Journal of Research and Applications in Mechanical Engineering
ISSN: 2229-2152 (Print); 2697-424x (Online)
(2025) Vol. 13, No. 2, Paper No. JRAME 25-13-029
[DOI: 10.14456/jrame.2025.29]

Research Article

Leveraging Generative Artificial Intelligence for Prototyping: Ambidextrous Thinking in Engineering Design

K. Miura[1]
D. Misaki[2,*]
[1] *Department of Mechanical Engineering, Graduate Student, Graduate School of Engineering, Kogakuin University, Tokyo, 163-8677, Japan*
[2] *Department of Mechanical Engineering, Associate Professor, Graduate School of Engineering, Kogakuin University, Tokyo, 163-8677, Japan*

Received 31 August 2024
Revised 18 November 2024
Accepted 30 November 2024

Abstract:
Rapid economic and technological changes make the future increasingly unpredictable in the current society. Consequently, an urgent need exists to cultivate human resources capable of making significant contributions, particularly in education, where design education programs play a crucial role. In engineering design, "ambidextrous thinking" has also garnered significant attention. This study has two main objectives: (1) To develop a new evaluation method and provide guidelines for future studies by performing large language model (LLM)-based evaluations of prototyping, specifically along the axes of exploration and exploitation. (2) To quantitatively and qualitatively analyze the impact of design education on engineering. This study evaluates 31 product redesigns by third-year students enrolled in the Design Engineering course at Kogakuin University. In the LLM-based evaluation of "exploitation," 70% of the top 10 proposals suggested by ChatGPT received the highest rating from the class instructor. In addition, in the "exploration" evaluation, incorporating the concept of "darkness" into the existing definition revealed the potential for a more effective evaluation of prototypes.

Keywords: *Generative AI, Engineering Design Education, Ambidextrous Thinking, Product Redesign, Large Language Models (LLMs)*

1. Introduction

Modern society has entered an era defined by "VUCA"—volatility, uncertainty, complexity, and ambiguity—where rapid economic and technological changes make the future unpredictable. In this context, product demand is becoming increasingly complex and varied, forcing companies to innovate beyond traditional frameworks. Consequently, an urgent need exists to develop human resources capable of making significant contributions, particularly in education. Design education programs that foster creativity and innovation are being widely implemented, with particular emphasis on engineering education, where these skills are increasingly critical. Consequently, educational policies for engineering students are evolving.

In our engineering education at Kogakuin University, design education is implemented using the design thinking approach, modeled after Stanford University's program, as proposed by Leifer et al. [1]. While traditional engineering focuses on solving known problems, design engineering is more exploratory because it seeks to identify new issues and develop solutions. Specifically, engineering adopts an approach that exploits existing knowledge, whereas design engineering ventures into uncharted territories to explore new possibilities. In this context, prototyping is crucial in engineering education. This helps identify and correct design issues early, reduces development risks, enhances stakeholder communication, and contributes to innovative solutions.

* Corresponding author: D. Misaki
E-mail address: misaki@cc.kogakuin.ac.jp

"Ambidextrous thinking" has also garnered attention in engineering design. Originally from business administration, "ambidexterity" refers to the simultaneous "exploitation" of existing assets and processes to increase efficiency and "explore" new ideas and markets to adapt to changing environments. This concept applies equally to engineering design, where Faste suggests that integrating "ambidexterity" can enhance creativity and flexibility [2].

Balancing "exploitation" and "exploration" in engineering design is essential for meeting the evolving needs of society. However, achieving this balance is challenging, as mentioned by Berger-Tal et al. [3], and the assumption that design thinking is inherently complex complicates efforts to quantify and qualitatively evaluate these concepts. Consequently, clear indicators for measuring "exploration" and "exploitation" are lacking, and a need exists for practical and easily applicable evaluation methods in educational settings.

This study focused on rapidly evolving large language models (LLMs). Since OpenAI released ChatGPT in November 2022, LLMs have garnered significant attention for their applicability across various fields, particularly for their creativity and natural language generation capabilities. LLMs have been actively explored as co-design tools, particularly in creative tasks. In non-engineering fields, Yen et al. [4] explored the use of LLMs to enhance students' mathematical problem-solving abilities by providing adaptive feedback and evaluated both quantitatively and qualitatively the challenges and effectiveness of the models in recognizing errors and delivering accurate feedback. Fede et al. proposed enhancing the idea generation process in creative tasks using LLMs as co-creation partners and developing systems that automatically extend, rewrite, and combine user input with other ideas [5]. From an engineering perspective, Pierson et al. developed an interactive interface to support engineering design and optimization using LLMs and evaluated how LLMs can assist in tasks such as Python code generation and design optimization [6]. Cheung et al. developed and tested a prototyping framework that integrates LLMs with the conversation theory into a rhino–grasshopper-based architectural design environment, using verbal and non-verbal feedback, including brainwave tracking, to support and enhance the workflow of the designer [7]. Studies on the use of LLMs as evaluation tools are underway. Bossou et al. explored how machines can evaluate creative work and provide human feedback in non-engineering domains [8]. Sanuvala and Fatima examined methods that utilize optical character recognition and natural language processing techniques to grade human-written answers [9]. Xu et al. compared the performance of ChatGPT on engineering design tasks with human results, evaluating its effectiveness in technical knowledge extraction and decision-based tasks [10]. Chiarello et al. systematically analyzed the effect of generative LLMs on various stages of engineering design, particularly in evaluation tasks, and highlighted their benefits and risks [11]. Figure 1. presents an analysis of related studies. This study focused on engineering design and the use of LLMs as evaluative tools.

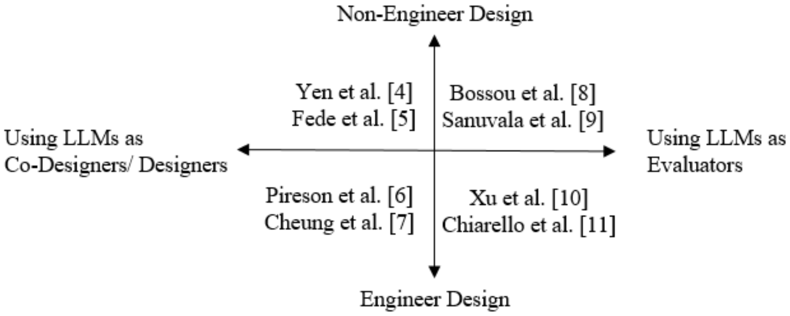


Fig. 1. Mapping of existing studies on LLMs in engineering and non-engineering design.

In this field, several studies, Freire et al. investigated the opportunities of LLM-powered cognitive assistants in manufacturing settings, focusing on the "exploitation" [12]. In contrast, researchers including Aikawa et al. have employed artificial intelligence (AI) to promote divergent thinking, aiming for "exploration" in product development [13]. However, studies that aim to integrate both approaches are scarce. For instance, Maher et al. attempted to quantitatively evaluate design creativity using three axes: "novelty," "value," and "surprise" [14]. However, these evaluation criteria are not based on ambidextrous thinking. Okamoto et al. evaluated the similarity of design documents by extracting keywords from previous documents [15]. However, this approach requires a rigorous comparison of design documents and is unsuitable for simplified evaluations such as prototype assessments in engineering education contexts. The uniqueness of this study is in the evaluation of prototyping using "exploration" and "exploitation" indicators. Further studies on the use of LLMs are required to evaluate these aspects, particularly in engineering education. Therefore, this study has two main objectives: (1) To develop a new evaluation method and provide guidelines for future studies by performing LLM-based evaluations of prototyping, specifically along the d exploration and exploitation axes. (2) To quantitatively and qualitatively analyze the impact of design education on engineering.

2. Leveraging generative AI for prototyping: Ambidextrous thinking in engineering design

2.1 Evaluation Method for Ambidextrous Thinking in Engineering Design Education

This study proposes an evaluation method for ambidextrous thinking in engineering design education that uses generative AI for prototyping. In particular, as illustrated in Figure 2., we suggest using LLMs as an evaluation tool for product design and development to address the existing challenges. The focus was on the quantitative evaluation of product design and development proposals created by students using ChatGPT.

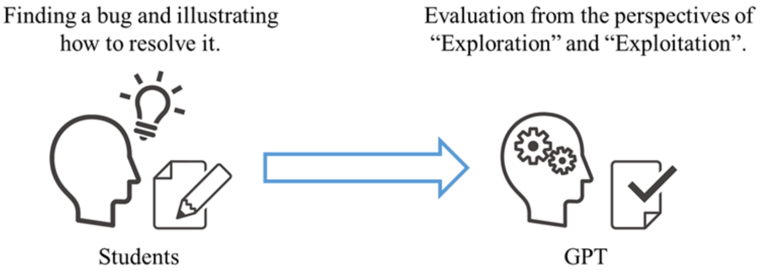

Finding a bug and illustrating how to resolve it.
Evaluation from the perspectives of "Exploration" and "Exploitation".
Students
GPT

Fig. 2. Leveraging Generative AI for Prototyping: Ambidextrous Thinking in Engineering Design.

The design engineering course at Kogakuin University integrates design thinking, an approach centered on human-centered design and the generation of creative and innovative solutions, into the engineering curriculum. This course was modeled after ME101: Visual thinking course at Stanford University, a long-standing and foundational class for undergraduate design majors [16]. In a semester-long course for third-year students in the Department of Mechanical Systems Engineering at Kogakuin University, the students learn about empathy, problem definition, ideation, prototyping, and testing related to design thinking. The final assignment challenges students to identify and solve a "bug" daily, to which they design and present new solutions.

Traditionally, only instructors evaluate the outcomes of such projects. However, the proposed evaluation method can significantly benefit both students and teachers by offering a more comprehensive assessment. To implement this evaluation method, we quantitatively and qualitatively analyzed sketches from a student's final project, "Solving Everyday Bugs through Product Redesign," using AI. ChatGPT was selected from among the various LLMs because it is currently the most widely used model, and its engineering studies are more advanced than those of other generative AI models. In this study, we created a customized GPT using the custom GPT features of ChatGPT. This feature enhances the accuracy of responses by writing prompts in advance, uploading reference files as "knowledge," and allowing the integration of third-party APIs, making it easy to create personalized GPT models. In this study, we used only the pre-prompt input feature.

2.2 Definition of "Exploration" and "Exploitation."

The context of engineering design and engineering design education can be defined as follows [17] [18]:

- Exploration: Design ways of knowing-doing-acting (Reach a "wow" performance)
- Exploitation: Engineering ways of knowing-doing-acting (Reach "thank you" performance)

In the context of design engineering, Okamoto et al. [19] defined "exploration" and "deepening" to relatively position multiple products and evaluate the novelty and creativity of a design based on design documents. One of the definitions of "exploration" and "exploitation" in product design, as used in this paper, is as shown in Figure 3.

However, using these definitions alone makes it difficult to capture the "Jamp of ideas" as an indicator of exploration. For instance, if exploration is defined as a design involving quantitative change, the mere act of generating several ideas can be evaluated as exploratory. Here, we attempt to evaluate "exploration" from the perspective of the "dark horse prototype." According to Bushnell et al. [20], "dark" in the dark horse prototype is described as follows:

- "*The prototype vision is "dark.": The prototype must explore a space that is "dark," meaning that it is risky, radical, infeasible, and/or in a direction orthogonal to previously explored solutions. The technical implementation, questionable user perception, or its departure from a plausible direction already converged on, should make the team feel uncomfortable in pursuing it as a direction.*"

By integrating the definitions by Okamoto et al. and Cutkosky et al., exploration can also be defined as follows.

- Exploration: Design involves qualitative change in a design space that is "dark," meaning that it is risky, radical, infeasible, and/or in a direction orthogonal to previously explored solutions.

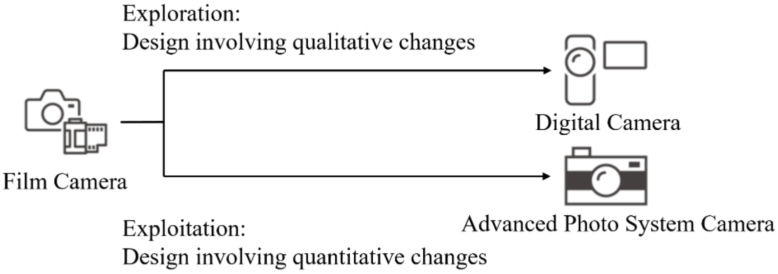
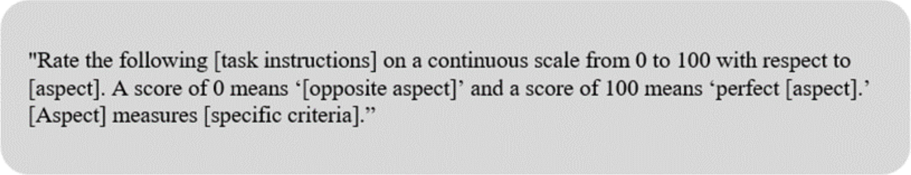

Exploration:
Design involving qualitative changes
Digital Camera
Film Camera
Advanced Photo System Camera
Exploitation:
Design involving quantitative changes

Fig. 3. Exploration and Exploitation of product design [19].

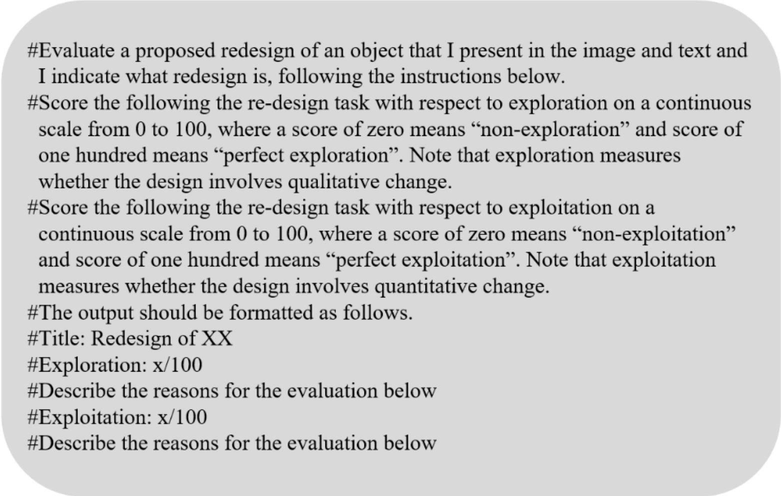
2.3 Method to Evaluate Using "DA Prompt"

To evaluate "exploration" and "exploitation" using ChatGPT, we applied "direct assessment (DA) prompt," drawing on the work of Wang et al. and Kocmi et al. DA prompts are designed to quantitatively evaluate a specific quality on a scale of 0–100, where the score reflects the degree of meaning retention and grammatical accuracy [21] [22]. In this study, this prompt was used to assess the generative quality of ChatGPT. Using this prompt, we can evaluate product design proposals with finer granularity instead of a simple binary evaluation of good or bad. The DA prompt is shown in Figure 4.



"Rate the following [task instructions] on a continuous scale from 0 to 100 with respect to [aspect]. A score of 0 means '[opposite aspect]' and a score of 100 means 'perfect [aspect].' [Aspect] measures [specific criteria]."

Fig. 4. DA prompt [21] [22].



```
#Evaluate a proposed redesign of an object that I present in the image and text and
  I indicate what redesign is, following the instructions below.
#Score the following the re-design task with respect to exploration on a continuous
  scale from 0 to 100, where a score of zero means "non-exploration" and score of
  one hundred means "perfect exploration". Note that exploration measures
  whether the design involves qualitative change.
#Score the following the re-design task with respect to exploitation on a
  continuous scale from 0 to 100, where a score of zero means "non-exploitation"
  and score of one hundred means "perfect exploitation". Note that exploitation
  measures whether the design involves quantitative change.
#The output should be formatted as follows.
#Title: Redesign of XX
#Exploration: x/100
#Describe the reasons for the evaluation below
#Exploitation: x/100
#Describe the reasons for the evaluation below
```

Fig. 5. Prompt used in the "redesign evaluation GPT."

Based on the prompt shown in Figure 4, we proposed a GPT using ChatGPT to evaluate redesign proposals that consider general ambidextrous thinking in engineering design described by Okamoto et al. in Section 2.2, as shown in Figure 5. This setup enabled the presentation of images of the redesign, accompanying descriptive text, and details of the proposed redesign. This allowed a quantitative evaluation of the redesigns and provided the underlying logic that led to each evaluation.

3. Experiments

3.1 Using ChatGPT to Evaluate Students' Prototypes

To evaluate the proposed methodology, 31 product redesign proposals addressing everyday problems sketched by engineering students in design engineering classes at Kogakuin University from 2017 to 2019 and in 2022 were analyzed. The "redesign proposal evaluation GPT" shown in Figure 5. was used to input the one- to two-page sketches prepared as the final assignment of individual design projects of "bug" fix design in the design engineering class at our university, along with a description of the redesign proposal. Third-year students enrolled in the design engineering course at Kogakuin University provided the final assignment as an experimental dataset, following the process shown in Figure 6.

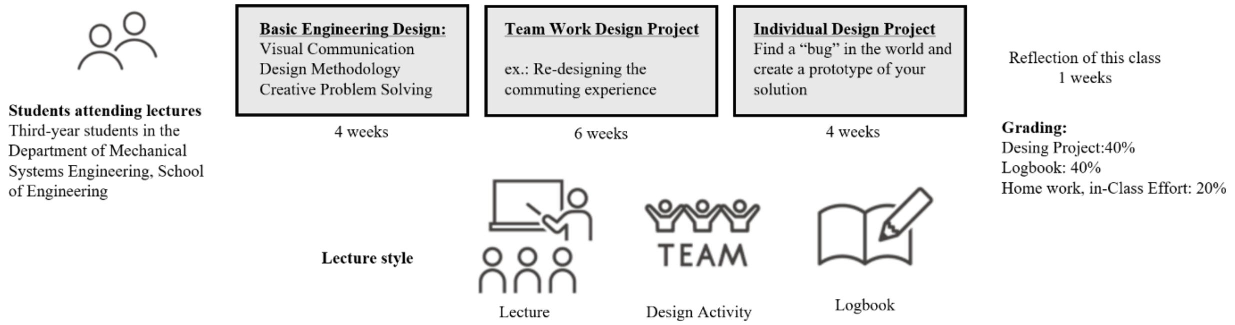
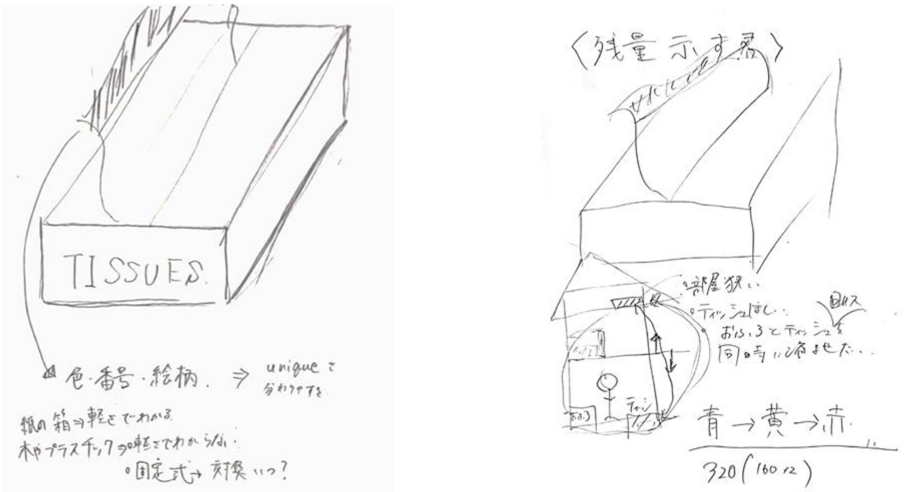

Students attending lectures
Third-year students in the Department of Mechanical Systems Engineering, School of Engineering
Basic Engineering Design:
Visual Communication
Design Methodology
Creative Problem Solving
4 weeks
Team Work Design Project
ex.: Re-designing the commuting experience
6 weeks
Individual Design Project
Find a "bug" in the world and create a prototype of your solution
4 weeks
Reflection of this class
1 weeks
Grading:
Desing Project:40%
Logbook: 40%
Home work, in-Class Effort: 20%
Lecture style
Lecture
TEAM
Design Activity
Logbook

Fig. 6. Typical lecture content of the Design Engineering course at Kogakuin University.

Each task was performed thrice to account for variations in the ChatGPT output, and the average score was calculated from these iterations. To prevent bias from the order of responses, memory was disabled, and a new ChatGPT session was initiated for each evaluation. Additionally, when ChatGPT presented two responses and a preferred answer was requested, the data were considered invalid and the evaluation was performed again. The GPT version used in this study was 4.



(a) Styling design of the tissue box.

(b) Use of tissue box as a new solution.

Fig. 7. Redesign proposal was conceived by a design class student: tissue box described in Japanese.

Figure 7. shows the idea conceived by the students. Figure 7. (a) aims to create a unique appearance by adding colors, numbers, and patterns. This suggests the use of lightweight materials in box construction. Figure 7. (b) proposes specific use cases, introducing a mechanism in which the color changes from blue to yellow to red as the number of remaining tissues decreases, making it easy to determine the remaining quantity at a glance.

Additionally, when ChatGPT presented two responses and a preferred answer was requested, the data were considered invalid and the evaluation was performed again.

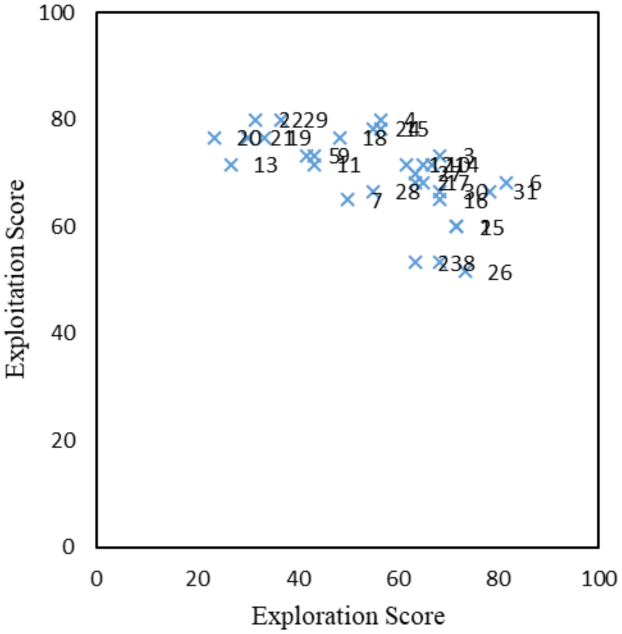


Fig. 8. Average scores for exploration and exploitation by using proposal method.

In the exploration evaluation, redesign proposals that applied other technologies to existing products and integrated them, such as Proposal No.38, which suggested adding an automatic age verification function to cameras in convenience stores, and Proposal No.6, which suggested redesigning escalators by incorporating multiple ideas, received high scores. Conversely, proposals that focused on extending or improving existing products, such as Proposal No.13, which suggested creating a case to suppress vibrations in washing machines, tended to score lower. In the exploitation evaluation, ideas that effectively utilized existing technology and design, such as Proposal No.22, which suggested changing the shape of the ruler to make it easier to pick up or design a charging cable with a fixed plug, in No.29, were rated highly. On the other hand, proposals that needed more technical considerations in our proposed evaluation method received lower scores. For example, the student's proposal No. 26, a shoulder-mounted umbrella design, received a low score. Because ChatGPT identified this as an issue related to attachment mechanisms and stability. ChatGPT evaluated this design: “The design does not delve into specific mechanisms for how the umbrella would attach to the shoulder or how stability and comfort would be maintained.” Human evaluation also identified these points as potential design flaws, supporting the chat GPT's conclusions. As shown in Figure 8, a few ideas scored highly for both “exploration” and “exploitation.” Additionally, ChatGPT occasionally misinterpreted the text in the logbooks. For instance, it misread a note indicating “145°” as “195°.”

3.2 Comparison Between ChatGPT's Evaluation and the Evaluation by Course Instructor

We conducted a study to investigate the differences between evaluations made by ChatGPT and those made by educators. In this study, the course instructor and associate professor Daigo Misaki, who was unaware of the evaluation by ChatGPT, evaluated the “redesign of a product to solve everyday problems” by assigning it a score on a three-point scale for each of the exploration and use stages.

When comparing the evaluations of the top 10 proposals suggested by ChatGPT for exploration evaluation, the proportion of proposals that received the highest evaluation from the course instructor was 10%. In contrast, of the top 10 proposals suggested by ChatGPT for deepening evaluation, the percentage of proposals that received the highest evaluation from the course instructor was 70%. In contrast, for the four ideas that the course instructor assigned the highest rating in “exploration,” ChatGPT ranked them 6th, 19th, 25th, and 29th.

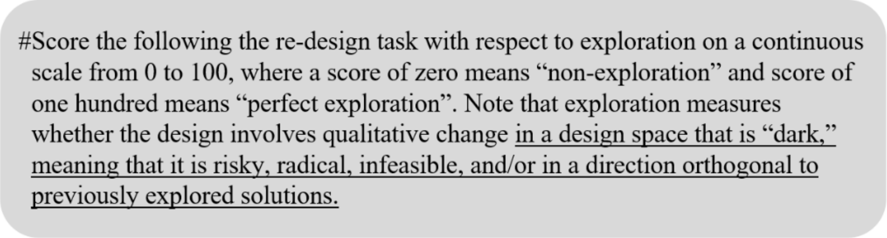
From the course instructor's comments, the results of the “deepening” evaluation were consistent with his own evaluation. It was suggested that this was because both focused on the dispersion of keywords.

From the course instructor 's comments, a divergence was observed in the evaluation of “exploration.” For instance, the course instructor assigned a high evaluation to the redesign proposal of the clear file, where the solution to the issue of corner folding was to round them such that it would be acceptable if they folded. This solution went beyond conventional thinking by addressing the problem in a manner that made folding acceptable. In contrast, ChatGPT assigned a score of 41.7, as the basic functionality and materials of the product did not change.

Additionally, in the evaluation of “exploration” in the redesign of the escalator, ChatGPT focused on the number of ideas and assigned a high score of 81.7, but the course instructor highlighted that a lack of “jump” was observed in the ideas and refrained from assigning the highest score.

3.3 Considering from the perspective of “dark”

Based on the teacher’s comments, we realized that the evaluation of “exploration” lacked sufficient assessment of idea jumps. Therefore, we decided to use the definition of “exploration” that considers “darkness” from Section 2.3. The prompt was created by adding the underlined part to the prompt shown in Figure 5, as shown in Figure 9.



#Score the following the re-design task with respect to exploration on a continuous scale from 0 to 100, where a score of zero means “non-exploration” and score of one hundred means “perfect exploration”. Note that exploration measures whether the design involves qualitative change in a design space that is “dark,” meaning that it is risky, radical, infeasible, and/or in a direction orthogonal to previously explored solutions.

Fig. 9. Prompt for evaluating exploration from perspective of being “dark.”

We evaluated a dataset similar to that in Section 3.1 and investigated the change in the exploration score when a prompt evaluation of darkness was input. The results are shown in Figure 10, and the redesign numbers correspond to those shown in Figure 8.

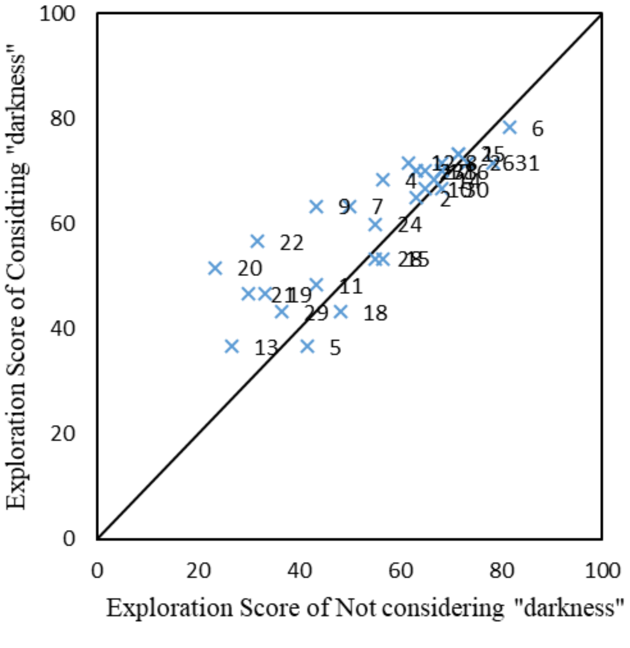


Fig. 10. Comparison of exploration scores with and without the evaluation of “darkness.”

In the following section, we evaluate the cases in which differences were observed in the evaluation of exploration for each evaluation method. Owing to the difficulty in grasping a ruler (No. 22) placed on a desk, the idea of rounding the corners of the ruler and making it easier to grasp resulted in a 25.0 point difference in rating. When “dark” was considered, rounding the corners was praised for improving usability and incorporating ergonomic elements, receiving high marks as a subtle but innovative change. In contrast, when “dark” was not considered, the minimal change in shape and lack of impact on the primary function of the ruler led to a lower score.

Similarly, hands-free, foot-operated toilet door No.20 exhibited a 23-point difference in the rating. When “darkness” was considered, the concept, while not entirely novel, was moderately rated for providing a practical method of opening doors without using hands. In contrast, when “darkness” was not evaluated, a change in interaction was acknowledged; however, limited functional improvement resulted in a lower score. The redesign of transparent file No.5, which was highly praised by the course instructor in Section 3.2, received a low score of 36.7 points using this evaluation method.

4. Discussions

Notably, LLMs such as the ChatGPT used in this study are associated with uncertainty in the evaluation. Specifically, LLMs occasionally struggle to make accurate judgments when evaluating design concepts such as “exploration” and “utilization.” For instance, even if a prototype includes an exploratory element that tests new functions, the LLM may misinterpret it as an extension of existing functions and incorrectly assess it as “utilization.” These misunderstandings are likely owing to the LLM's incomplete grasp of the evaluation criteria, which can affect the consistency of its evaluations. Although the AI focused on the number of ideas in the exploration evaluation, the course instructor prioritized whether the ideas exhibited significant “leaps.” In this study, “exploration” and “utilization” were defined as “design with qualitative change” and “design with quantitative change,” respectively. To better assess “exploration,” it is necessary to establish the criteria for evaluating idea leaps more accurately. Enhancing the prompt design to enable a more precise interpretation by an LLM is also a future challenge.

Additionally, superficial evaluations such as prototype assessments can be problematic when characters are misread. Moreover, achieving a balance between “utilization” and “exploration” is not always essential for product development. The ideal balance between these scores may vary, depending on the purpose of the product. Although this method is effective for certain types of design problems, it may be less precise than the traditional expert evaluations for tasks that emphasize technical specifications or require advanced expertise.
The method used in this study allowed us to investigate the impact of engineering education on students. For instance, by comparing the results with those of first-year undergraduates, the extent to which university-level engineering education is reflected in the design process can be observed.

In this experiment, each student redesigned a different object, contributing to variations in the “utilization” and “exploration” scores. Thus, LLM-based evaluations should currently be considered methods that are limited to specific conditions. In the future, it will be necessary to design LLMs that account for these factors during evaluation. The correlation coefficient determined in the Section 3.1 experimental results was -0.602, indicating a negative correlation between “exploration” and “exploitation” in product redesign proposals. This result quantitatively demonstrates that students tend to trade exploration and exploitation in their redesign thinking. Overcoming this trade-off will be an essential focus for future studies. Students and ChatGPT did not interact in this experiment, which resulted in a one-sided evaluation. Therefore, we aim to develop an interactive system.

Additionally, the experimental results using the prompt that considered darkness in Section 3.3 revealed that the “exploration” evaluation was closer to the course instructor's assessment than the prompt used in Section 3.1, which did not consider darkness. However, because it cannot be assumed that the evaluations fully matched those of the course instructor, a more refined and prompt design is needed.

Although this method may not be precise in the evaluations performed by professional designers or course instructors, the evaluation approach presented in this study, particularly for measuring the extent of idea exploration, can serve as a valuable guideline for assessing design proposals.

Additionally, course instructor evaluations may be influenced by external factors. In class, students presented their drawings after completing them. It cannot be ruled out that the high evaluations provided during these presentations may have affected the assessment in this experiment, which was based solely on drawings. Therefore, it is necessary to perform a more comprehensive evaluation involving more designers and educators.

5. Conclusion

This study performed quantitative and qualitative analyses to develop a new evaluation method using LLMs for prototyping, with a specific focus on the axes of exploration and exploitation. The goal was to establish guidelines for future applications and assess the impact of design education on the engineering domain. The analysis demonstrated that ChatGPT was effective in evaluating both exploration and exploitation.

Acknowledgments

This study was partially supported by JSPS KAKENHI (Grant Number: 23 K 02796.)

References

[1] Leifer L, Steinert M. Dancing with ambiguity: Causality behavior, design thinking, and triple-loop-learning. Inf Knowl Syst Manag. 2011;10(1–4):151–173.

[2] Faste R. Ambidextrous thinking. Innov Mech Eng Curric 1990s. 1994.

[3] Berger-Tal O, Nathan J, Meron E, Saltz D. The exploration-exploitation dilemma: a multidisciplinary framework. PLoS One. 2014;9(4):e95693.

[4] Yen AZ, Hsu WL. Three questions concerning the use of large language models to facilitate mathematics learning. Conf Empir Methods Nat Lang Process. 2023.

[5] Fede G, Rocchesso D, Dow S, Andolina S. The idea machine: LLM-based expansion, rewriting, combination, and suggestion of ideas. Proc 14th Conf Creativity Cognition. 2022 Jun;623–627.

[6] Pierson KC, Ha MJ. Usage of ChatGPT for engineering design and analysis tool development. AIAA SciTech 2024 Forum. 2024;0914.

[7] Cheung LH, Dall'Asta JC, Di Marco G. Exploring large language model as a design partner through verbal and non-verbal conversation in architectural design process. Proc SIGraDi 2023 Conf. 2023 Nov 20–24; São Paulo, Brazil. São Paulo: Blucher; 2024. p. 1049–1060.

[8] Bossou K, Ackerman M. Should machines evaluate us? Opportunities and challenges. International Conference Innov Comput Cloud Comput. 2021.

[9] Sanwuala G, Fatima SS. A study of automated evaluation of student’s examination paper using machine learning techniques. International Conference Comput Commun Intell Syst (ICCCIS). 2021;1049–1054.

[10] Xu W, Kotecha MC, McAdams DA. How good is ChatGPT? An exploratory study on ChatGPT’s performance in engineering design tasks and subjective decision-making. Proc Des Soc. 2024;2307–2316.

[11] Chiarello F, Barandoni S, Majda Škec M, Fantoni G. Generative large language models in engineering design: Opportunities and challenges. Proc Des Soc. 2024;4:1959–1968.

[12] Freire SK, Foosherian M, Wang C, Niforatos E. Harnessing large language models for cognitive assistants in factories. 5th International Conference Convers User Interfaces; 2023 Jul 19–21; Eindhoven, Netherlands. New York: ACM; 2023. Article 44, p. 1–6.

[13] Aikawa Y, Tamura R, Xu C, Ge X, Misaki D. Introducing augmented Post-it: an AR prototype for engaging body movements in online GPT-supported brainstorming. Adjunct Proc 36th ACM Symp User Interface Softw Technol (UIST '23 Adjunct). 2023. Article 6, p. 1–3.

[14] Maher ML, Fisher DH. Using AI to evaluate creative designs. In: Duffy A, Nagai Y, Taura T, eds. DS 73-1 Proc 2nd International Conference Design Creativity Vol 1. 2012. p. 45–54.

[15] Okamoto M, Murakami T. Proposal of defining exploration and exploitation in engineering design and evaluating the degree of exploration by natural language processing. Proc ASME 2022 Int Des Eng Tech Conf Comput Inf Eng Conf. Vol 6: 34th International Conference Des Theory Methodol (DTM); 2022 Aug 14–17; St. Louis, MO, USA. ASME; 2022.

[16] Leifer L, Steinert M. Dancing with ambiguity: Causality behavior, design thinking, and triple-loop-learning. Inf Knowl Syst Manag. 2011;10(1–4):151–173.

[17] Ge X, Leifer L. Design thinking at the core: Learn new ways of thinking and doing by reframing. Proc Am Soc Mech Eng. 2017.

[18] Lande M. Catalysts for design thinking and engineering thinking: Fostering ambidextrous mindsets for innovation. Int J Eng Educ. 2016;32(3):1356–1363.

[19] Okamoto M, Murakami T. Proposal of cluster analysis method for products considering exploration and exploitation in engineering design. Proc Des Soc. 2023;2995–3004.

[20] Bushnell T, Steber S, Matta A, Cutkosky M, Leifer L. Using a ‘dark horse’ prototype to manage innovative teams. 3rd International Conference Integr Des Eng Manag Innov; 2013 Sep; Delft, Netherlands. p. 8.

[21] Wang J, Liang Y, Meng F, Sun Z, Shi H, Li Z, Xu J, Qu J, Zhou J. Is ChatGPT a good NLG evaluator? A preliminary study. Proc 4th New Frontiers Summarization Workshop. 2023;1–11.

[22] Kocmi T, Federmann C. Large language models are state-of-the-art evaluators of translation quality. Proc 24th Annu Conf Eur Assoc Mach Transl. 2023 Jun 12–16; Tampere, Finland. Tampere: EAMT; 2023. p. 193–203.