Proceedings of the ASME 2024
International Design Engineering Technical Conferences and Computers and Information in Engineering Conference
IDETC/CIE2024
August 25-28, 2024, Washington, DC

DETC2024-143765

AUGMENTING ENGINEERING DESIGN WITH AI: INTRODUCING THE AI DESIGN ASSISTANT (AIDA)

Naveen Mathews Renji **Sagar Chakravarthy Mathada Veera** **Bei Yan** **Ting Liao***

Stevens Institute of Technology
Hoboken, NJ

ABSTRACT

It's critical to understand how to use artificial intelligence (AI) to foster innovation in the modern world as AI becomes more integrated into creative and problem-solving tasks. Using the sustainable washing machine as a primary example, this study designed and developed AI design assistant AIDA as a web-based chatbot to facilitate design ideation, leveraging large language models. AIDA prompts design tasks and assesses user-generated ideas for validity, novelty, and feasibility using RoBERTa-based models. As in the initial phase of an ongoing project, we conducted a human-subject experiment to validate a baseline version of AIDA and examined user performance and perceptions. The participants demonstrated smooth interaction with AIDA and consistent performance. They reported mostly positive perceived usefulness, enjoyment, and trust. Moreover, females and participants equal to or over 25 showed a comparable level of trust for general automated systems and AIDA, whereas male and under 25 participants were more skeptical about AIDA. This research offers a framework for technical development, tailored interactions, and real-time feedback, as well as insights into the use of AI chatbots to mediate engineering design. By analyzing user behavior and survey responses, we identified future directions in designing AI systems in engineering education and early-stage design.

Keywords: Human-AI collaboration, large language models, engineering design, design ideation, design education

1. INTRODUCTION

The advent of artificial intelligence (AI) has revolutionized numerous aspects of our lives, transcending various industries and domains. Doctors now rely on algorithms and surgical bots to diagnose and perform medical operations [1-2]. AI has also been incorporated into Microsoft Teams to schedule meetings, summarize meeting discussions, and assign tasks [3]. Blending AI into creative and problem-solving activities brings new possibilities for augmenting human capability, particularly in engineering design [4-6]. The emerging AI technologies offer engineering designers new tools at the early stages of need finding [7-9], brainstorming [10,11], and concept generation [11-15] to later stages of design evaluation [16,17], prototyping [18], as well as design education [19]. It also marks a significant shift in how we approach innovation, leveraging generative design and data-driven design. Despite its rapid advancement, modern AI systems become more interactive compared to legacy automated systems and even act as social actors [20,21]. When AI systems form new interaction dynamics with human users, there is still much to understand regarding how human users react to AI systems' feedback, recommendations, and even decisions, as well as how user behavior evolves when teaming up with AI systems.

To explore how AI can not only complement but actively boost human capabilities in engineering design, we proposed AIDA – AI Design Assistant – to act as a facilitator in the engineering design process and to address a simple yet profound question: *How can we make AI a teammate in the design process, rather than just a tool?*

*Corresponding author: tliao@stevens.edu

This question led us to design a multi-stage development approach for AIDA, where each stage introduces more advanced AI capabilities and seamless interaction. Building upon the proposed technological development, we aimed to understand how users interact with and perceive AIDA and how well users performed design activities along with AIDA.

The AIDA system is implemented as a chatbot. Chatbots that simulate natural human conversations are now broadly used to support essential activities [22-24], enhance education [25], assist team collaboration [26], streamline software development and testing [27,28], and reduce caregivers' workloads in clinical practice [29]. While acknowledging the various forms of design beyond text-based description, we chose chatbots to leverage the application of AI in natural language processing (NLP), which stands out as a transformative force. Yet, transitioning NLP systems to cognitive design teammates necessitates in-depth research and advancements in contextual understanding, creative problem-solving, collaboration and adaptability, emotional intelligence, and ethical reasoning.

This paper presents AIDA's design framework and development process *at its initial stage*. We conducted a human-subject study with novice designers and tested its functionality in generating design ideas for sustainable washing machines. The interaction flow was developed to mimic the standard design process, and AIDA was designed to evaluate user-generated ideas and provide feedback on their ideation outcomes to facilitate ideation. While learning design is challenging due to its iterative nature and the requirement for guidance, AIDA has great potential to benefit users with limited experience or domain knowledge by providing guidance and feedback at a relatively low cost. Meanwhile, AIDA provides an automatic, efficient, and potentially objective evaluation method for design ideas. In this paper, we validated AIDA's ability to interact with users and support design activities through a post-interaction survey and expert review. To guide further design iterations of AIDA, we discussed the observations, the takeaways, and the challenges we faced regarding development.

This paper is organized as follows: Section 2 reviews AI-aided concept generation, design idea evaluation, and the state-of-art of LLM. Section 3 presents the process of developing and implementing AIDA. Section 4 describes the experiment for testing and validating the design of AIDA, with exploratory analysis in Section 5. To provide guidelines for future phases of this project, we discuss current limitations and future directions in Section 6 and conclusions in Section 7.

2. BACKGROUND

2.1 AI-aided Concept Generation

AI-aided design, leveraging advanced algorithms and machine learning capabilities, has gained substantial popularity and is increasingly recognized as an invaluable tool across various industries and disciplines [5,30,31]. This widespread adoption attests to AI's transformative impact on the design process, contributing to increased efficiency, innovative solutions, and enhanced outcomes in diverse applications [32,33]. For example, Autodesk's Fusion 360 uses generative design to propose multiple design alternatives based on specified constraints and goals to explore a wider range of creative solutions [34]. AI is also utilized in topology optimization to enhance structural efficiency [35] and simulation software to predict stress points, thermal properties, and fluid dynamics [36].

Moreover, many emerging research studies have embodied AI algorithms as virtual design assistants. For instance, a context-aware design assistant by Pinquié et al. [37] leverages a property graph data model to manage and utilize design rules that were traditionally managed in unstructured documents. The assistant was able to retrieve rules based on specific documents, recommend appropriate rules during the design process, verify design solutions, and automate routine design tasks. Recent research in human-AI collaboration has also outlined interactions between human designers and AI assistants. A framework was proposed to improve design assistants, enhancing design quality, diversity, and efficiency for Design Space Exploration (DSE) in the early phases of complex systems [38]. Combining insights from design cognition, teams, and human-machine collaboration, the study highlights the importance of understanding their dynamics to improve design tools and outcomes.

In addition, social presence has been shown to influence motivation in human-AI collaboration settings. The study by Siemon et al. [39] investigates the role of social presence in group ideation processes involving an AI bot named GenBo, where the human participants were involved in brainstorming ideas to increase the number of zoo visitors. The study found that the presence of AI teammates significantly affected participants' motivation compared to the condition of an all-human team.

2.2 Design Ideas Evaluation

Evaluating design concepts is an integral part of the design process, contributing to the final product's overall success, efficiency, and relevance [40,41]. Establishing a standard system for evaluation is the top priority of evaluation. The evaluation metrics are commonly grounded in outcomes and encompass the quantity, quality, novelty, and variety of generated ideas [42,43]. Yet, some existing research points out that the conventional design evaluation method is constrained by its inability to directly measure attribute performance levels in a way that reflects their subsequent value to the designer and by the inaccurate quantification of beneficial attribute tradeoffs. In addition, concept assessment is commonly carried out through human expertise, but there is limited agreement on the design and reporting standards for human evaluations. The methods used and the information provided encompass issues such as incomplete details (e.g., number of evaluators, outputs assessed, and ratings gathered), insufficient analysis of obtained results (e.g., effect size and statistical significance), and substantial difference in the terminology and definitions used for assessed aspects of output quality [44,45]. Considering this situation, successfully implementing more automated approaches would furnish designers and the design research community with a benchmarking tool for consistently evaluating the outcomes of design ideation. Recently, Bradley et al. [46] proposed and

empirically tested an automated method for design concept assessment, utilizing machine learning to extract ontological data from a large set of crowd-generated concepts, introducing a filtering strategy and quantitative metrics for creativity rating. The results highlight the efficacy of the automated approach in outperforming human-selected subsets during the design concept selection.

2.3 State-of-the-art of LLMs

Large Language Models (LLMs) constitute a significant advancement in the field of NLP by pushing the limits of what machines can comprehend and produce in terms of human language. Transformer architectures [47] are the basis for many models, including RoBERTa and GPT (including ChatGPT) [48-50], which have completely changed how NLP tasks are approached. RoBERTa, or robustly optimized BERT, pushes the boundaries of pre-trained language representation models by improving their robustness, optimization, and generalization capabilities [49,50].

The state-of-the-art in LLMs is characterized by continuous advancements in model architecture, training techniques, and the scale of training data. These developments have led to notable improvements in language understanding and generation capabilities, making LLMs versatile tools for a wide range of applications [51], from chatbots like AIDA to content creation, translation, and more. As LLMs continue to evolve, they are set to redefine the landscape of human-computer interaction, making it more natural, intuitive, and effective.

3 SYSTEM DEVELOPMENT

3.1 System Design

The AIDA system was designed to ensure a seamless user experience and mimic a conventional brainstorming session on an individual basis. In the form of a web-based chatbot, AIDA first commences with a cordial greeting when a user initiates interaction. Subsequently, the user is provided with a design prompt and asked to brainstorm design ideas and input textual descriptions. This input will serve as the pivotal component in the subsequent interaction process.

To play a supportive role, the AIDA chatbot was particularly designed to provide feedback on ideas' novelty and feasibility – two essential and widely accepted criteria for assessing ideation outcomes. The assessments are automatically generated using LLMs and an existing dataset of sustainable design ideas (see details in Section 3.2). In addition, prior to the novelty and feasibility assessments, these input ideas undergo validation to ensure they are relevant to the given design prompt. The outcomes of these assessments cumulatively guide the AIDA chatbot's responses to users. If the input idea passes the validation, AIDA proceeds to check whether the idea is novel. If the idea is classified as novel, AIDA continues to assess its feasibility. Once the input idea passes all three assessment

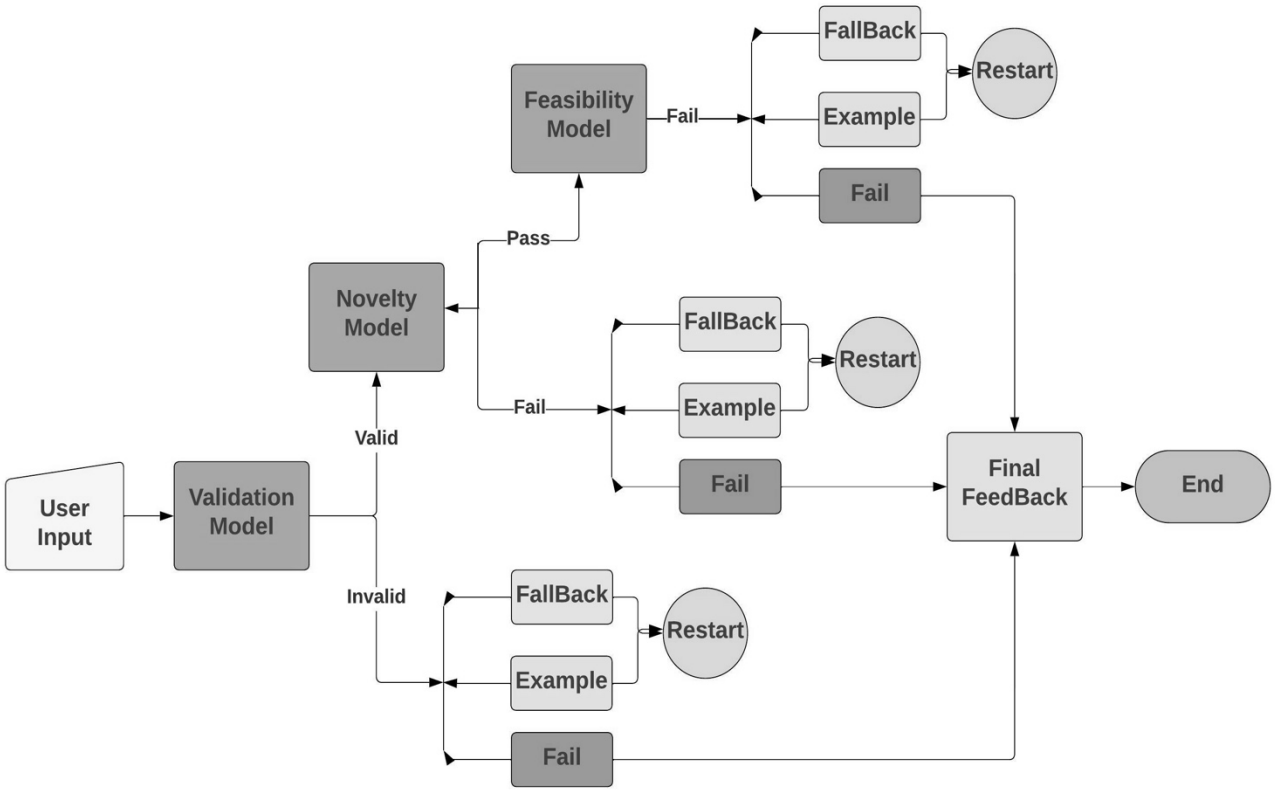

FIGURE 1 THE PROCESS FLOW CHART OF THE AIDA CHATBOT

models, it is considered a successful input, and the user is then informed of this. This process is depicted in Figure 1.

If an input idea fails to meet the established criteria of validity, novelty, and feasibility, the AIDA chatbot employs fallback messages to instruct users for revised submissions, as shown by the restart in the flow chart in Figure 1.

If user-generated ideas fail twice, a selected idea from the existing dataset is provided as an inspirational example. The examples provide guidelines for novice designers to refine their input, presumably enhancing ideation quality and assessment results. If a user fails to qualify for a particular criterium three times, their input is considered a failed idea, and they are given final feedback before the AIDA chatbot's termination. This iterative flow aims to engage users in a collaborative task and augment their performance via timely feedback.

3.2 Data Collection and Preprocessing

The datasets of design ideas and evaluation scores are very hard to come by, and training the assessment models is difficult compared to developing LLMs to classify sentiments using open databases. To overcome this challenge, we tailored the prompt and trained models using a dataset collected in a previous design study by the author [52]. That study prompted participants to generate ideas on tablets via a survey. The existing dataset contains 672 idea descriptions and their respective feasibility and novelty scores on a scale of 1-5. Table 1 shows examples of idea descriptions and their respective novelty and feasibility scores. Feasibility refers to the ease of implementation without major changes or violation of known constraints, financially and physically; novelty refers to not being expressed before and ingenious [52].

TABLE 1 EXAMPLE IDEAS FROM THE EXISTING DATASET

Idea Description	Novelty Score	Feasibility Score
Have a washing machine that can detect if clothes have stains and let users know which items have stains before you dry them.	3	3
Have a removable laundry basket that the clothes can be added or removed.	5	3
Display has a "time remaining" indicator for the wash cycle.	1	5

For the validation model, we needed data that could train the model to distinguish between ideas about washing machines and anything else. We leveraged chatGPT and generated descriptions about random items and descriptions of washing machines – not just sustainable washers. Then, we compiled them into a dataset with scores of 0 assigned to non-washing machine descriptions and 1 to washing machine descriptions. This dataset contained 238 rows, which were equally balanced between the valid (1) and invalid (0) design descriptions. Table 2 shows two example ideas and their respective validity scores.

TABLE 2 EXAMPLE IDEAS AND VALIDITY SCORES

Idea Description	Validity Score
have a removable laundry basket that the clothes can be added/removed in.	1
develop a video game that teaches players about history and culture.	0

The datasets were then split into three models of various compositions:

1. Validation Dataset: design ideas and validation score
2. Novelty Dataset: design ideas and novelty score
3. Feasibility Dataset: design ideas and feasibility score

To further expand the training dataset, we used the few-shot prompting of chatGPT, which involves providing a couple of original examples to chatGPT to steer results. We produced 742 new pairs of ideas and their respective scores to increase the variety of descriptions (Table 3). This potentially enhanced the dataset by increasing the quantity and variety and balancing the existing imbalance of the feasible and non-feasible ideas. In total, we had the novelty dataset with 1395 rows and the feasibility dataset with 1216 rows.

TABLE 3 COUNTS OF NOVEL AND FEASIBLE IDEAS

Data Source	Novel	Non-novel	Feasible	Non-feasible
Existing Dataset	362	291	575	77
ChatGPT	360	382	88	475
Total	722	673	663	553

3.3 Model Training and Fine Tuning

We adopted the RoBERTa architecture to classify the engineering design ideas based on their validity, novelty, and feasibility. Each model went through a similar process for fine-tuning, and the fine-tuned version of the 'roberta-base' pre-trained model was created. Model training and fine-tuning processes included tokenization, dataset split, training setup, model initialization, training, and evaluation.

Tokenization: Using the RobertaTokenizer, training design ideas were tokenized, ensuring the inclusion of special tokens and attention masks. We set a consistent padding strategy with a max length of 256 tokens.

Dataset Split: We employed a 90-10 train-validation split, ensuring a broad training base while retaining a separate validation set for model evaluation.

Training Setup: Training was orchestrated using HuggingFace's TrainingArguments class. Our setup comprised 50 epochs, a batch size of 16, and an evaluation strategy contingent on accuracy. The training process was tailored to load the model with the best accuracy at the conclusion of training. This ensured that if the models peaked prior to the end of the 50th epoch, we would still retain the most accurate version of the model.

Model Initialization and Training: We utilized the 'Roberta for Sequence Classification' for model initialization. The

training was facilitated using the Hugging Faces Trainer class, which integrated seamlessly with the model, datasets, and metric functions.

Evaluation: Post-training, the model's performance was gauged on the validation dataset, which was obtained by splitting the original data into a 1:9 ratio. The accuracy of the validation model is 94%; the novelty and feasibility models reach approximately 85% accuracy level, as shown in Table 4.

TABLE 4 ASSESSMENT MODELS' ACCURACY

Model	Validation	Novelty	Feasibility
Accuracy	95%	84%	86%

3.4 System Implementation

The implementation involved the development of the front end to allow seamless interactions and the back end to assess and respond to user inputs.

Frontend Development:

Web Platform: A web-based interface was constructed using the React framework. This interface offers users a dynamic platform to input, converse, and receive feedback on their idea input.

User Interface: A simple, easy-to-understand chat box layout was adopted (Figure 2). Users can initiate a chat, type in their ideas, and instantly view the system's responses, fostering an engaging interaction. Users can also create a login, allowing us to track the user inputs and interactions and store information in our database.

Backend Development:

Framework: We adopted Flask, a lightweight yet powerful web framework, as the backbone of our system's backend. This choice ensures swift responses and seamless integration with the machine learning components. The front end communicates with the back end using REST APIs.

Model Evaluation: Upon receiving a user's input from the front end, the backend triggers the fine-tuned models to evaluate the input ideas sequentially. The system assesses the ideas based on validation, novelty, and then feasibility metrics, utilizing the trained and fine-tuned RoBERTa models as in Section 3.2.

Response Generation: Once the evaluation is complete, the backend ascertains whether the input ideas surpass the established thresholds. Based on this, a pass or fail response is generated. This information is relayed back to the front end, which is rendered and presented to the user in a comprehensible format.

3.5 System Deployment

The deployment of the AIDA system involved integrating both backend and frontend components within a controlled and scalable environment. This was achieved through a streamlined process utilizing Docker for containerization, ensuring an isolated and consistent deployment setting.

Server Setup and Docker Integration: The initial step in the deployment involved preparing the server environment. This included updating server packages and integrating Docker, a key component for managing containerized applications. Docker's role was crucial in creating a stable environment for the backend Flask application and the React-based front end.

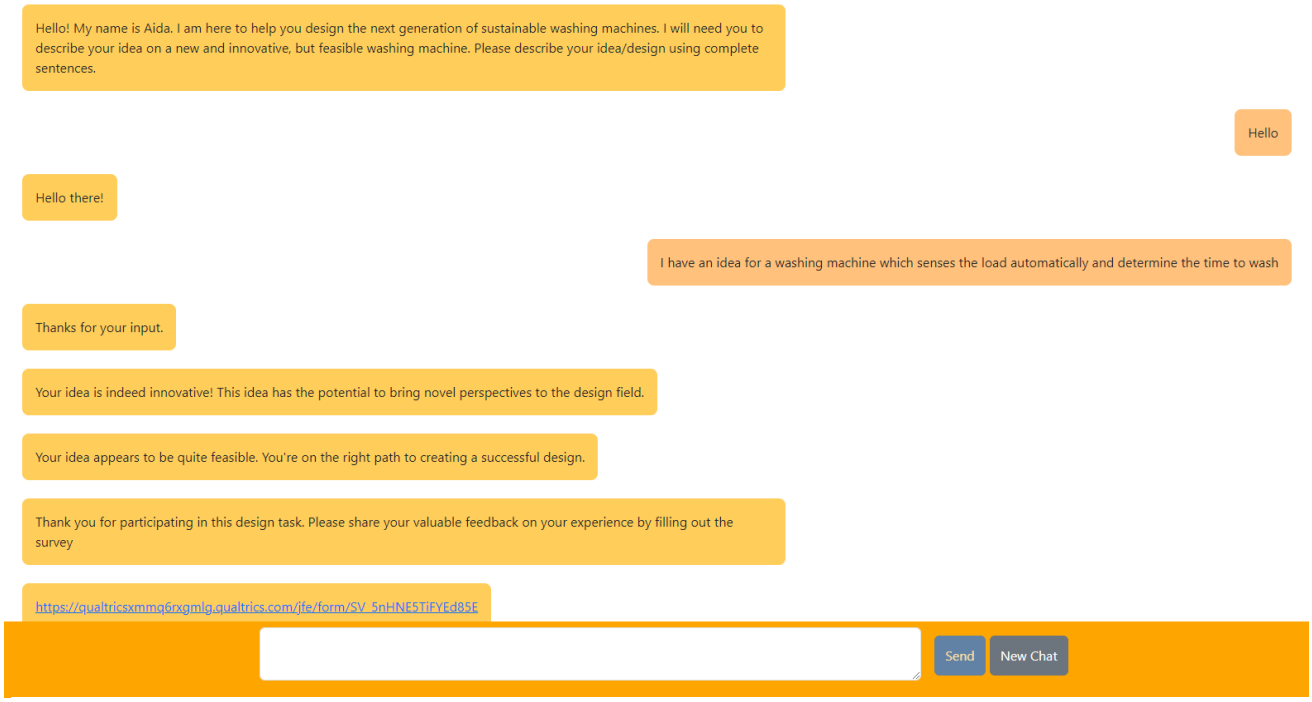

FIGURE 2 ILLUSTRATION OF THE AIDA CHATBOT INTERFACE

Backend Deployment: The deployment of the backend focused on transferring the Flask application and associated machine learning models to the server. The backend code was containerized using Docker, facilitating scalability and ease of future updates. The containerization of the backend ensured that the application could handle varying loads and could be easily maintained.

Frontend Deployment: Following the backend, the frontend was deployed using a similar Docker-based approach. The process involved building a Docker image for the React application, pushing this image to a Docker repository, and subsequently pulling and running it on the server. This approach ensured that the frontend deployment was consistent with the backend and equally scalable.

Security and Configuration: To ensure application security, SSL certificates were implemented to establish connections, and Nginx was configured as the web server. These steps were instrumental in safeguarding the application and efficiently managing client requests, thereby enhancing the overall reliability and performance of the system.

In summary, our system is a confluence of modern frontend and backend components meticulously designed to offer users real-time feedback on their design ideas. Through the interactive chatbot interface and the assessment models at its core, the AIDA system represents a significant stride in automating design evaluation and enhancing the ideation process.

4 EXPERIMENT DESIGN

4.1 Experiment Procedure

At its initial stage, we conducted an experiment to test and validate the proposed AIDA system. The experiment consists of three steps: 1) a pre-interaction survey about participants' general feelings, beliefs, and demographic information, 2) users' interaction with AIDA, and 3) a post-interaction survey about their perception of the AIDA system.

Once participants log in to the web application (Figure 2), AIDA initiates the design task with a message: "Hello! My name is Aida. I am here to help you design the next generation of sustainable washing machines. I will need you to describe your idea of a new and innovative but feasible washing machine. Please describe your idea/design using complete sentences".

Once the interaction with the chatbot concludes, we direct the participants to the post-interaction survey to understand their perception of AIDA during the interaction and gather feedback.

4.2 Human-Chatbot Interaction

After receiving the input idea and assessing it using the fine-tuned models, the AIDA chatbot provides three primary types of interaction: *feedback*, *fallback*, and *example*, shown in Figure 1.

When the input idea is recognized as valid, novel, and feasible, AIDA prompts *feedback*, such as "Your idea is indeed innovative! This idea has the potential to bring novel perspectives to the design field." The input needs to satisfy all three criteria; otherwise, it will proceed to fallback.

If the input fails any assessment, AIDA prompts *fallback* of that criterium, such as "The idea is robust, but similar concepts have been explored. You may consider using different methodologies, materials, or technologies. Please try again" for novelty.

If the second attempt fails, an *example* randomly drawn from the existing dataset with a mean score of three is provided, such as "The idea still lacks uniqueness. Have you considered pivoting the idea's focus? Here is an example: 'Sense the load automatically and automatically determine the time to wash.' Please elaborate on the idea or brainstorm another one."

If a participant fails the same assessment three times, they will be prompted to terminate ideation, as "Thanks for your effort. There are some notions like the one you proposed" and brought to the end of the interaction as "Thank you for participating in this design task. Please share your valuable feedback on your experience by filling out the survey." Passing the novelty assessment but failing the feasibility will trigger a fallback of the feasibility accordingly.

If the participants input a qualified idea, they will be prompted to repeat the ideation to gather more data, such as "You did a great job! Your design is novel and feasible. Please come up with another idea of sustainable washing machines, potentially with a different focus."

A mediocre example is chosen to set the baseline in this design stage. We acknowledge that people may tend to converge with examples. The assessment threshold is set to 0.5 so that most novice designers can complete two rounds of ideation. Definitions of novelty and feasibility were not explicitly explained during the session.

A timer is set to ensure that participants are engaged. If they idle for more than 5 minutes, the check-in message is prompted with an example idea.

4.3 Experiment Setup

Even though the AIDA chatbot can be accessed online, we decided to conduct an in-person experiment at this stage to better observe participants' behavior and obtain feedback for further design iterations. The controlled lab setup allowed participants to eliminate distractions and the influence of using online resources for design.

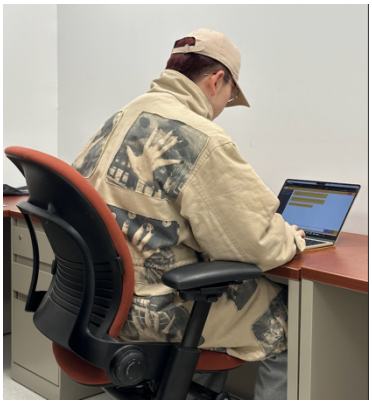

Figure 3 ILLUSTRATION OF EXPERIMENT SETUP

The experiment took place in a room with an office setup (Figure 3). A laptop was provided to access the surveys and AIDA chatbot. Participants were recruited at the authors' home

institution, acknowledging that most participants were novice designers. Each of the participants was provided compensation through a gift card worth $10 dollars.

4.4 Measuring Metrics

We recorded participants' textual input and self-reported beliefs and perceptions of AIDA. During the interaction, textual conversations and ideas generated by participants were recorded to understand participants' reactions and performance. After the interaction, participants were surveyed about their trust in AIDA, perceived effort, usefulness, and satisfaction with the chatbot using the existing metrics [53,54], validated in previous studies [55,56]. They reported their demographic information, previous experience, and knowledge of chatbots.

5 RESULTS AND ANALYSIS

In this exploratory study, the analysis aims to first verify the system design and then explore the user perceptions that guide the further system development and design iterations. With the approval of the Institutional Review Board (IRB), we conducted an experiment at a lab with 30 participants.

All 30 participants completed the prior interaction survey and interaction with AIDA. Four participants' post-interaction survey responses were deemed invalid due to their failure to clear the survey's attention verification questions. Of the remaining 26 participants, 12 self-identified as male, 13 as female, and one chose not to answer. All participants were under 35 years old; 19 were under 25, and 7 were equal to or greater than 25.

5.1 Task Completion

The maximum completion time of the experiment was 35 minutes, and the minimum was 18 minutes; the average time taken by all participants was 26 minutes. Each participant was prompted to conduct two rounds of idea generation unless they failed the validation or assessments. Among 26 respondents, one failed the validation test, and two encountered a fallback due to failure of the novelty test. The average number of ideas for each participant during the interaction was 1.93 ideas. Therefore, the system performed as expected to provide feedback and fallback and facilitate participants through the ideation process.

5.2 Ideation Outcomes

The ideas generated during the experiment were extracted from the conversational scripts, cleaned, and evaluated by subject experts for novelty and feasibility on a scale of 1-5, following the guidelines in the original design study [52]. Both expert reviewers – a Ph.D. candidate in sustainable design and a design faculty – had previous experience in sustainable design and obtained training on the rating scales. The rating was done separately by two reviewers (R1, R2). They both volunteered for this task. Table 5 shows the mean and inter-rater reliability (IRR) scores measured by Cohen's Kappa.

The low IRR scores for both metrics indicate a relatively low level of agreement regarding novelty and feasibility scores. Thus, the mean values are shown separately in Table 5. Two experts will be asked to discuss and resolve the discrepancies as part of future work in a standard process. Yet, the results highlight the inherent challenge of evaluating design ideas and generating novel ideas by novice designers.

TABLE 5 MEAN VALUES AND IRR OF EXPERT REVIEW

		Reviewer 1	Reviewer 2
Novelty	Mean	2.57	2.55
	IRR	0.103	
Feasibility	Mean	3.87	2.96
	IRR	0.176	

TABLE 6 MEAN VALUES BETWEEN TWO ROUNDS

		Reviewer 1	Reviewer 2
Novelty	First idea	2.51	2.55
	Second idea	2.62	2.44
Feasibility	First idea	3.70	3.03
	Second idea	4.03	2.92

The mean scores for each round of ideation were also calculated to examine if participants' performance varied, as shown in Table 6. Similarly, significant discrepancies exist between the reviewers. We observe a small improvement from the first idea to the second in the novelty and feasibility scores by R1 However, there is a slight decline in novelty and feasibility scores by R2. As the current design of AIDA does not provide further guidance between two rounds, it is anticipated that participants did not show significant improvement. Their performance did not deteriorate, indicating they were engaged in both rounds.

5.3 User Perception

To gauge participants' perceptions during their interaction with the AIDA chatbot, we measured participants' attitudes prior to their interaction, such as initial trust, anthropomorphism (on a scale of 10), prior experience with chatbots, and technical knowledge of chatbots. After their interaction with AIDA, we measured user perception, including trust in AIDA, perceived usefulness, and enjoyment. In general, our participants showed a relatively high trust propensity and neutral intention to anthropomorphize chatbots.

Overall, participants showed positive experience interacting with AIDA: the reported mean values of trust, usefulness, and enjoyment were greater than three, the neutral value (Table 7). This suggests that people can work with chatbots in the design process and find the chatbot's feedback useful for their ideation.

However, interacting with AIDA seems to decrease people's trust in chatbots. We observed differences between participants' initial trust and trust in AIDA and ran a paired t-test. The result shows a significant difference between initial trust and trust in AIDA (t-stats=2.26, p-value=0.03) while acknowledging that the questions about initial trust and trust were different. The decline in trust after the interaction may be attributed to the advent of highly sophisticated chatbots like ChatGPT and Bard, where most of the participants have used ChatGPT and Bard as an AI teammate/tutor for various tasks.

Participants' gender and age are common demographic factors for perceptions, and they were tested to see if they moderate how the interaction experience alters their trust in

AIDA. Table 8 shows mean values and paired t-tests between males and females under 25 and the other participants with p-values. The results indicate significant differences in trust before and after interaction for male and under-25-year-old participants. Female participants reported a relatively high and steady level of trust for general automated systems and the AIDA, whereas male participants were potentially keen to change when the focal product changed. Similarly, participants under 25 years old were sensitive to the specific interaction when they decided to trust. In addition, female participants showed slightly higher levels of perceived usefulness and enjoyment of AIDA, yet the results are not significant.

TABLE 7 CRONBACH'S ALPHA, MEAN, AND DEVIATION

Metrics	α	Mean	SD
Initial Trust	0.88	3.65	0.76
Anthropomorphism	0.88	4.90	1.76
Trust in AIDA	0.87	3.20	0.78
Usefulness	0.87	3.21	0.98
Enjoyment	0.87	3.62	1.11

TABLE 8 MEAN AND T-TEST BY GENDER AND AGE

	Gender		Age	
	Male	Female	<25	≥25
Initial trust	3.85	3.50	3.65	3.57
Trust in AIDA	3.00	3.50	3.10	3.36
T-statistic (p-value)	3.03 (0.01*)	0.07 (0.94)	2.40 (0.03*)	0.54 (0.61)

6 LIMITATION AND FUTURE WORK

As the initial stage of the project, this study laid a foundation for the system design and development, where we also examined user performance with the AIDA system and perception toward it. This study has several limitations. While we conducted it in a controlled lab setting, the participants were homogeneous, primarily students and novice designers. They demonstrated difficulty in generating novel ideas. While AIDA can assist inexperienced designers by enhancing their skills through immediate feedback and inspirational examples, their struggles underscore the value of introducing AIDA and similar tools in engineering education. The future version of AIDA will be hosted and distributed online, with the intention of accessing a diverse population beyond novice designers.

We acknowledge that the capability of our model to classify design descriptions is controlled by the available computational power and size of the training data. The scores in the training dataset were originally generated by human reviewers and inevitably incorporated bias. We acknowledge that the models may propagate these biases during training. Yet, utilizing LLMs with large-scale public data may overcome the inherent human bias in traditional design evaluation.

Most of the chatbot responses were prescribed to control the interaction. Based on the feedback from a participant, such as, "The chatbot seems good. Although the responses of the chatbot feel a bit robotic and could be better", more natural conversations need to be enabled for the next version of the design.

This version of AIDA was designed to set a baseline, while many participants expected it to be less automatic or more intelligent. For example, we received feedback: "It didn't feel like it detected my emotions during the process" and "While the chatbot was quick to respond, its replies felt too generic and didn't really dive into the specifics of my questions."

Another limitation stems from AIDA overestimating the novelty and feasibility of an idea due to its limited capability owing to a model fine-tuned on a small sample size. Though we used the few-shot technique to increase the sample size, we acknowledge the intricacies involved in a design description being termed novel or feasible.

As we conclude this initial phase, we are poised to embark on subsequent phases that promise to significantly elevate AIDA's functionality and user experience by providing, e.g., empathetic feedback, detailed explanations of scores, and natural conversations.

Advanced NLP Models: In later stages, we plan to enhance AIDA's capabilities by incorporating OpenAI's GPT-4 or Meta's LLaMa2. This upgrade will significantly improve AIDA's ability to deeply understand and critique design ideas and offer a personalized interaction experience based on each user's inputs and previous interactions.

Broadened Assessment Criteria: While the current system predominantly focuses on validity, novelty, and feasibility, future iterations could encompass a more extensive set of assessment criteria, enabling a more holistic evaluation approach.

Versatile Interaction Designs: The current AIDA provides feedback from a neutral perspective. More interactive designs, for example, empathetic vs. relentless, will be incorporated to examine their impact on user perception and performance.

By enhancing and refining AIDA's capabilities, we aim to extend its applicability by experimenting with more language models, understanding the impact of various factors on design outcomes, and setting new standards in human-AI teaming within and potentially beyond engineering design.

7 CONCLUSION

This study presents the design and development process of the AI Design Assistant (AIDA), which proactively prompts design tasks and provides real-time feedback leveraging LLMs. The proposed system lays a foundation for further development of AI-powered assistive systems, heralding a significant shift from traditional processes by human facilitators and reviewers to an automatic approach with virtual agents, such as chatbots. This automatic approach promises to simultaneously handle high volumes of brainstorming and evaluations and overcome obstacles such as location constraints and lack of guidance. Implementing emerging technologies, the assessment models can expedite the design evaluation process for improved reliability and objectivity.

In the experiment with a baseline AIDA system, participants successfully completed two rounds of ideation as instructed, whereas they maintained a similar level of performance and demonstrated difficulty generating novel sustainable ideas. Participants generally found the application easy to use, useful,

and trustworthy. Particularly, male participants and participants under 25 were sensitive to the interactions of the focal system, and their trust in the AIDA system was lower than their initial trust toward legacy automated systems. The results highlight the urgent need to examine and enhance intelligent design assistants with a focus on trust.

This study's contributions lie in the successful creation of an AI-powered design assistant with its chatbot interface and assessment models for instantaneous feedback. By merging sophisticated AI models with a user-centric interface, we have showcased an approach that prompts and encourages users to conduct design, with its capability of customizing interaction. The potential of AI-driven assessment systems using LLMs is vast, promising a future where design evaluations are quicker, more accurate, and devoid of human biases.

Beyond the realm of engineering design, the potential applications of this AI-powered system span a wide range of domains and industries. The majority of other chatbots are responders compared to the AIDA system, which plays the role of a prompter. Whether it's evaluating product designs in manufacturing, assessing creative content in media, or gauging business strategies in the corporate sector, the foundational principles of this system can be adapted to various paradigms, augmenting human capabilities across diverse fields.

ACKNOWLEDGEMENTS

We thank Hossein Basereh Taramsari for rating the design ideas as an expert in sustainable design.

This research was supported by National Science Foundation Grant 2301846 and 2105169.

REFERENCE

[1] Beane, M. “Shadow Learning: Building Robotic Surgical Skill When Approved Means Fail.” *Administrative Science Quarterly* Vol. 64 No. 1 (2018): pp. 87–123. DOI: 10.1177/0001839217751692.

[2] Lebovitz, S., Lifshitz-Assaf, H., and Levina, N. “To Engage or Not to Engage with AI for Critical Judgments: How Professionals Deal with Opacity When Using AI for Medical Diagnosis.” *Organization Science* Vol. 33 No. 1 (2022): 126–148. DOI: 10.1287/orsc.2021.1549.

[3] Velush, L. “How We’re Recapping Our Meetings with AI and Microsoft Teams Premium at Microsoft.” Inside Track, 2024. URL: https://www.microsoft.com/insidetrack/blog/how-were-recapping-our-meetings-with-ai-and-microsoft-teams-premium-at-microsoft.

[4] Zhang, G., Raina, A., Cagan, J., and McComb, C. “A Cautionary Tale About the Impact of AI on Human Design Teams.” *Design Study* Vol. 72 (2021). DOI: 10.1016/j.destud.2021.100990.

[5] Liu, Z., Gao, F., and Wang, Y. “A Generative Adversarial Network For AI-Aided Chair Design.” *Proceedings of the 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 486–490. IEEE, 2019. DOI: 10.1109/MIPR.2019.00098.

[6] Gyory, J. T., Song, B., Cagan, J., and McComb, C. “Communication In AI-Assisted Teams During an Interdisciplinary Drone Design Problem.” Proceedings of the Design Society Vol 1 (2021): pp. 651–660. DOI: 10.1017/pds.2021.65.

[7] Han, Y., and Moghaddam, M. “Analysis of Sentiment Expressions for User-Centered Design.” *Expert Systems with Applications* Vol. 171 (2021). DOI: 10.1016/j.eswa.2021.114604.

[8] Park, S., and Kim, H. M. “Phrase Embedding and Clustering for Sub-Feature Extraction from Online Data.” *ASME Journal of Mechanical Design*, Transactions of the ASME Vol. 144 (2022). DOI: 10.1115/1.4052904.

[9] Chandrasegaran, S. K., Ramani, K., Sriram, R. D., Horváth, I., Bernard, A., Harik, R. F., and Gao, W. “The Evolution, Challenges, And Future of Knowledge Representation in Product Design Systems.” *Computer-Aided Design* Vol. 45 (2013): pp. 204–228. DOI: 10.1016/j.cad.2012.08.006.

[10] Fu, K., Chan, J., Cagan, J., Kotovsky, K., Schunn, C., and Wood, K. “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*, Vol. 135 (2013). DOI: 10.1115/1.4023158.

[11] Lopez, C. E., Miller, S. R., and Tucker, C. S. “Exploring Biases Between Human and Machine Generated Designs.” *ASME Journal of Mechanical Design* Vol. 141 (2019). DOI: 10.1115/1.4041857.

[12] He, Y., Camburn, B., Luo, J., Yang, M. C., and Wood, K. L. “Visual Sensemaking of Massive Crowdsourced Data for Design Ideation.” *Proceedings of the International Conference on Engineering Design*, ICED 2019-August, pp. 409–418, Cambridge University Press, 2019. DOI: 10.1017/dsi.2019.44.

[13] Zhu, Q., and J. Luo. “Generative Pre-Trained Transformer for Design Concept Generation: An Exploration.” *Proceedings of the Design Society* Vol. 2 (2022): pp. 1825–34. DOI: 10.1017/pds.2022.185.

[14] Vlah, D., Žavbi, R., and Vukašinović, N. “Evaluation of Topology Optimization and Generative Design Tools as Support for Conceptual Design.” *Proceedings of the Design Society* Vol. 1, (2022): pp. 451–460. DOI: 10.1017/dsd.2020.165.

[15] Pillai, P. P., Burnell, E., Wang, X., and Yang, M. C. “Effects of Robust Convex Optimization on Early-Stage Design Space Exploratory Behavior.” *ASME Journal of Mechanical Design*, (2020). DOI: 10.1115/1.4048580.

[16] Yuan, C., Marion, T., and Moghaddam, M. “Leveraging End-User Data for Enhanced Design Concept Evaluation: A Multimodal Deep Regression Model.” *ASME Journal of Mechanical Design* Vol. 144 No. 2 (2022). DOI: 10.1115/1.4052366.

[17] Alcaide-Marzal, J., Diego-Mas, J. A., and Acosta-Zazueta, G. “A 3D Shape Generative Method for Aesthetic Product Design.” *Design Study* Vol. 66 (2020): pp. 144–176. DOI: 10.1016/j.destud.2019.11.003.

[18] Dering, M. L., Tucker, C. S., and Kumara, S. "An Unsupervised Machine Learning Approach to Assessing Designer Performance During Physical Prototyping." *Journal of Computing and Information Science in Engineering* Vol. 18 (2018). DOI: 10.1115/1.4037434.

[19] Chien, Y.-H., and Yao, C.-K. "Development of An AI User Bot for Engineering Design Education Using an Intent and Flow Combined Framework." *Applied Sciences* Vol. 10 (2020): p.7970. DOI: 10.3390/app10227970.

[20] Parasuraman, R., and Riley, V. "Humans and Automation: Use, Misuse, Disuse, Abuse." *Human Factors: The Journal of the Human Factors and Ergonomics Society* Vol. 39 (2006): pp. 230–253. DOI: 10.1518/001872097778543886.

[21] Nass, C., and Moon, Y. "Machines and Mindlessness: Social Responses to Computers." *Journal of Social Issues* Vol. 1 (2000): pp. 81–103.

[22] Fitzpatrick, K.K., Darcy, A., and Vierhile, M. "Delivering Cognitive Behavior Therapy to Young Adults with Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial." *JMIR Mental Health* Vol. 4 No. 2 (2017). DOI: 10.2196/mental.7785.

[23] Williams, A. C., Kaur, H., Mark, G., Thompson, A. L., Iqbal, S. T., and Teevan, J. "Supporting Workplace Detachment and Reattachment with Conversational Intelligence." *Conference on Human Factors in Computing Systems - Proceedings 2018-April* (2018): pp. 1–13. DOI: 10.1145/3173574.3173662.

[24] Toxtli, C., Monroy-Hernández, A., and Cranshaw, J. "Understanding Chatbot-Mediated Task Management." *Conference on Human Factors in Computing Systems - Proceedings 2018-April* (2018): pp. 1–6. DOI: 10.1145/3173574.3173632.

[25] Khang, A., Shah, V, and Rani, S, eds. *Handbook of Research on AI-Based Technologies and Applications in the Era of the Metaverse*. Hershey, PA: IGI Global, 2023. DOI: 10.4018/978-1-6684-8851-5.

[26] Fadhil, A., Schiavo, G., and Wang, Y. "CoachAI: A Conversational Agent Assisted Health Coaching Platform." (2019). URL: http://arxiv.org/abs/1904.11961.

[27] Lebeuf, C., Zagalsky, A., Foucault, M., and Storey, M. A. "Defining and Classifying Software Bots: A Faceted Taxonomy." *Proceedings of 2019 IEEE/ACM 1st International Workshop on Bots in Software Engineering*, BotSE 2019, pp. 1–6. Institute of Electrical and Electronics Engineers Inc., 2019. DOI: 10.1109/BotSE.2019.00008.

[28] Lebeuf, C., Storey, M. A., and Zagalsky, A. "Software Bots." IEEE Software Vol. 35 (2017): pp. 18–23. DOI: 10.1109/MS.2017.4541027.

[29] Habib, F., Shakil, G., Iqbal, S., and Sajid, S. "Self-Diagnosis Medical Chatbot Using Artificial Intelligence." (2021). DOI: 10.1007/978-981-15-6707-0_57.

[30] Tang, B., He, F., Liu, D., He, F., Wu, T., Fang, M., Niu, Z., Wu, Z., and Xu, D. "AI-Aided Design of Novel Targeted Covalent Inhibitors Against SARS-CoV-2." *Biomolecules* Vol. 12 (2022): 746. DOI: 10.3390/biom12060746.

[31] Mazzocco, G., Niemiec, I., Myronov, A., Skoczylas, P., Kaczmarczyk, J., Sanecka-Duin, A., Gruba, K., Król, P., Drwal, M., Szczepanik, M., Pyrc, K., and Stępniak, P. "AI Aided Design of Epitope-Based Vaccine for the Induction of Cellular Immune Responses Against SARS-CoV-2." *Frontiers in Genetics* Vol. 12 (2021). DOI: 10.3389/fgene.2021.602196.

[32] Bermejillo Barrera, M. D., Franco-Martínez, F., and Díaz Lantada, A. "Artificial Intelligence Aided Design of Tissue Engineering Scaffolds Employing Virtual Tomography and 3D Convolutional Neural Networks." *Materials* Vol. 14 No. 18 (2021): p. 5278. DOI: 10.3390/ma14185278.

[33] Wang, B., Asan, O., Liao, T., and Mansouri, M. "The Future Role of Clinical Artificial Intelligence: View of Chronic Patients." *IEEE Transactions on Technology and Society* (2024). DOI: 10.1109/TTS.2024.3374647.

[34] Gerhard, D., Köring, T., and Neges, M. "Generative Engineering and Design–A Comparison of Different Approaches to Utilize Artificial Intelligence in CAD Software Tools." *IFIP International Conference on Product Lifecycle Management* pp. 206-215. Cham: Springer Nature Switzerland.

[35] Carney, D. "Siemens Updates NX Software with AI Tools." DesignNews, 2022. URL: https://www.designnews.com/automotive-engineering/siemens-updates-nx-software-with-ai-tools

[36] Maduabuchi, C. "Thermo-mechanical Optimization of Thermoelectric Generators Using Deep Learning Artificial Intelligence Algorithms Fed with Verified Finite Element Simulation Data." *Applied Energy*, Vol. *315* (2022): p. 118943. DOI: 10.1016/j.apenergy.2022.118943.

[37] Pinquié, R., Véron, P., Segonds, F., and Zynda, T. "A Property Graph Data Model for a Context-Aware Design Assistant*." IFIP International Federation for Information Processing* (2019).

[38] Viros-i-Martin, A, and Selva, D. "A Framework to Study Human-AI Collaborative Design Space Exploration." *Proceedings of the ASME 2021 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. Online. August 17–19, 2021. DOI: 10.1115/DETC2021-67619

[39] Siemon, D., Elshan, E., de Vreede, T., and Oeste-Reiß, S. "Examining the Antecedents of Creative Collaboration with an AI Teammate." *Forty-Third International Conference on Information Systems*, Copenhagen 2022.

[40] Jing, L., Wang, J., Xie, J., Feng, D., Wang, J., Peng, X., and Jiang, S. "A Quantitative Simulation-Based Conceptual Design Evaluation Approach Integrating Bond Graph and Rough Vikor Under Uncertainty." *Journal of Cleaner Production* Vol. 380 (2022): p. 134928. DOI: 10.1016/j.jclepro.2022.134928.

[41] Schenke, K., Redman, E. J. K. H., Chung, G. K. W. K., Chang, S. M., Feng, T., Parks, C. B., and Roberts, J. D. "Does "Measure Up!" Measure up? Evaluation of an iPad App to Teach Preschoolers Measurement Concepts."

Computers and Education Vol. 146 (2020): p. 103749. DOI: 10.1016/j.compedu.2019.103749.

[42] Shah, J. J., Kulkarni, S. V., and Vargas-Hernandez, N. "Evaluation of Idea Generation Methods for Conceptual Design: Effectiveness Metrics and Design of Experiments." *ASME Journal of Mechanical Design* Vol. 122 (2000): pp. 377–384. DOI: 10.1115/1.1315592.

[43] Shah, J. J., Smith, S. M., and Vargas-Hernandez, N. "Metrics for Measuring Ideation Effectiveness." *Design Study* Vol. 24 (2003): pp. 111–134. DOI: 10.1016/S0142-694X(02)00034-0.

[44] Santhanam, S., and Shaikh, S. "Towards Best Experiment Design for Evaluating Dialogue System Output. *Association for Computational Linguistics* (2019). DOI: 10.18653/v1/W19-8610.

[45] Celikyilmaz, A., Clark, E., and Gao, J. "Evaluation of Text Generation: A Survey." (2020). DOI: 10.48550/arXiv.2006.14799.

[46] Camburn, B., He, Y., Raviselvam, S., Luo, J., and Wood, K. "Machine Learning-Based Design Concept Evaluation." *ASME Journal of Mechanical Design* Vol. 142 (2020). DOI: 10.1115/1.4045126.

[47] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. "Attention Is All You Need." *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010 (2017). DOI: 10.48550/arXiv.1706.03762.

[48] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., and Avila, R. "GPT-4 Technical Report." arXiv:2303.08774 (2023).

[49] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." arXiv:1907.11692 (2019). DOI: 10.48550/arXiv.1907.11692.

[50] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv:1810.04805 (2019). DOI: 10.48550/arXiv.1810.04805.

[51] Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., and Mian, A. "A Comprehensive Overview of Large Language Models." arXiv:2307.06435 (2023). DOI: 10.48550/arXiv.2307.06435.

[52] Liao, T., and Macdonald, E. F. "Priming on Sustainable Design Idea Creation and Evaluation." *Sustainability* Vol. 13 (2021): pp. 1–23. DOI: 10.3390/su13095227.

[53] Yu, A., Berg, J. M., and Zlatev, J. J. "Emotional Acknowledgment: How Verbalizing Others' Emotions Fosters Interpersonal Trust." *Organizational Behavior and Human Decision Processes* Vol. 164 (2021): pp. 116–135. DOI: 10.1016/j.obhdp.2021.02.002.

[54] Van den Broeck, E., Zarouali, B., and Poels, K. "Chatbot Advertising Effectiveness: When Does the Message Get Through?" *Computers in Human Behavior* Vol. 98 (2019): pp. 150–157. DOI: 10.1016/J.CHB.2019.04.009.

[55] Liao, T, and Yan, B. "Are You Feeling Happy? the Effect of Emotions on People's Interaction Experience Toward Empathetic Chatbots." *Proceedings of the ASME 2022 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. St. Louis, Missouri, USA. August 14–17, 2022. DOI: 10.1115/DETC2022-91059

[56] Liao, T, and Yan, B. "Let's Chat If You Are Unhappy – The Effect of Emotions on Interaction Experience and Trust Toward Empathetic Chatbots." *Proceedings of the ASME 2023 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. Boston, Massachusetts, USA. August 20–23, 2023. DOI: 10.1115/DETC2023-115318