Do Large Language Models Produce Diverse Design Concepts? A Comparative Study with Human Crowdsourced Solutions

Kevin Ma
Department of Mechanical Engineering,
University of California, Berkeley,
Berkeley, CA 94720
kevinma1515@berkeley.edu

Daniele Grandi
Autodesk Research
San Francisco, CA, 94111
daniele.grandi@autodesk.com

Christopher McComb
Department of Mechanical Engineering,
Carnegie Mellon University,
Pittsburgh, PA 15289
ccm@cmu.edu

Kosa Goucher-Lambert[1]
Department of Mechanical Engineering,
University of California, Berkeley,
Berkeley, CA 94720
kosa@berkeley.edu

Access to large amounts of diverse design solutions can support designers during the early stage of the design process. In this paper, we explored the efficacy of large language models (LLM) in producing diverse design solutions, investigating the level of impact that parameter tuning and various prompt engineering techniques can have on the diversity of LLM-generated design solutions. Specifically, we used an LLM (GPT-4) to generate a total of 4,000 design solutions across five distinct design topics, eight combinations of parameters, and eight different types of prompt engineering techniques, leading to 50 LLM-generated solutions for each combination of method and design topic. Those LLM-generated design solutions were compared against 100 human-crowdsourced solutions in each design topic using the same set of diversity metrics. Results indicated that, across the five design topics tested, human-generated solutions consistently have greater diversity scores. Using a post hoc logistic regression analysis we also found that there is a meaningful semantic divide between humans and LLM-generated solutions in some design topics, but not in others. Taken together, these results contribute to the understanding of LLMs' capabilities and limitations in generating a large volume of diverse design solutions and offer insights for future research that leverages LLMs to generate diverse design solutions for a broad range of design tasks (e.g., inspirational stimuli).

Keywords: large language models, concept design generation

1 Introduction

Inspirational stimuli have been widely shown to support designers during the early stage design process by serving as a catalyst for creativity and innovation [1–3]. Among the various methods employed to elicit such stimuli, the use of design examples has proven to be particularly effective [4]. In the past, studies have explored the use of crowdsourcing to retrieve these design examples by leveraging the collective intelligence and diverse perspectives of a large number of individuals to generate a large set of design examples [5,6]. However, with the recent advances in large language models (LLMs), there has been an increased interest in exploring how LLMs can be used to generate candidate design solutions [7,8].

Recent advancements in LLMs (e.g., GPT-4) have opened new avenues for research into their application within the design domain. Through the use of prompt engineering techniques, researchers have demonstrated that LLMs have the capability to produce design solutions that are similar to crowdsourced human solutions [7]. Despite this potential, solutions generated by LLMs are often less diverse than human-generated solutions, which poses a significant challenge given the importance on novelty and diversity in the context of inspirational stimuli [7,9]. Therefore, it is essential to identify methods for generating diverse outputs from LLMs if they are to be used as sources of inspiration.

Thus, our research was guided by two primary research questions:

(1) How do parameters that tune LLMs affect the output diversity of the design solutions?

(2) How do different prompt engineering techniques impact the output diversity of the design examples?

In this paper, we explored the use of an LLM, in our case GPT-4, to generate a diverse set of design solutions. We first investigated whether systematically varying several parameters within the LLM would affect the diversity of the generated design solutions. Then we explored whether or not different prompt engineering techniques could enhance the diversity of the generated design solutions. Our research then involved a comparative analysis of these LLM-generated solutions with those obtained from crowdsourcing platforms, in our case Amazon Mechanical Turk, across five distinct design topics.

We then evaluated the diversity of the generated design solutions through a comprehensive set of computational metrics, comparing various methods of prompt engineering techniques and parameters. Due to initial findings discussed in this paper, our team hypothesized that there may be distinguishable semantic differences between the human and LLM generated design solutions. To investigate this hypothesis, we used logistic regression to analyze whether this was true. The outcomes of this analysis and their implications are discussed in this paper.

2 Background

2.1 Generating Design Examples for Inspirational Stimuli.

During the early stages of design, overcoming design fixation and enhancing creative outcomes are critical, leading to a large body of research on generating design solutions to be used as inspirational stimuli [10]. In particular, crowdsourcing has emerged as a relevant method for gathering diverse design ideas by leveraging the collective creativity of a distributed group of individuals [11].

[1]Corresponding Author.
2024

Copyright © 2026 by ASME

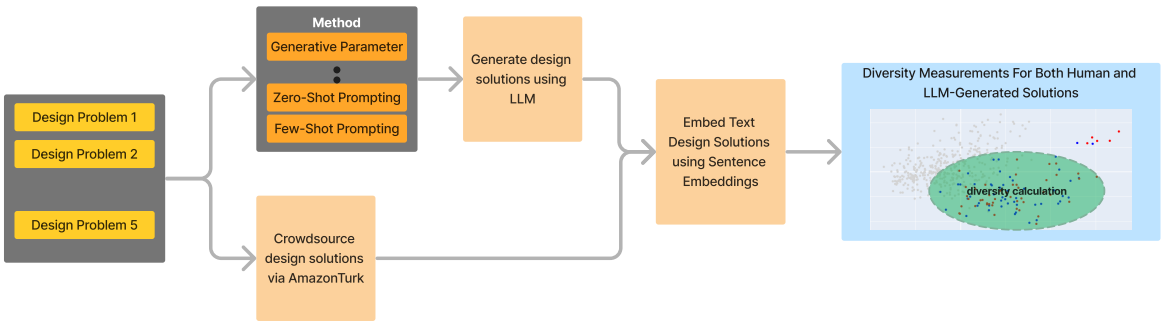

Fig. 1 Our overall objective is to better understand LLMs' ability to generate diverse design solutions – tested across a range of design problems and LLM input parameters. For each design topic, we generated 800 total design solutions using GPT-4 across different generative parameters and different prompt engineering techniques. For each design topic, we retrieved 100 design solutions via crowdsourcing. All the solutions were then converted into vector embeddings, which were used to measure diversity for quantitative comparisons. This was conducted 5 times across 5 different design problems, leading to a total of 4000 design solutions generated by an LLM and 500 design solutions retrieved via crowdsourcing.

Tools have been developed to harness crowdsourcing for retrieving analogical ideas, with empirical evidence suggesting a positive impact on creativity and ideation [6,12]. In addition, platforms like Amazon Mechanical Turk have proven effective in quickly generating large arrays of design solutions for inspiration [5]. However, crowdsourcing faces challenges, such as the lack of specialized skills among crowd workers, which can lead to impractical solutions being suggested [5]. Recent advancements in LLMs, particularly those using transformer architectures, have shown promise in generating high quality design solutions for inspirational stimuli [13]. Our paper focuses on using GPT-4, a model by OpenAI, which has demonstrated human-level performance across various benchmarks, exploring its applications within the design domain [14].

Initial studies employing generative transformers, such as GPT-3.5 and GPT-2, have shown that these models are capable of generating novel and useful design concepts [15]. The potential of LLMs has also been extended to the generation of biologically inspired design concepts, and recent studies have begun integrating established design frameworks, such as the Function-Behavior-Structure (FBS) framework, into LLMs [8,16]. In addition, our comparative study between human crowdsourced and LLM-generated solutions found that while LLMs can mimic human-like solutions, human participants often provide more novel and diverse ideas [7]. This highlights a limitation in LLMs because, in general, we strive for design examples that are diverse and novel to support designers during the ideation process [17]. Our study builds on insights from prior work on LLMs by exploring the impact a large range of prompt engineering methods and LLM parameters has on the quality of generated solutions.

2.2 Large Language Models. In this section, we present a short background regarding LLMs, prompt engineering for LLMs, and its associated parameters that can be fine-tuned.

2.2.1 Pre-Trained Large Language Models. The development of transformer-based architectures and advancements in computational power have enabled LLMs to be trained on large datasets, allowing them to emulate human-like responses and reasoning ability [13,18,19]. Numerous pre-trained LLMs exist, each with varying text generation capabilities [14,20,21]. Generally, there is a positive correlation between a model's size, measured by its parameters, and the accuracy of its output [13]. Thus, we selected GPT-4, a language model developed by OpenAI with a purported model size of over 1 trillion parameters, for our experimental framework. We chose GPT-4 primarily for this reason, as well as the model's popularity and its reported accuracy and performance at the time of writing [14].

2.2.2 Prompt Engineering and Finetuning Parameters. Recent advancements in LLMs have led to the emergence of prompt engineering as a technique, which involves manipulating LLM input prompts to enhance the accuracy and quality of the output. Within this field, there exists few-shot prompting, which involves providing examples to LLMs as inputs to improve the content accuracy, which is in contrast to zero-shot prompting whereby the LLM is queried without examples [22]. Likewise, research has shown that minor modifications to zero-shot prompts like adding the text *"You are an expert"* can enhance outputs similar to few-shot prompting [23]. Our study explores various prompt engineering strategies and their effects on the diversity of generated design solutions. In addition to prompt engineering, we also investigated other parameters directly linked to the LLMs, which in our model of choice, GPT-4, include temperature and top-P.

Temperature in GPT-4 controls the randomness of text generation, with lower settings favoring highly probable text and higher settings increasing variability [14]. Likewise, Top-P determines the breadth of text consideration based on cumulative probability, with lower values restricting selection to highly probable text and higher values allowing more variability [14]. Despite being specific to GPT-4, we argue that studying the impact of these parameters on the generated output diversity is nonetheless crucial. One of the less explored areas in current research surrounding LLMs is the impact that fine-tuning such parameters can have on the LLM's generated output. Our paper contributes to this field by assessing whether adjustments to these parameters significantly influence output generation. We argue that a comprehensive understanding of these effects is essential for designers, particularly when calibrating LLMs to support them during the design process.

2.3 Measuring Diversity for a Set of Designs. In our paper, we aimed to quantify the diversity of design solutions generated by both crowdsourcing human workers and LLMs. Quantifying these metrics is important because past research has indicated that inspirational stimuli comprising a mix of near and far analogies are most conducive to supporting designers in the early stage design process [24,25]. Therefore, to assess how broad of a spectrum the generated design solutions cover, we must measure their diversity.

To computationally measure diversity within a collection of design concepts, we focused on dataset coverage, so concepts that span a wider conceptual space should have higher diversity scores [26]. One straightforward method to evaluate this is by calculating the average distance to the nearest neighbor, also called nearest

generated sample [26]. This involves measuring the distance from each original datapoint to its closest generated counterpart and then averaging these distances across the dataset to gauge coverage. Another metric for assessing diversity is the convex hull. This method is predicated on the extent of the spread of the design solutions by using the total hypervolume encapsulated by the convex hull as an indicator of diversity [27]. Additionally, determinantal point processes (DPPs) offer an alternative approach. DPPs have been explored in prior research as a suitable metric for evaluating diversity in engineering design and more general machine learning contexts [28,29]. Lastly, we considered the average distance to the centroid of all generated design solutions as an additional potential measure of diversity [26]. In this paper, we applied all these methods to our dataset of design solutions to ensure methodological rigor and consistency across our findings.

3 Methods

We begin by discussing how we retrieved the human crowdsourced design solutions in Section 3.1. Following this, Section 3.2 details the prompt engineering techniques we utilized to generate the design solutions using a LLM. In that same section, we outline the various parameter combinations and prompt engineering methods we elected to test. We utilized GPT-4 (*gpt-4-0613*) for all experiments. In Section 3.3, we describe the diversity metrics selected for analyzing our study. For additional information along with access to the code, we have made the GitHub repository publicly available [2].

3.1 Crowdsourcing Design Solutions from AmazonTurk Workers. All the crowdsourced design solutions were extracted from a previous study conducted in 2019, prior to any large scale use of LLMs, as reported by Goucher-Lambert *et al.*, where Amazon Mechanical Turk was utilized to solicit design solutions from Amazon Turk workers [5]. The goal was to crowdsource a minimum of 100 responses from the workers for a variety of design problems, five of which were selected for this study and shown in Table 1.

Table 1 Design Problems Selected from Historical Data

Design Problems
1. A lightweight exercise device that can be used while traveling [30]
2. A device that disperses a light coating of powdered substance over a surface [4]
3. A new way to measure the passage of time [31]
4. An innovative product to froth milk [32]
5. A device to fold washcloths, hand towels, and small bath towels [33]

3.2 Engineering Zero- and Few-Shot Prompts. In this section, we outline the zero- and few-shot prompt inputs into the LLM.

3.2.1 Baseline Iterative Zero-Shot Prompting & Parameter Sweeps. In the baseline zero-shot prompting approach, the initial prompt is formulated as *"Generate 5 design solutions for "* followed by the specific design problem as outlined in Table 1. For example, a complete prompt input for a design problem would be *"Generate 5 design solutions for a lightweight exercise device that can be used while traveling"*. Upon receiving the initial prompt input, the LLM generates five design solutions, and both the initial prompt input and the output are recorded in a structured data repository. Subsequently, the LLM receives the prompt input *"Generate 5 more design solutions for"* followed by the same design problem as the initial prompt (e.g., *"Generate 5 more design solutions for a lightweight exercise device that can be used while traveling."*). The responses from this prompt and the prompt input are then added to the data repository. The initial prompt input is conducted just once, but the subsequent prompt inputs are conducted nine additional times, resulting in a cumulative total of 50 design solutions for each problem. This iterative prompting strategy was selected to afford the LLM greater latitude in elaborating on each design solution.

To address research question 1, we proceeded to use this style of baseline zero-shot prompting for each design problem across several combinations of top-P and temperature parameters. As noted in the GPT-4 API [14], the temperature values can vary from 0 to 2 and the top-P values can vary from 0 to 1. In deciding between which parameters to test, we divided the temperature and top-P values to low *(temperature = 0 / top-P = 0)*, medium *(temperature = 1 / top-P = 0.5)*, and high *(temperature = 2 / top-P = 1)*. We then tested the parameters on all possible combinations of low, medium, and high, which is shown in Table 2. Note, additional parameters for the GPT-4 model were set to the default parameters by OpenAI (Frequency penalty = 0 and Presence penalty = 0). In addition, the maximum tokens parameter was set to the maximum possible length.

Table 2 Temperature and Top-P combination

Temperature	Top-P
0	0
0	0.5
0	1
1	0
1	0.5
1	1
2	0
2	0.5

We excluded the high-high combination of temperature and top-P due to the output being incoherent. For each combination, we ran the baseline zero-shot prompting method as discussed in this section. This led to the LLM generating a total of 50 design solutions for each of the 8 different combinations of temperature and top-P.

3.2.2 Prompt Engineered Baseline Zero-shot Prompting. For this prompt engineering method, we leveraged the zero-shot reasoning capabilities of LLMs [23]. Prior studies have demonstrated that prefacing a query with phrases such as *"Let's think step-by-step"* can significantly enhance the LLMs' problem-solving process, yielding results comparable to those achieved through sophisticated prompt engineering methods [23]. We sought to determine whether subtle refinements to our baseline zero-shot prompts proposed in Section 3.2.1 could stimulate LLMs to produce a more diverse set of solutions. To this end, we experimented with the addition of certain phrases and adjectives to the original prompts. For instance, we introduced statements like *"You are a design expert."* or *"You are a design expert who is excellent at ideating far-fetched design ideas."*. Additionally, we incorporated adjectives such as *"novel"*, *"unique"*, *"creative"*, and *"diverse"* within the zero-shot prompts. An example of these prompt modifications is shown in Table 3 for design problem 1. We conducted this same process of prompt modification for all the other design problems listed in Table 1, starting with design problem 1.

3.2.3 Critique Prompt Engineering. We borrowed from prior literature that has shown there are benefits in having LLMs critique their initial answers to iterate on the solution and allow the LLM to 'reflect' on the answer, provide more rationale behind it, clarify any points of confusion, and add detail to the answer [34,35]. From the qualitative evaluations in our prior work, we found that the LLM-generated design solutions tended to lack details compared to the

[2] See the source code at GitHub Repository

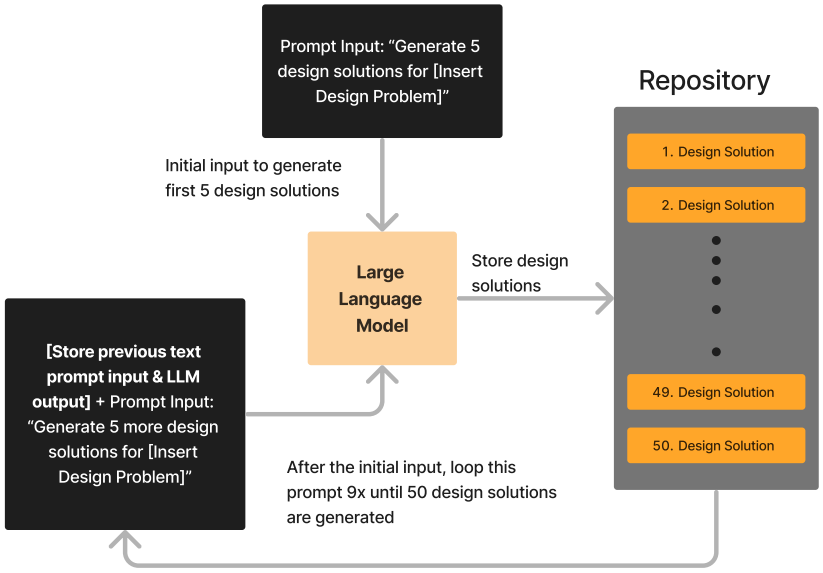

Fig. 2 Methodology for zero-shot baseline prompting. To generate a total of 50 design solutions, there was an initial input of *"Generate 5 design solutions for* [design problem]" (see Table 1 for list of design problems input and Table 3 for an example of how the prompts were input). After the LLM (GPT-4 in our case) generated 5 design solutions, they were stored in a data structure. Using the stored data structure, we conditioned the next generation of 5 more design solutions subject to the design solutions already generated as seen in the figure. We performed this loop 9 times until there was a total of 50 design solutions generated.

human-crowdsourced design solutions [7]. This lack of detail made it difficult for experts to evaluate the design solutions for feasibility, novelty, and usefulness. Moreover, design is an iterative process, and we suspected that allowing the LLM to iterate on the solution might result in a more diverse set of solutions.

To implement this critique method we first prompted the LLM to *"Generate 50 design solutions for [design problem]"*, where *[design problem]* was replaced by each of the five design problems in Table 1. Then, for each of the 50 design solutions, we prompted the LLM to *"please expand the design solution to add more detail and explain the reasoning and assumptions behind the solution"*. This yielded 50 critiqued design solutions for each of the design problems.

3.2.4 Few-Shot Prompting. In past literature, results showed that the accuracy and quality of the LLMs response could be improved by providing examples within the input prompt prior to requesting a specific output from the LLM [22]. This technique was also referred to as "few-shot learning". Our research sought to empirically evaluate this approach for the task of generating a diverse set of design examples, so we employed a similar methodology where we added examples to the initial baseline zero-shot prompt with some modifications (see Table 3 under the "Few-Shot" category for a specific example). For the selection of the design examples, we opted to randomly sample three design solutions from the human crowdsourced solutions corresponding to its respective design problem. The LLM generated a total of 50 design solutions via this method, and these solutions were then subject to comparative analysis against other prompt engineering strategies discussed in previous sections.

3.3 Computationally Measuring Diversity. We explored the diversity of design solutions generated through human crowdsourcing and LLM through a variety of prompt engineering techniques. We first converted the textual design solutions into vector embeddings using SentenceBERT [36], a model that had been used in past research to capture semantic similarity [37,38]. Through this embedding model, each design solution was represented as a 384-dimensional vector, resulting in a 50x384 dimensional embedding space for a set of 50 design solutions.

Our primary objective was to quantify the diversity within this set of solutions. Because there is no standardized way of measuring diversity, we employed a variety of computational metrics, including determinantal point processes (DPP), nearest generated sample, convex hull volume, and average distance to centroid to ensure a holistic assessment of diversity for the generated design solutions. Thus, each of these metrics provides a unique perspective on "space coverage" [26]: the nearest generated sample focuses on local density, the convex hull volume assesses the overall extent of the design coverage, the average distance to centroid evaluates coverage relative to the geometric central point, and DPP examines density through the determinant of the kernel matrix. While DPP and nearest generated sample calculations can be performed directly on high-dimensional data, other diversity metrics that we used, such as convex hull volume and average distance to centroid, required dimensionality reduction to facilitate computation. As a result, we used principal component analysis (PCA) to dimensionally reduce the embeddings to 20-dimensions for average distance to centroid calculations and to 13-dimensions for convex hull volume calculations. We chose this dimensionality because higher dimensions had presented computational limitations for running the convex hull and average distance to centroid calculations.

We used percentage change to compare the diversity of each set of parameters for generating design solutions relative to a baseline. Specifically, for each design topic and across all the diversity metrics, we calculated the percentage change for each set of LLM-generated design solutions relative to the second set of human crowdsourced solutions, referred to as 'Human 50 v2' in Figures 3 and 4. It is important to note that we divided the 100 total human crowdsourced solutions for each design topic into two groups of 50, which are labelled as 'Human 50 v1' and 'Human 50 v2' in the heatmaps, respectively. This division was done to ensure a fair comparison between each generated set of LLM-generated design solutions and human crowdsourced solutions. Thus, we used the following equation to calculate the percentage change:

Table 3 Examples of Prompt Engineered Zero- and Few-Shot Prompting for Design Problem 1

Example Prompt	Prompt Engineering Type
Initial Prompt: Generate 5 design solutions for a lightweight exercise device that can be used while traveling. **Subsequent Prompts:** Generate 5 more design solutions for a lightweight exercise device that can be used while traveling.	Baseline
Initial Prompt: Generate 5 **novel** design solutions for a lightweight exercise device that can be used while traveling. **Subsequent Prompts:** Generate 5 more **novel** design solutions for a lightweight exercise device that can be used while traveling.	Adjective - Novel
Initial Prompt: Generate 5 **unique** design solutions for a lightweight exercise device that can be used while traveling. **Subsequent Prompts:** Generate 5 more **unique** design solutions for a lightweight exercise device that can be used while traveling.	Adjective - Unique
Initial Prompt: Generate 5 **creative** design solutions for a lightweight exercise device that can be used while traveling. **Subsequent Prompts:** Generate 5 more **creative** design solutions for a lightweight exercise device that can be used while traveling.	Adjective - Creative
Initial Prompt: Generate 5 **diverse** design solutions for a lightweight exercise device that can be used while traveling. **Subsequent Prompts:** Generate 5 more **diverse** design solutions for a lightweight exercise device that can be used while traveling.	Adjective - Diverse
Initial Prompt: You are a design expert. Generate 5 design solutions for a lightweight exercise device that can be used while traveling. **Subsequent Prompts: You are a design expert.** Generate 5 more design solutions for a lightweight exercise device that can be used while traveling.	Phrase - You are a design expert
Initial Prompt: You are a design expert who is excellent at ideating far-fetched design ideas. Generate 5 design solutions for a lightweight exercise device that can be used while traveling. **Subsequent Prompts: You are a design expert who is excellent at ideating far-fetched design ideas.** Generate 5 more design solutions for a lightweight exercise device that can be used while traveling.	Phrase - You are a design expert who is excellent at ideating far-fetched design ideas
Initial Prompt: Generate 5 design solutions for a lightweight exercise device that can be used while traveling. **Here are some example design solutions [...]. Note, the example design solutions are just for guidance. You do not have to mimic the solutions.** **Subsequent Prompts:** Generate 5 more design solutions for a lightweight exercise device that can be used while traveling. **Here are some example design solutions [...]. Note, the example design solutions are just for guidance. You do not have to mimic the solution.**	Few-Shot

$$\Delta_{x_1 to x_2} = \frac{x_1 - x_2}{|x_2|} \times 100\% \quad (1)$$

where x_2 is the score of the second set of the human crowdsourced solution (referred to as 'Human 50 v2' in Figure 3 and 4) and x_1 is the diversity score for a specific design solution generation method. We opted to use the absolute value of x_2 in the denominator because we wanted to avoid instances where the result of the relationship between x_1 and x_2 is negative due to a negative denominator.

4 Results

4.1 Diversity Results.

4.1.1 Parameters Impact on Diversity of Generated Solutions. Findings from Figure 3 reveal that human crowdsourced solutions ('Human 50 v1' and 'Human 50 v2' along the x-axis) have higher diversity scores than LLM-generated solutions regardless of the evaluation method. In addition, the parameter combination of temperature = 1 and top-P = 1 yields the most diverse set of LLM-generated solutions as evidenced in Figure 3. Coincidentally, this parameter combination was also the default configuration provided by OpenAI's GPT-4 in its playground interface.

In addition, we observe that a definitive trend is not apparent across both the temperature and top-P parameter settings. This observation is supported by Figure 3, wherein, upon maintaining a constant temperature and analyzing the variations in diversity score with an incremental increase in top-P (across the x-axis), a consistent monotonic trend fails to emerge. Similarly, with top-P held constant, the variation in diversity score with an increase in temperature does not exhibit a clear monotonic pattern either. This observation is of interest, given the general expectation that an increase in temperature should correlate with enhanced "randomness" or "diversity", and a similar outcome is anticipated with an increase in top-P [14]. However, the data do not strongly support these assumptions.

4.1.2 Prompt Engineering Impact on Diversity of Generated Solutions. Similar to findings presented in Section 4.1.1, the heatmaps in Figure 4 reveal that human crowdsourced solutions, on average, exhibit higher diversity scores than those generated by the LLMs regardless of the prompt engineering technique applied or the diversity metric used to calculate the score. Notably, the critique-based prompt engineering method yields the highest average diversity score relative to the other prompt engineering approaches.

Furthermore, we observe that the diversity metric associated with the LLM-generated design solutions exhibit variability contingent upon the adjectives employed within the prompt inputs. Notably, the adjectives *"creative"*, *"unique"*, *"novel"*, and *"diverse"*, despite their semantic similarities, have differing outcomes in terms of the diversity scores. Specifically, the use of the word *"unique"* and *"diverse"* results in slightly comparatively lower diversity scores relative to the words *"creative"* and *"novel"*. Additionally, the inclusion of the phrase *"You are a design expert who is excellent at ideating far-fetched design ideas."* at the beginning of the baseline prompt leads to a considerable improvement in the diversity scores. However, a similar enhancement is not observed

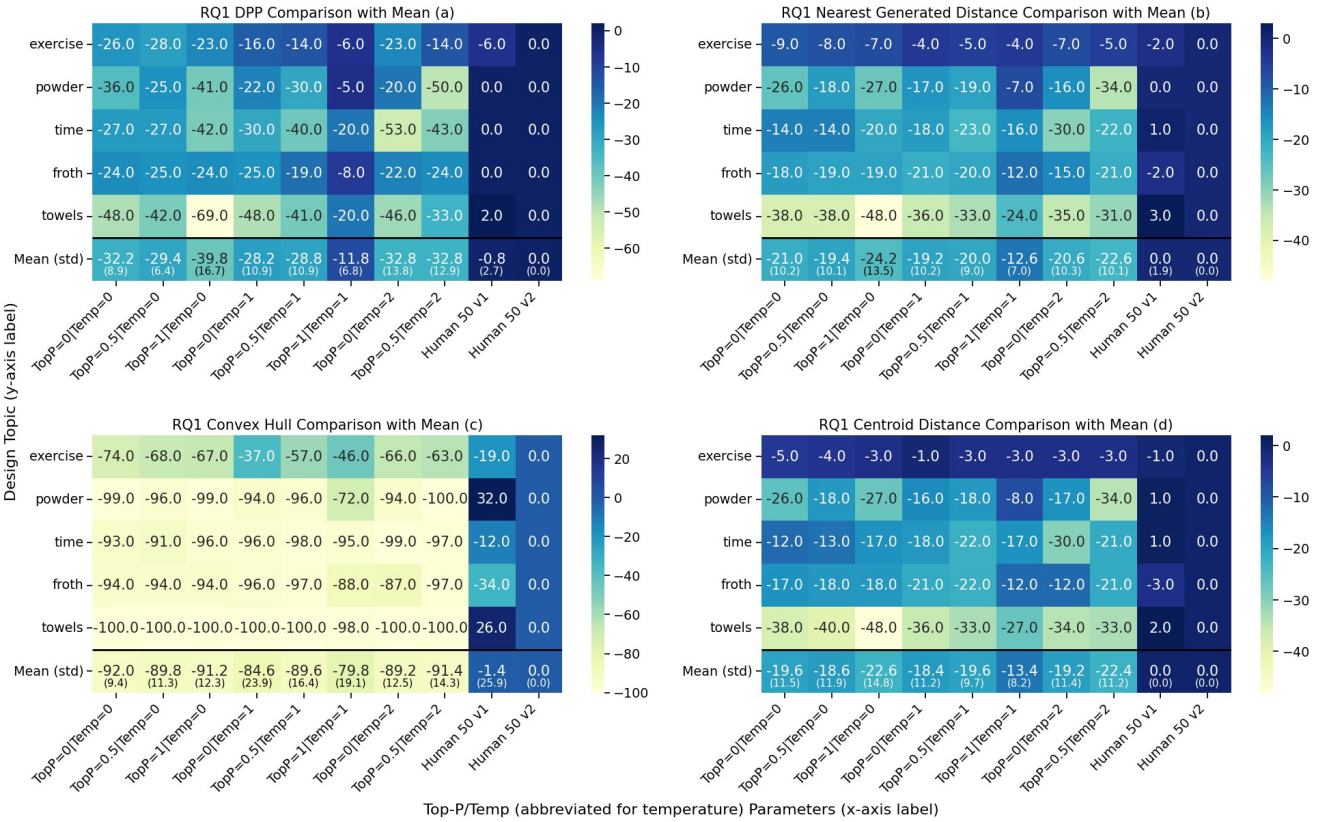

Fig. 3 Each heatmap represents one diversity metric. On the x-axis are the temperature and top-P values, and the y-axis are the corresponding design topics. The tabular value was calculated via percent difference in diversity to 'Human 50 v2' measured for the 50 design solutions.

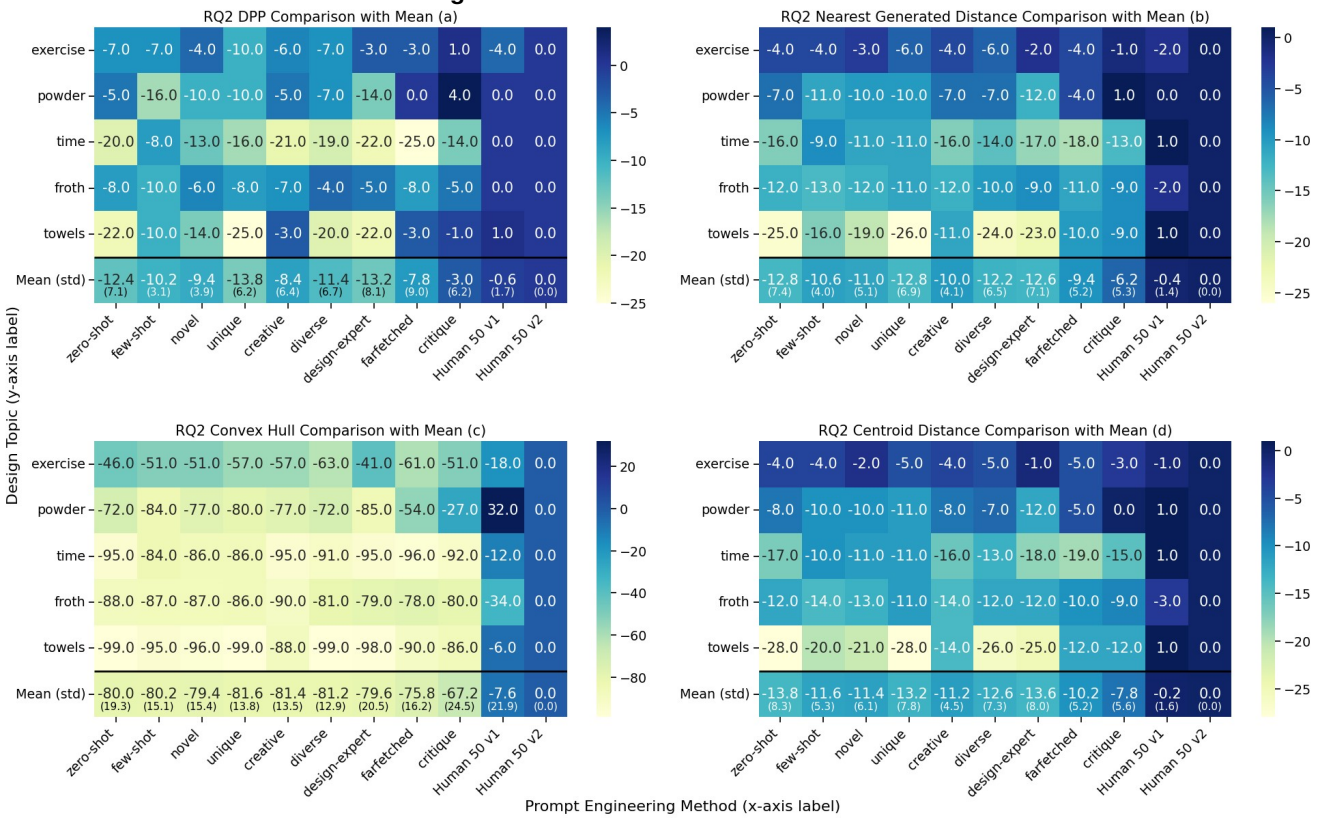

Fig. 4 Each heatmap represents one diversity metric. On the x-axis are the different ways of prompt-engineering, and the y-axis are the corresponding design topics. The tabular value was calculated via percent difference in diversity to 'Human 50 v2' measured for the 50 design solutions.

Table 4 Logistic regression model results with confusion matrix and statistical information.

Confusion Matrix Topics	TP/FN	FP/TN	Precision	Recall	F-1 Score
Froth	80	0	0.95	1.00	0.98
	4	16	1.00	0.80	0.89
Exercise Device	80	0	0.84	1.00	0.91
	15	5	1.00	0.25	0.40
Powder	80	0	0.87	1.00	0.93
	12	8	1.00	0.40	0.57
Time	80	0	0.89	1.00	0.94
	10	10	1.00	0.50	0.67
Towels	80	0	0.91	1.00	0.95
	8	12	1.00	0.60	0.75

when the phrase *"You are a design expert"* is included at the beginning of the baseline prompt. Moreover, our critique-critique method, which involves adding only one extra prompt sequence that edits the initial 50 generated design solutions, leads to the most notable improvement in the diversity scores. These observations suggest that even minor modifications to the prompt's structure or sequence can potentially lead to significant enhancements in diversity scores. These results also hint at the sensitivity of LLMs to subtle adjustments and modifications in the prompts and prompting sequences.

Despite the improvements in the diversity scores of the LLM-generated design solutions, we observed that these solutions did not achieve the level of diversity present in solutions from crowdsourced workers. This observation leads us to suspect that there might be semantic differences between design solutions produced by humans and those produced by LLMs. Thus, a subsequent examination of our dataset is undertaken to examine this question, which we detail in Section 4.2.

4.2 Logistic Regression Results. In the preceding section, we suggested that there may exist semantic differences between LLM-generated design solutions and human-crowdsourced design solutions. To investigate this, we trained a logistic regression model to determine if a decision boundary existed that could distinguish between embeddings of human-crowdsourced and LLM-generated design solutions [39].

We selected the prompt engineering dataset from RQ2 for our logistic regression analysis due to the results of the LLM-generated design solutions exhibiting the best diversity value relative to the human-crowdsourced design solutions. To train our model, we first categorized the design solutions as either LLM-generated or human-crowdsourced. Thus, for each design topic, the dataset comprised of 400 LLM-generated design solutions and 100 human-crowdsourced design solutions. We subsequently divided the data for each design topic into training and test sets, allocating 80% for the training set (320 LLM-generated design solutions and 80 human-crowdsourced design solutions) and 20% for the test set (80 LLM-generated design solutions and 20 human-crowdsourced design solutions). A logistic regression model was then trained on the training set for each design topic, and its accuracy was evaluated on the test set. The outcomes, including the confusion matrix, precision, recall, and F-1 score for each design topic, are presented in Table 4. The confusion matrix, as shown in Table 4, can be interpreted as follows: true positive means the LLM-generated solutions are correctly classified, false positive means LLM-generated solutions were incorrectly classified as human-crowdsourced solutions, true negative means the human-crowdsourced solutions were classified correctly, and false negative means the human-crowdsourced solutions were classified incorrectly.

Interestingly, we observed that all the LLM-generated design solutions were correctly classified, whereas the accuracy in correctly classifying the human-crowdsourced design solutions varied significantly across design topics, as evidenced by the fluctuating recall scores. Given the data imbalance, we argue that the accuracy in classifying human-crowdsourced design solutions serve as a more reliable metric for determining the presence of a significant semantic difference between the embeddings of human-crowdsourced and LLM-generated design solutions. As a result, the findings suggests that it is not definitive whether a clear distinction exists between human-crowdsourced and LLM-generated design solutions. The degree of separation appears to vary by design topic, with the milk froth design topic demonstrating the highest accuracy in distinguishing the human-crowdsourced solutions from those generated by LLMs.

5 Discussion

In this section, we discuss the findings of our results and their potential implications for design.

5.1 Enhancing Diversity in LLM-Generated Design Concepts through LLM Parameters. The influence of model parameters on concept design output diversity has not been extensively explored within a formal research context. Various informal sources have posited that parameters such as temperature and top-P in GPT-4 can influence the diversity and creativity of the generated text, based on the idea that these parameters can control the likelihood of the subsequent text generation [14]. Thus, the logic was that the higher the temperature, the more creative or diverse the output. However, findings from our study indicated that this relationship may not be as linear as previously thought. When controlling for one variable and observing changes in temperature or top-P across low, medium, and high settings, our analysis did not reveal a consistent trend in any specific direction. Despite this, our research did identify a parameter combination – medium temperae ture and high top-P (temperature = 1 and top-P = 1) – as having the highest score for diversity. Interestingly, this combination aligned with the default settings provided by OpenAI for GPT-4.

Future approaches could frame temperature selection as an optimization problem, using algorithms like sequential greedy search, or use machine learning techniques like reinforcement learning or model-agnostic meta-learning [40–42]. Alternatively, developing a new interface mechanism with a precise definition of diversity and a parameter to control it could enhance output diversity.

5.2 Enhancing Diversity in LLM-Generated Design Concepts through Prompt Engineering. Our study aimed to understand the impact of prompt engineering on the diversity of generated outputs by exploring three types of prompt manipulation. We found that minor adjustments to the prompt like adding the phrase *"'You are a design expert who is excellent at ideating far-fetched design solutions."* or prompting the LLM to critique and modify its own generated solution yielded the highest diversity score among the method tested. These results contrast with the minimal changes in impact that different parameter combinations of top-P and temperature had on the diversity score. Thus, we propose that while LLM parameters could enhance diversity, prompt engineering is more likely to have a greater impact on the diversity of outputs.

5.3 Future Directions in Enhancing Diversity using both LLM and Crowdsourcing Data. In Section 4.2, we explored the disparity in diversity scores between human-crowdsourced and LLM-generated design solutions, as noted in Sections 4.1.1 and 4.1.2. A logistic regression was trained to determine if a hyperplane could distinguish between these two types of design solutions. The results were mixed; for some topics like milk froth, the hyperplane effectively separated the solutions, while for others, it did not. This variation might be due to the specific nature of the design topics, with LLMs potentially having a more comprehensive understanding of certain concepts than the average crowdsourced worker. Overall, the results were inconclusive, preventing significant conclusions.

Despite this, interesting insights emerged. In the design topics where human crowdsourced solutions are semantically aligned

with those generated by LLMs, results suggested that designers could achieve a broad spectrum of diverse solutions simply by experimenting with various prompt engineering techniques. On the other hand, in the cases where there was a clear semantic gap between human and LLM outputs, we saw an opportunity for a synergistic approach. By leveraging human crowdsourced solutions as a source of inspirational stimuli, designers could utilize far-fetched solutions from human crowdsourcing in conjunction with LLM-generated solutions to enhance their own conceptual design space. This concept of combining AI with human crowdsourcing is supported by recent research that demonstrated AI, when combined with crowdsource evaluation, can boost innovative idea generation [43].

6 Limitations

We acknowledge several limitations that restricts the generalizability of our findings to other LLMs. Firstly, our investigation on LLM parameters is primarily predicated on settings specific to the GPT-4 model, and these findings may not apply to other pre-trained LLMs. Likewise, our comparison test for different prompt engineering techniques were tested solely on GPT-4, limiting their applicability to other pre-trained LLMs. Despite this, we argue that while our study is limited to a single model, it lays the groundwork for subsequent research to investigate whether these effects are consistent across different LLMs. Additionally, our study's focus on specific design topics under certain LLM settings may not fully capture real-world design complexity.

7 Conclusion

In this study, we explored the impact of tuning parameters in LLMs and the effectiveness of various prompt engineering strategies on the diversity of the generated design solutions. We found that there existed optimal parameters that led to the highest diversity, and we found that, surprisingly, the relationship between those parameters did not follow a clear pattern. Among the prompt engineering strategies tested, we found that various prompt engineering techniques ranging from modification of the prompt sequence to slight adjective modifications led to varying effects on the diversity score, indicating that the diversity of the LLM outputs may be highly responsive to prompt structure, phrasing, and sequencing. We also conducted a subsequent follow-on investigation to test whether human-crowdsourced and LLM-generated design solutions were semantically different and found mixed results across design topics. Notably, we were able to perfectly classify LLM-generated solutions from the human-crowdsourced solutions. However, the ability to use a linear classifier to correctly classify human crowdsourced solutions from LLM-generated solutions varied greatly based on the specific design topic, indicating that while LLM-generated solutions were consistently identified, human-crowdsourced solutions were sometimes misclassified. In summary, this study provides a comparison between LLM-generated and human-crowdsourced solutions, establishing a benchmark for assessing new models or frameworks in engineering design. This connection is vital for future studies in LLM applications for design diversity.

Acknowledgement

This work builds upon prior work published at the ASME International Design Engineering Technical Conferences and Computers and Information in Engineering Conference 2023 [7]. We also thank the members of the Berkeley Institute of Design for their feedback and support.

References

[1] Dahl, D. W. and Moreau, P., 2002, "The Influence and Value of Analogical Thinking during New Product Ideation," Journal of Marketing Research, **39**(1), pp. 47–60.

[2] Kwon, E., Huang, F., and Goucher-Lambert, K., 2022, "Enabling multi-modal search for inspirational design stimuli using deep learning," Artificial Intelligence for Engineering Design, Analysis and Manufacturing, **36**, p. e22.

[3] Fu, K., Murphy, J., Yang, M., Otto, K., Jensen, D., and Wood, K., 2015, "Design-by-analogy: experimental evaluation of a functional analogy search methodology for concept generation improvement," Research in Engineering Design, **26**, pp. 77–95.

[4] Linsey, J. S., Wood, K. L., and Markman, A. B., 2008, "Modality and representation in analogy," Artificial Intelligence for Engineering Design, Analysis and Manufacturing, **22**(2), p. 85–100.

[5] Goucher-Lambert, K. and Cagan, J., 2019, "Crowdsourcing inspiration: Using crowd generated inspirational stimuli to support designer ideation," Design Studies, **61**, pp. 1–29.

[6] Yu, L., Kittur, A., and Kraut, R. E., 2014, "Searching for analogical ideas with crowds," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, New York, NY, USA, p. 1225–1234, doi: 10.1145/2556288.2557378.

[7] Ma, K., Grandi, D., McComb, C., and Goucher-Lambert, K., 2023, "Conceptual Design Generation Using Large Language Models," *Proceedings of the ASME 2023 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Vol. 6: 35th International Conference on Design Theory and Methodology (DTM), ASME, Boston, Massachusetts, USA, p. V006T06A021, doi: 10.1115/DETC2023-116838.

[8] Zhu, Q., Zhang, X., and Luo, J., 2023, "Biologically Inspired Design Concept Generation Using Generative Pre-Trained Transformers," ASME Journal of Mechanical Design, **145**(4), p. 041409.

[9] Chan, J., Fu, K., Schunn, C., Cagan, J., Wood, K., and Kotovsky, K., 2011, "On the Benefits and Pitfalls of Analogies for Innovative Design: Ideation Performance Based on Analogical Distance, Commonness, and Modality of Examples," ASME Journal of Mechanical Design, **133**(8), p. 081004.

[10] Jiang, S., Hu, J., Wood, K. L., and Luo, J., 2022, "Data-Driven Design-By-Analogy: State-of-the-Art and Future Directions," ASME Journal of Mechanical Design, **144**(2), p. 020801.

[11] Poetz, M. K. and Schreier, M., 2012, "The Value of Crowdsourcing: Can Users Really Compete with Professionals in Generating New Product Ideas," Journal of Product Innovation Management, **29**, pp. 245–256.

[12] Yu, L., Kittur, A., and Kraut, R. E., 2014, "Distributed analogical idea generation: inventing with crowds," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, New York, NY, USA, p. 1245–1254, doi: 10.1145/2556288.2557371.

[13] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D., 2020, "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds., Vol. 33, Curran Associates, Inc., pp. 1877–1901.

[14] OpenAI (2023), 2024, "GPT-4 Technical Report," 2303.08774, https://arxiv.org/abs/2303.08774

[15] Zhu, Q. and Luo, J., 2023, "Generative Transformers for Design Concept Generation," ASME Journal of Computing and Information Science in Engineering, **23**(4), p. 041003.

[16] Wang, B., Zuo, H., Cai, Z., Yin, Y., Childs, P., Sun, L., and Chen, L., "A Task-Decomposed AI-Aided Approach for Generative Conceptual Design," *Proceedings of the ASME 2023 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 35th International Conference on Design Theory and Methodology (DTM), Vol. 6, ASME, Boston, Massachusetts, USA, p. V006T06A009, doi: 10.1115/DETC2023-109087.

[17] Goucher-Lambert, K., Gyory, J. T., Kotovsky, K., and Cagan, J., 2020, "Adaptive Inspirational Design Stimuli: Using Design Output to Computationally Search for Stimuli That Impact Concept Generation," ASME Journal of Mechanical Design, **142**(9), p. 091401.

[18] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I., 2017, "Attention is All you Need," *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., Vol. 30, Curran Associates, Inc.

[19] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W., 2022, "Emergent Abilities of Large Language Models," Transactions on Machine Learning Research.

[20] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G., 2023, "LLaMA: Open and Efficient Foundation Language Models," 2302.13971, https://arxiv.org/abs/2302.13971

[21] Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., Mordatch, I., and Florence, P., 2023, "PaLM-E: An Embodied Multimodal Language Model," 2303.03378, https://arxiv.org/abs/2303.03378

[22] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D., 2024, "Chain-of-thought prompting elicits reasoning in large language models," *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Curran Associates Inc., Red Hook, NY, USA.

[23] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y., 2024, "Large language models are zero-shot reasoners," *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Curran Associates Inc., Red Hook, NY, USA.

[24] Fu, K., Chan, J., Cagan, J., Kotovsky, K., Schunn, C., and Wood, K., 2013, "The Meaning of "Near" and "Far": The Impact of Structuring Design Databases and the Effect of Distance of Analogy on Design Output," ASME Journal of Mechanical Design, **135**(2), p. 021007.

[25] Chan, J. and Schunn, C., 2015, "The Impact of Analogies on Creative Concept Generation: Lessons From an In Vivo Study in Engineering Design," Cognitive Science, **39**(1), pp. 126–155.

[26] Regenwetter, L., Srivastava, A., Gutfreund, D., and Ahmed, F., 2023, "Beyond Statistical Similarity: Rethinking Metrics for Deep Generative Models in Engineering Design," Computer-Aided Design, **165**, p. 103609.

[27] Podani, J., 2009, "Convex hulls, habitat filtering, and functional diversity: mathematical elegance versus ecological interpretability," Community Ecology, **10**(2), pp. 244–250.

[28] Kulesza, A. and Taskar, B., 2012, "Determinantal Point Processes for Machine Learning," Foundations and Trends® in Machine Learning, **5**(2–3), pp. 123–286.

[29] Chen, W. and Ahmed, F., 2020, "PaDGAN: Learning to Generate High-Quality Novel Designs," ASME Journal of Mechanical Design, **143**(3), p. 031703.

[30] Linsey, J. S. and Viswanathan, V. K., 2014, "Overcoming Cognitive Challenges in Bioinspired Design and Analogy," *Biologically Inspired Design: Computational Methods and Tools*, A. Goel, D. McAdams, and R. Stone, eds., Springer, London, Chap. 9.

[31] Tseng, I., Moss, J., Cagan, J., and Kotovsky, K., 2008, "The role of timing and analogical similarity in the stimulation of idea generation in design," Design Studies, **29**(3), pp. 203–221.

[32] Toh, C. A. and Miller, S. R., 2014, "The Impact of Example Modality and Physical Interactions on Design Creativity," ASME Journal of Mechanical Design, **136**(9), p. 091004.

[33] Linsey, J. S., Markman, A. B., and Wood, K. L., 2012, "Design by Analogy: A Study of the WordTree Method for Problem Re-Representation," ASME Journal of Mechanical Design, **134**(4), p. 041009.

[34] Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., and Yao, S., 2023, "Reflexion: Language Agents with Verbal Reinforcement Learning," 2303.11366, https://arxiv.org/abs/2303.11366

[35] White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and Schmidt, D. C., 2023, "A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT," 2302.11382, https://arxiv.org/abs/2302.11382

[36] Reimers, N. and Gurevych, I., 2019, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, eds., Association for Computational Linguistics, Hong Kong, China, pp. 3982–3992, doi: 10.18653/v1/D19-1410.

[37] Walsh, H. S. and Andrade, S. R., 2022, "Semantic Search With Sentence-BERT for Design Information Retrieval," *Proceedings of the ASME 2022 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Vol. 2, St. Louis, Missouri, USA, p. V002T02A066, doi: 10.1115/DETC2022-89557.

[38] Tsumuraya, K., Amano, M., Uehara, M., and Adachi, Y., 2022, "Topic-Based Clustering of Japanese Sentences Using Sentence-BERT," *2022 Tenth International Symposium on Computing and Networking Workshops (CANDARW)*, pp. 255–260, doi: 10.1109/CANDARW57323.2022.00044.

[39] Hastie, T., Tibshirani, R., and Friedman, J., 2009, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer Series in Statistics, Springer New York, NY, Published: 09 February 2009 (Hardcover), 26 August 2009 (eBook).

[40] Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B., 2011, "Algorithms for Hyper-Parameter Optimization," *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, eds., Vol. 24, Curran Associates, Inc.

[41] Barto, A. G., 1997, "Chapter 2 - Reinforcement Learning," *Neural Systems for Control*, O. Omidvar and D. L. Elliott, eds., Academic Press, San Diego, pp. 7–30.

[42] Finn, C., Abbeel, P., and Levine, S., 2017, "Model-agnostic meta-learning for fast adaptation of deep networks," *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, JMLR.org, p. 1126–1135.

[43] Kittur, A., Yu, L., Hope, T., Chan, J., Lifshitz-Assaf, H., Gilon, K., Ng, F., Kraut, R. E., and Shahaf, D., 2019, "Scaling up analogical innovation with crowds and AI," Proceedings of the National Academy of Sciences, **116**(6), pp. 1870–1877.