asked to use text to describe novel ideas of a product design that can fulfil this task. Since human participants come from different backgrounds, they can be unfamiliar with design. Therefore, the collected text design samples can also be varied. Some participants can provide very specific descriptions of their design ideas. For instance, an anonymous participant with a professional design background describes his/her idea as: "Container with an accelerometer to detect potential drop or spill situation and then shuts close". On the other hand, participants with non-design backgrounds may provide a more simplified description of the design, such as "insulated pouch with a straw". To compare humans and LLM, we also used ChatGPT-4 to perform the same design tasks and collected generated design ideas. In this test, a total of 78 ideas were selected for this study by random sampling, of which 37 samples were generated by humans and 41 samples were generated by ChatGPT. The creativity, more specifically novelty and usefulness, of the idea samples were then evaluated following the CAT method employing human experts.
To assess the accuracy of using LLMs for creativity evaluation, two design experts with over 10 years of experience participated in the human evaluation task voluntarily. They were provided with the evaluation instructions, measuring the novelty and usefulness of the design idea samples individually using a 7-point Likert scale. A Cronbach Alpha test was performed to analyse the evaluation rating reliability, which shows there is a "Good" (α=.80) and an "Excellent" (α=.92) reliability for novelty and usefulness measures, respectively, between the two experts. The two design experts then discussed their evaluation results and adjusted the ratings to achieve consensus agreements.
We use LLMs to assess the novelty and usefulness of design samples and compare the results provided by human experts to see if the proposed evaluation scheme can improve the alignment of LLMs. Since LLMs differ in architecture design and training data, using the same prompt across various models may produce diverse and sometimes unexpected responses. To verify the generalization of the proposed evaluation scheme, this experiment incorporates multiple LLMs for comparison. We selected ten different LLMs for evaluation, including widely used models such as GPT-4, Gemini, and Llama 3.1 and its variants. We also incorporated open-source and lightweight models, such as Gemma 2 and OpenChat 3.6, to ensure the diversity of LLMs in the experiment. Details of the LLMs we used in the experiment are listed in Table 3. All selected LLMs were provided with the evaluation procedure and tasked with independently assessing the novelty and usefulness of the idea samples. Their results were then compared with human scores to evaluate the degree of alignment. The input prompt with the evaluation procedures has the following components: definition of the evaluation task, design requirement that the samples need to fulfil, presence of results with a 7-point scale, questions for the evaluation process and the rule table. As for the baseline comparison in this test, the basic prompt is trimmed from the input prompt with only the evaluation task, design goals, and the 7-point scale. This allows LLMs to assess the idea samples based solely on their understanding of novelty and usefulness. Detailed results of the experiment are presented below.

Table 3. List of LLM model details

Model ID	Model Name	Parameters	Base Model	Open Source
0	GPT-4	1.76T	-	×
1	Gemini	Closed-Source	-	×
2	Llama 3.1	70B	Llama 2	✓
3	Qwen 2.5	72B	-	✓
4	Mixtral	8*7B	-	✓
5	WizardLM-2	8*22B	Mistral-7B	✓
6	Phi3-medium-4k	14B	-	✓
7	Gemma 2	27B	-	✓
8	Llama 3.1	405B	Llama 2	✓
9	OpenChat 3.6	8B	Llama 3	✓

4.2. Experiment results

As shown in Figure 3, each point represents the standard deviation of sample scores. Blue points indicate novelty, while orange points are for sample usefulness. A lower standard deviation implies that the sample scores across the ten LLMs are close, meaning that the models have achieved a consensus in evaluating the creativity level of this idea sample. Compared to the baseline, the standard deviations of the proposed method are lower than the baseline. The average standard deviation in sample novelty scores is 0.822 and 0.881 in usefulness, lower than the standard deviation of LLM scores when using the basic prompt (0.970 for novelty and 0.946 for usefulness).

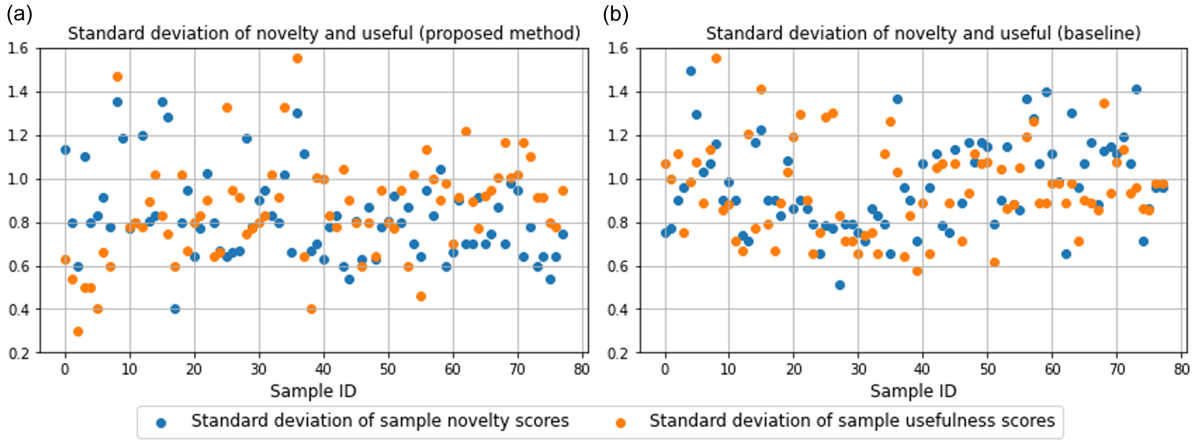

Figure 3. Standard deviation of the idea novelty and usefulness scores

Since the LLM is inherently a black box with a high level of randomness and uncertainty, its responses may not strictly align with human scores. To facilitate comparison, the 7-point scale was grouped into four categories: 0 represents no creativity in the idea sample, 1 and 2 indicate low creativity, 3 and 4 represent medium creativity, and 5 and 6 correspond to high creativity samples. If the human and LLM scores are in these four categories, their evaluation will be considered as an alignment.

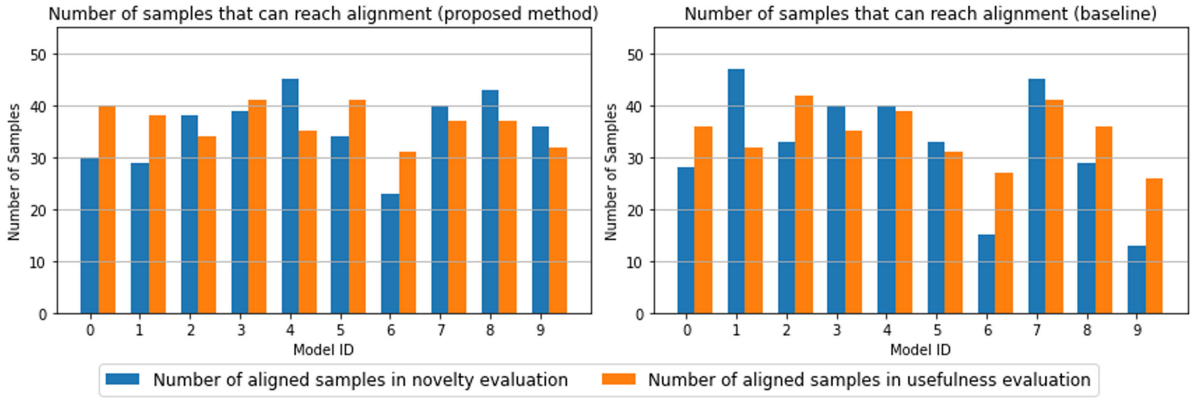

Figure 4. Number of idea results from LLMs that can reach alignments with human experts

Figure 4 illustrates the performance of LLMs in creativity evaluation tasks, comparing their alignment with human experts when using the proposed method and the baseline prompt. The basic prompt can perform well in specific models, such as Model 7 (Gemma 2), whose number of aligned samples can be over 40. This may be because this model is trained with sufficient designing data and can be suitable for creativity evaluation tasks. In contrast, the proposed method demonstrates significant improvements in most models as their responses can be better aligned with human experts. For instance, Models 6, 8, and 9 align poorly with human scores when using the baseline prompt. However, these models can follow the proposed evaluation scheme to improve performance. This indicates that the proposed method can be broadly applicable across LLMs and can generate more consistent and reliable evaluation results.

5. Discussion

To gain insight into better supporting design creativity evaluation by leveraging advanced AI capabilities, this study has explored the use of LLMs for evaluating ideas generated by both humans and ChatGPT, as well as compared the evaluation results with human experts. Compared to baseline prompts, the proposed procedure enhances the capabilities of LLMs in creativity evaluation tasks and achieves more accurate evaluation results that can align with human experts. Still, several open research issues and challenges still exist that can be addressed in the future.

In this work, we designed the evaluation scheme and provided the prompt by humans. Human experts with expertise and knowledge in design participate in the prompt writing process. Still, different evaluators can have different writing preferences and styles. These factors can potentially affect the responses from LLMs. To develop a more reliable evaluation scheme, we aim to improve our prompt by using LLMs to assist the prompt optimization procedure and enhance creativity evaluation capabilities. An evaluation procedure is introduced within our prompt to enable different LLMs to achieve alignment. However, LLMs may still produce varied results due to differences in their design knowledge and performance in evaluation tasks. A promising approach to address this variability is to consider results from multiple LLMs comprehensively. By collaborating on evaluation tasks, LLMs can generate more reliable and robust results to achieve a consensus among all participants.

The LLMs utilized in this work are built for general purposes. They can handle various text-based natural language tasks but are not specifically designed for evaluation. Additionally, the proposed method includes a step that compares the design ideas with existing works. However, some models may lack up-to-date knowledge due to their training data limitations. Therefore, we plan to fine-tune LLMs with a specialized design product database or enable web access for these models to reference the latest products and generate more trustworthy responses.

6. Conclusions

With the advancements of AI and LLM, there has been an abundance of tools to support concept generation, whereas the evaluation of ideas still primarily relies upon human evaluation. This study contributes to the use of LLM to evaluate ideas. We targeted the novelty and usefulness and proposed an evaluation procedure to standardize outputs and improve alignment. The experiment used 10 LLMs to evaluate 78 text-based design idea samples and compare the results with human experts. The initial findings demonstrate that the proposed method can improve evaluation performance, enabling LLM results to align better with human scores.

Acknowledgements

This work is funded by DIGITLab, UKRI Next Staged Digital Economy Centre (EP/T022566/1).

References

Acar, S., Burnett, C., & Cabra, J. F. (2017). Ingredients of Creativity: Originality and More. *Creativity Research Journal*, 29(2), 133-144. https://doi.org/10.1080/10400419.2017.1302776

Amabile, T. M. (1983). The Social Psychology of Creativity. Springer.

Becattini, N., Borgianni, Y., Cascini, G., & Rotini, F. (2017). Surprise and design creativity: investigating the drivers of unexpectedness. *International journal of design creativity and innovation*, 5(1-2), 29-47

Chakrabarti, A., Sarkar, P., Leelavathamma, B., & Nataraju, B. S. (2005). A functional representation for aiding biomimetic and artificial inspiration of new ideas. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 19(2), 113-132. https://doi.org/10.1017/S0890060405050109

Chen, L., Cai, Z., Jiang, Z., Luo, J., Sun, L., Childs, P., & Zuo, H. (2024). AskNatureNet: A divergent thinking tool based on bio-inspired design knowledge. *Advanced Engineering Informatics*, 62, 102593. https://doi.org/10.1016/j.aei.2024.102593

Chen, L., Xiao, S., Chen, Y., Sun, L., Childs, P. R. N., & Han, J. (2024). An artificial intelligence approach for interpreting creative combinational designs. *Journal of Engineering Design*, 1-28. https://doi.org/10.1080/09544828.2024.2377068

Chiu, I., & Shu, L. H. (2012). Investigating effects of oppositely related semantic stimuli on design concept creativity. *Journal of Engineering Design*, 23(4), 271-296. https://doi.org/10.1080/09544828.2011.603298

Christensen, B. T., & Ball, L. J. (2016). Dimensions of creative evaluation: Distinct design and reasoning strategies for aesthetic, functional and originality judgments. *Design Studies*, 45, 116-136. https://doi.org/10.1016/j.destud.2015.12.005

Chulvi, V., Sonseca, Á., Mulet, E., & Chakrabarti, A. (2012). Assessment of the Relationships Among Design Methods, Design Activities, and Creativity. *Journal of Mechanical Design*, 134(11). https://doi.org/10.1115/1.4007362

Cramer-Petersen, C. L., & Ahmed-Kristensen, S. (2015). Reasoning in Design: Idea Generation Condition Effects on Reasoning Processes and Evaluation of Ideas. *In Proceedings of the 22nd Innovation and Product Development Management Conference European Institute for Advanced Studies in Management*.

Cropley, D. H., & Kaufman, J. C. (2019). The siren song of aesthetics? Domain differences and creativity in engineering and design. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 233(2), 451-464. https://doi.org/10.1177/0954406218778311

De Bono, E. (2009). Lateral thinking: A textbook of creativity. Penguin UK.

Gero, J., Yu, R., & Wells, J. (2019). The effect of design education on creative design cognition of high school students. *International Journal of Design Creativity and Innovation*, 7(4), 196-212. https://doi.org/10.1080/21650349.2019.1628664

Grace, K., Maher, M. L., Fisher, D., & Brady, K. (2015). Data-intensive evaluation of design creativity using novelty, value, and surprise. *International Journal of Design Creativity and Innovation*, 3(3-4), 125-147. https://doi.org/10.1080/21650349.2014.943295

Han, J., Forbes, H., & Schaefer, D. (2021). An exploration of how creativity, functionality, and aesthetics are related in design. *Research in Engineering Design*, 32(3), 289-307. https://doi.org/10.1007/s00163-021-00366-9

Han, J., Forbes, H., Shi, F., Hao, J., & Schaefer, D. (2020). A DATA-DRIVEN APPROACH FOR CREATIVE CONCEPT GENERATION AND EVALUATION. *Proceedings of the Design Society: DESIGN Conference*, 1, 167-176. https://doi.org/10.1017/dsd.2020.5

Han, J., Park, D., Shi, F., Chen, L., Hua, M., & Childs, P. R. N. (2017). Three driven approaches to combinational creativity: Problem-, similarity- and inspiration-driven. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 233(2), 373-384. https://doi.org/10.1177/0954406217750189

Han, J., Shi, F., Chen, L., & Childs, P. R. N. (2018a). The Combinator – a computer-based tool for creative idea generation based on a simulation approach. *Design Science*, 4, e11, Article e11. https://doi.org/10.1017/dsj.2018.7

Han, J., Shi, F., Chen, L., & Childs, P. R. N. (2018b). A computational tool for creative idea generation based on analogical reasoning and ontology. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 32(4), 462-477. https://doi.org/10.1017/S0890060418000082

Horn, D., & Salvendy, G. (2009). Measuring consumer perception of product creativity: Impact on satisfaction and purchasability. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 19(3), 223-240. https://doi.org/10.1002/hfm.20150

Jiang, S., Xie, M., & Luo, J. (2024). Large Language Models for Combinatorial Optimization of Design Structure Matrix. *arXiv preprint arXiv:2411.12571*.

Keshwani, S., Lenau, T. A., Ahmed-Kristensen, S., & Chakrabarti, A. (2017). Comparing novelty of designs from biological-inspiration with those from brainstorming. *Journal of Engineering Design*, 28(10–12), 654–680. https://doi.org/10.1080/09544828.2017.1393504

Lee, J. H., Gu, N., & Ostwald, M. J. (2015). Creativity and parametric design? Comparing designer's cognitive approaches with assessed levels of creativity. *International Journal of Design Creativity and Innovation*, 3(2), 78-94. https://doi.org/10.1080/21650349.2014.931826

Linsey, J. S., Markman, A. B., & Wood, K. L. (2012). Design by Analogy: A Study of the WordTree Method for Problem Re-Representation. *Journal of Mechanical Design*, 134(4). https://doi.org/10.1115/1.4006145

Lopez, R., Linsey, J. S., & Smith, S. M. (2011). Characterizing the Effect of Domain Distance in Design-by-Analogy. *ASME 2011 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*,

Luo, J., Sarica, S., & Wood, K. L. (2021). Guiding data-driven design ideation by knowledge distance. *Knowledge-Based Systems*, 218, 106873. https://doi.org/10.1016/j.knosys.2021.106873

O'Quin, K., & Besemer, S. P. (1989). The development, reliability, and validity of the revised creative product semantic scale. *Creativity Research Journal*, 2(4), 267-278. https://doi.org/10.1080/10400418909534323

Obieke, C. C., Milisavljevic-Syed, J., Silva, A., & Han, J. (2023). A Computational Approach to Identifying Engineering Design Problems. *Journal of Mechanical Design*, 145(4). https://doi.org/10.1115/1.4056496

Plucker, J. A., & Makel, M. C. (2010). Assessment of Creativity. In J. C. Kaufman & R. J. Sternberg (Eds.), *The Cambridge Handbook of Creativity* (pp. 48-73). Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9780511763205.005

Sarkar, P., & Chakrabarti, A. (2011). Assessing design creativity. *Design Studies*, 32(4), 348-383. https://doi.org/10.1016/j.destud.2011.01.002

Shah, J. J., Smith, S. M., & Vargas-Hernandez, N. (2003). Metrics for measuring ideation effectiveness. *Design Studies*, 24(2), 111-134. https://doi.org/10.1016/S0142-694X(02)00034-0

Shi, F., Chen, L., Han, J., & Childs, P. (2017). A Data-Driven Text Mining and Semantic Network Analysis for Design Information Retrieval. *Journal of Mechanical Design*, 139(11). https://doi.org/10.1115/1.4037649

Siddharth, L., & Chakrabarti, A. (2018). Evaluating the impact of Idea-Inspire 4.0 on analogical transfer of concepts. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 32(4), 431-448. https://doi.org/10.1017/S0890060418000136

Siddharth, L., & Luo, J. (2024). Retrieval augmented generation using engineering design knowledge. *Knowledge-Based Systems*, 303, 112410. https://doi.org/10.1016/j.knosys.2024.112410

Song, B., Miller, S., & Ahmed, F. (2023). Attention-Enhanced Multimodal Learning for Conceptual Design Evaluations. *Journal of Mechanical Design*, 145(4). https://doi.org/10.1115/1.4056669

Srinivasan, V., Song, B., Luo, J., Subburaj, K., Elara, M. R., Blessing, L., & Wood, K. (2018). Does Analogical Distance Affect Performance of Ideation? *Journal of Mechanical Design*, 140(7). https://doi.org/10.1115/1.4040165

Starkey, E. M., Menold, J., & Miller, S. R. (2019). When Are Designers Willing to Take Risks? How Concept Creativity and Prototype Fidelity Influence Perceived Risk. *Journal of Mechanical Design*, 141(3). https://doi.org/10.1115/1.4042339

Taura, T., & Nagai, Y. (2017). Creativity in Innovation Design: the roles of intuition, synthesis, and hypothesis. *International Journal of Design Creativity and Innovation*, 5(3-4), 131-148. https://doi.org/10.1080/21650349.2017.1313132

The Design Council (2005). *The Double Diamond*. Available from: https://www.designcouncil.org.uk/our-resources/the-double-diamond/.

Wang, B., Zhu, Y., Chen, L., Liu, J., Sun, L., & Childs, P. (2023). A study of the evaluation metrics for generative images containing combinational creativity. *AI EDAM*, 37, e11, Article e11.

Yilmaz, S., Daly, S. R., Seifert, C. M., & Gonzalez, R. (2016). Evidence-based design heuristics for idea generation. Design Studies, 46, 95-124. https://doi.org/10.1016/j.destud.2016.05.001

Zheng, X., & Miller, S. R. (2020). Out in the Field Versus Inside in the Lab: A Comparison of Design Professionals' Concept Screening Practices. *Journal of Mechanical Design*, 143(4).

Zhu, Q., Zhang, X., & Luo, J. (2023). Biologically Inspired Design Concept Generation Using Generative Pre-Trained Transformers. *Journal of Mechanical Design*, 145(4). https://doi.org/10.1115/1.4056598