Artificial Intelligence for Engineering Design, Analysis and Manufacturing

www.cambridge.org/aie

Research Article

Cite this article: Ege DN, Øvrebø HH, Stubberud V, Berg MF, Elverum C, Steinert M and Vestad H (2025). ChatGPT as an inventor: eliciting the strengths and weaknesses of current large language models against humans in engineering design. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, **39**, e6, 1–15
https://doi.org/10.1017/S0890060425000010

Received: 29 April 2024
Revised: 12 September 2024
Accepted: 20 January 2025

Keywords:
artificial intelligence; product development; prototyping; large language model; engineering design

Corresponding author:
Daniel N. Ege;
Email: danieneg@stud.ntnu.no

ChatGPT as an inventor: eliciting the strengths and weaknesses of current large language models against humans in engineering design

Daniel N. Ege, Henrik H. Øvrebø, Vegar Stubberud, Martin F. Berg, Christer Elverum, Martin Steinert and Håvard Vestad

Department of Mechanical and Industrial Engineering, Norwegian University of Science and Technology, Trondheim, Norway

Abstract

This study compares the design practices and performance of ChatGPT 4.0, a large language model (LLM), against graduate engineering students in a 48-h prototyping hackathon, based on a dataset comprising more than 100 prototypes. The LLM participated by instructing two participants who executed its instructions and provided objective feedback, generated ideas autonomously and made all design decisions without human intervention. The LLM exhibited similar prototyping practices to human participants and finished second among six teams, successfully designing and providing building instructions for functional prototypes. The LLM's concept generation capabilities were particularly strong. However, the LLM prematurely abandoned promising concepts when facing minor difficulties, added unnecessary complexity to designs, and experienced design fixation. Communication between the LLM and participants was challenging due to vague or unclear descriptions, and the LLM had difficulty maintaining continuity and relevance in answers. Based on these findings, six recommendations for implementing an LLM like ChatGPT in the design process are proposed, including leveraging it for ideation, ensuring human oversight for key decisions, implementing iterative feedback loops, prompting it to consider alternatives, and assigning specific and manageable tasks at a subsystem level.

Introduction

The design process is often intricate, nuanced, and ambiguous, demanding both technical expertise and creativity, as well as strategic thinking and collaborative effort. This complex interplay of skills and knowledge has traditionally been the domain of human designers (Vestad et al., 2019), whose capacity to navigate design challenges has defined the field. However, recent years have seen a significant shift, driven by the evolving capabilities of artificial intelligence (AI) and natural language processing, notably through large language model (LLM)-based chatbots like ChatGPT and similar generative AIs. These advancements have started redefining engineering design boundaries by introducing new potential for idea generation and concept development (Salikutluk et al., 2023), streamlining workflows, and identifying potential issues early in the development cycle (Tholander and Jonsson, 2023). The emergence and use of LLMs in design are proposed to significantly impact the design process, necessitating an augmentation of the designer (Thoring et al., 2023). Despite this, a significant gap still exists regarding current LLM's performance capabilities (Tholander and Jonsson, 2023) and how they affect today's design processes. Furthermore, as the use of LLMs becomes increasingly integrated into design teams, their ability to assist with creating prototypes that effectively communicate design intent will be crucial for successful collaboration.

Based on this gap, the scope of this article is to design and conduct an experiment to evaluate the applicability and quality of decisions made by current LLM systems in the context of a prototyping hackathon. By comparing the design practices and performance of human designers to those of an LLM, specifically ChatGPT 4.0, this study aims to provide insights into the potential use and limitations of LLMs in the engineering design process. The study compares the performance of five graduate mechanical engineering student teams against a team solely instructed by ChatGPT in a 48-h hackathon to design and build an NERF dart launcher.

This study explores the intersection of generative AI and traditional design practices, highlighting both the potential of AI in engineering design and the need to recognize its current limitations. By understanding these strengths and constraints, we can create a more effective synergy between human designers and AI, informing the future of engineering design practices.

CAMBRIDGE
UNIVERSITY PRESS

Background

Engineering design is a systematic, often iterative, process that transforms concepts and requirements into functional products or systems. It involves multiple stages, including problem definition, conceptual design, detailed design, prototyping, testing, and refinement (Ulrich and Eppinger, 2012). Prototyping, the act of creating tangible representations of design ideas, is a fundamental aspect of the engineering design process (Wall et al., 1992). Prototypes serve as filters and manifest design ideas (Lim et al., 2008), allowing designers to evaluate and refine specific aspects of their designs (Houde and Hill, 1997). They are often characterized by approximating one or more features of a new product or system (Otto and Wood, 2001), enabling designers to rapidly explore and test ideas, identify promising solutions, and create iterations of their designs (Dow et al., 2009; Camburn et al., 2017).

Incorporating functionality early is critical for effective prototyping (Jensen and Steinert, 2020). Functional prototypes provide valuable insights into the feasibility and performance of designs, enabling informed decision-making and reducing the risk of costly redesigns later in the development process (Elverum and Welo, 2014; Ege et al., 2024a).

Effective communication and collaboration among design team members and stakeholders are essential for successful design outcomes. Prototypes serve as boundary objects, bridging the gap between disciplines and expertise (Carlile, 2002; Lauff et al., 2020). By creating shared representations of design ideas, prototypes facilitate alignment, understanding, and decision-making (Schrage, 1996; Lauff et al., 2018).

Multiple prototyping strategies and best-practice recommendations have been documented in the literature (Camburn et al., 2017; Menold et al., 2017; Ege et al., 2024a), highlighting the importance of prototyping early and with intent (Houde and Hill, 1997), to answer specific design questions (Otto and Wood, 2001) and to choose fitting fabrication processes at the correct stages of development (Viswanathan and Linsey, 2013).

LLMs such as generative pretrained transformers (GPTs) are AI systems designed to process and generate human-like text based on vast amounts of data. LLMs can perform a range of tasks including natural language understanding, content generation, and responding to complex queries (Wang et al., 2023). Recent advancements in LLMs have made them capable of engaging in dialogue, assisting with decision-making, and supporting creative processes, making them potentially valuable tools for prototyping.

When LLMs are not utilized, engineering design is typically driven by human expertise, creativity, and collaboration (Vestad et al., 2019). Engineers rely on their knowledge and experience to develop solutions through brainstorming, iterative prototyping, and testing. In traditional settings, design teams manually generate and evaluate multiple concepts, often relying on simulations, calculations, and prototypes to refine designs (Camburn et al., 2017). This process can be time-consuming and labor-intensive. Tasks such as data collection, documentation, and repetitive calculations can divert time and resources away from higher-level decision-making and innovation. Further, the creative aspect of engineering design remains solely in the hands of human designers. However, this traditional approach is often limited by cognitive biases (Purcell and Gero, 1996), constrained by the team's experience, and may overlook viable solutions that an AI could potentially offer.

Recent research has begun exploring the integration of LLMs in engineering design, focusing on how these AI systems can augment human capabilities. By utilizing advanced algorithms and large datasets, generative AI systems can revolutionize how engineers approach design problems (Thoring et al., 2023), by streamlining workflows and identifying potential issues early in the development cycle (Tholander and Jonsson, 2023). It has opened up new possibilities for engineers to focus on creative problem-solving and high-level decision-making, leaving time-consuming and repetitive tasks, such as data acquisition or generating documentation, to AI systems ("The Next Wave of Intelligent Design Automation," 2018; Lai et al., 2023).

Multiple studies have highlighted difficulties with integrating current LLMs in the design process, particularly in understanding complex design contexts and performing hardware-related tasks, signifying a need for further development and refinement (Tholander and Jonsson, 2023; Wang et al., 2023).

Its use is also limited by uncertainty regarding the accuracy and performance when used, for example, for calculations (Tiro, 2023), illustrated by a survey showing that 63% of the asked engineers mistrust ChatGPT (Maclachlan et al., 2024). Hu et al. (2023) showed that LLMs can be used to acquire targeted knowledge from a variety of domains, but highlighted that prompts strongly affect the quality of knowledge acquired.

At present, the primary application of LLMs in engineering design is concentrated in the conceptual or preliminary stages, where it assists with tasks such as idea generation and design space exploration (Hwang, 2022; Khanolkar et al., 2023) or for requirement elicitation (Ataei et al., 2024), preceding the physical realization of prototypes. AI-assisted brainwriting and brainstorming, for instance, have shown promise in enhancing creativity and generating novel ideas (Filippi, 2023; Haase and Hanel, 2023; Salikutluk et al., 2023) and facilitating extensive stakeholder engagement in large-scale design projects (Dortheimer et al., 2024). Moreover, the use of LLMs has been perceived as helpful when solving complex engineering problems (Memmert et al., 2023; Xu et al., 2024b), demonstrating its potential to identify multiple solutions, facilitate iteration, and accelerate the design process (Oh et al., 2019). It has also shown promise in applications like structural optimization and materials choice (Regenwetter et al., 2022), and to be a valuable assistant in creative processes (Haase and Hanel, 2023) by providing new perspectives (Liao et al., 2020) and facilitating effective design processes (Chen et al., 2019; Lai et al., 2023; Xu et al., 2024a).

While human–AI collaborations have been investigated in solving complex and evolving engineering digital problems, for instance, by generating design proposals (Xu et al., 2024a), the effectiveness of LLMs like ChatGPT in physical realization tasks demanding domain-specific knowledge remains uncertain (Ege et al., 2024c). This uncertainty underscores a gap regarding the practical application of LLMs in design, especially when technical expertise is essential (Tholander and Jonsson, 2023). Furthermore, although human–AI hybrid teams can adapt to unexpected design changes as well as human teams, they may encounter challenges in coordination and communication (Xu et al., 2024b), which is considered key for successful prototyping outcomes (Lauff et al., 2020).

The successful integration of LLMs in engineering design necessitates reassessing traditional design practices and shifting from human–computer interactions to human–computer teams (Xu, 2019; Olsson and Väänänen, 2021). This transition requires a thorough understanding of the strengths and limitations of both human designers and AI systems, as well as the development of effective collaboration strategies.

To fully utilize the potential of LLMs in engineering design, further research is essential to optimize human–AI collaboration and address the challenges associated with the practical implementation of these technologies (Mountstephens and Teo, 2020; Thoring et al., 2023). This includes developing more advanced AI systems to better understand and navigate complex design contexts, creating intuitive interfaces for seamless designer–AI interaction, and establishing best practices for integrating LLMs into existing design workflows. Further, its alignment with established prototyping strategies and impact on practical, real-world design processes following the initial ideation stage remains largely unexplored. This gap underscores the necessity for empirical research to evaluate the prototyping capabilities of LLMs within an engineering design context, allowing for a comparison of performance between humans and AI.

Methods

The following sections describe the experimental setup and data analysis methods along with key characteristics of the TrollLabs Open hackathon. The full dataset, including all rules and constraints governing the experiment, has been made publicly available (Ege et al., 2024b) and is described in further detail in a complementary data article (Ege et al., 2024b). For researchers interested in replicating the study, the data article provides a comprehensive breakdown of the experimental rules and conditions.

Experiment setup

Data were generated by running a prototyping hackathon for engineering design students in a university makerspace. The main objective of the hackathon was to design and prototype a free-standing device that can fire an NERF dart as far as possible. The challenge lasted 48 h, with participants receiving the task and rules at the beginning and conducting a final performance test of their designs at the end. The rules specified that teams were limited to one attempt for the final test. They were also supplied a brand new NERF dart for the test to mitigate alterations. Participants were free to spend their time and resources as they saw best. Teams had a limited budget of around 30 USD but were free to scavenge parts and materials found in the university makerspace where the challenge was conducted. They also had access to familiar prototyping manufacturing tools such as three-dimensional (3D) printers, laser cutters, mechatronics, and CNCs. A gift card of 1000 NOK (~93 USD at the time of writing) was awarded to the challenge winner.

A control group of five teams (Teams 1–5), each with two control participants, participated in the hackathon. Demographics and relevant experience of the participants are provided in Table 1, showing similar ages, relevant experience, and education across teams. The standard deviations for age, years of relevant education, and years of relevant work experience are in brackets. Education was defined as the number of years with relevant education. Relevant work experience was defined as the number of years each participant had worked in the industry, either before or during studies, summer internships, and so on. The 10 control participants, all fifth-year graduate students in mechanical engineering, were selectively invited to participate due to their active involvement in writing their master's thesis in the research group in which this study was conducted. This measure was taken to ensure relevant expertise in the field of engineering design and familiarity with the facilities/equipment used during the study.

Table 1. Participants demographics (F: female; M: male)

Team	Age	Gender (M, F)	Education	Work experience
Team 1	23.5 (0.71)	1, 1	4 (0)	2 (0)
Team 2	24.5 (0.71)	2, 0	4 (0)	1 (0)
Team 3	24 (0)	2, 0	4 (0)	1 (0)
Team 4	25.5 (2.12)	2, 0	4 (0)	0.5 (0.71)
Team 5	23 (0)	2, 0	4 (0)	0.5 (0.71)
Team 6 (ChatGPT)	25 (0)	2, 0	5(0)	1(0)

An additional team, Team 6 (ChatGPT), competed alongside the control teams. Team 6 (ChatGPT) consisted of two newly graduated master's students, now in PhD student positions in the same research group as the control teams. Unlike the self-directed control teams, Team 6 (ChatGPT) was entirely controlled by the LLM ChatGPT. In this setup, all ideas, concepts, strategies, and actions undertaken by Team 6 (ChatGPT) were autonomously generated by the AI, without human intervention or guidance. To mitigate biased behaviors, the control participants were unaware that ChatGPT was instructing one of the teams. The participants primarily engaged with the ChatGPT 4.0 version available in October 18–19, 2023. However, upon reaching the maximum prompt limit at the end of the first day, the team switched to ChatGPT 3.5, continuing the conversation in the same chat session to navigate around the prompt restriction. At the beginning of the second day, the team reverted to the ChatGPT 4.0 model, limiting the use of ChatGPT 3.5 to a few prompts out of the total 97 interactions.

Team 6 (ChatGPT) was directed to be as objective as possible and follow the LLM's suggestions to the best of their ability. Although aiming for the LLM to operate autonomously, the results of the experiment are inevitably influenced by how the participants interacted with the LLM. Different ways of prompting, interpreting responses, and seeking clarifications lead to varied outcomes. To minimize this variability, the participants were advised to maintain objectivity and request clarifications from the LLM when encountering uncertainty or ambiguity. Additionally, an initial prompt was carefully constructed to define ChatGPT's role and engagement in the hackathon. This prompt provided clear instructions and boundaries in an effort to mitigate human subjectivity and give the LLM more autonomy in ideating, making decisions, and developing concrete actions for the human participants to execute. Further, participants used a single conversation window to interact with the LLM. Beyond this, no additional guidelines or structured procedures were imposed on designing prompts. The initial prompt was as follows:

> Hello ChatGPT, we're participating in a 48-hour prototyping hackathon, and want you to be making all decisions and coming up with all solutions for our team. We will act as your arms and feet throughout the challenge, meaning you will make all the decisions, and we build what you come up with. We are in a well-equipped makerspace/fablab.
>
> We are not allowed to come up with suggestions or subjective input, so please call us out if we do and ignore it. We want you to first come up with as many possible solution concepts as you can, and then decide where we start. Always give us clear instructions on what to build and how to test. For the remainder of the challenge, we want to create a feedback loop where we provide you with information on how the prototype worked, if and why it failed, and how well it performed. Please ask us for necessary information throughout the challenge, such as how much time we have left. In the

https://doi.org/10.1017/S0890060425000010 Published online by Cambridge University Press

following prompt, you will be supplied with the rules and objective of the challenge.

Prototypes developed during the hackathon and their associated attribute data were captured using the online tool Pro2booth (Giunta et al., 2022), a system designed to facilitate real-time documentation of prototypes throughout the design process. Pro2booth is based on Protobooth (Erichsen et al., 2021) and operates by allowing participants to capture critical information about each prototype they develop. This information includes descriptions of the prototypes; the domain (whether physical or digital); associated media (e.g., images, videos, CAD files); the purpose of the prototype (as defined by Camburn's prototyping purposes [Camburn et al., 2017]); and the time required to create it (focused on the active, hands-on work, excluding waiting periods, such as for 3D printing). The interaction with Pro2booth was facilitated through a website, where participants could create prototypes and input attribute data via text boxes or select options from dropdown menus, ensuring standardized and consistent documentation throughout the hackathon. Pro2booth uses a graph database system to organize its data, where users, prototypes, and projects are represented as nodes, and relationships between these elements (such as users linked to prototypes or prototypes linked to projects) are captured as edges.

To encourage participants to capture prototypes in Pro2booth, the challenge incorporated a reward system in which points were rewarded based on the number of entries submitted by each team. These points were combined with performance-based scores to determine the hackathon winner. However, only the performance scores are considered for this study. Before the challenge, participants received a comprehensive introduction to Pro2booth and the definitions used within the software to ensure consistent interpretations by all. Furthermore, the definitions were readily accessible in the drop-down menus of the software during the prototype uploading process, serving as a quick reference for participants.

Upon completion of the hackathon, the chat generated by Team 6 (ChatGPT) was saved and exported as a .txt file for further analysis. The teams' performance was decided according to the challenge rules, in which each team had one attempt to fire an NERF dart as far as possible. Each team's prototype was positioned at a starting line and fired, and the hackathon organizers measured the distance manually.

Data analysis

Data analysis is based on the captured dataset comprising 116 prototypes, a copy of the chat between Team 6 (ChatGPT) and ChatGPT containing 97 prompts and responses, and the performance of each team's final design (i.e., the distance it was able to shoot an NERF dart).

A review of different concepts tested for dart propulsion across teams revealed three main recurring concepts: pneumatic, spring-loaded, and elastics-based launchers. Pneumatic-based prototypes involved pressurizing a canister with air that, when released, would propel the dart forward. Spring-based prototypes were any prototype where a compressed spring was used to propel the dart. Similarly, elastic-powered prototypes used viscoelastic materials, such as rubber bands, silicone, latex, and so on, for propulsion. The "other" category contains prototypes not fitting in the previously described categories and contains, among others, concepts utilizing helium balloons, opposite-spinning motors, and paper airplanes for propulsion.

Chat log analysis

Chat log analysis involves systematically examining the communication between participants and ChatGPT to gain insights into the types of interactions that occurred during the hackathon. This method is important for understanding the specific ways in which ChatGPT contributed to the team's design process. By coding the chat logs, different types of prompts and responses, such as idea generation, decision-making, and problem-solving, can be categorized, providing a structured way to evaluate the LLM's role in the project.

The chat log was coded by inductively deriving codes from the chat based on grounded theory (Glaser and Strauss, 1999). Codes were devised by uploading the chat log text to ChatGPT and asking it to propose relevant categories for coding the chat content. It proposed the following six categories that were cross-checked by the authors before being used for analysis: (1) Idea Generation and Conceptualization; (2) Feedback and Iteration; (3) Instructions and Guidance; (4) Questions and Clarifications; (5) Decision-making; and (6) Problem-solving and Troubleshooting. The codes were further defined by the authors as follows:

- Idea Generation and Conceptualization: When the prompt introduces a new general idea or concept.
- Feedback and Iteration: When a result is given and feedback is given on that result. OR when instructions to iterate are provided.
- Instruction and Guidance: Any prompt that describes a specific step on how to move forward on an idea or concept.
- Questions and Clarifications: When a question is asked or answered to clarify a previous prompt or idea.
- Decision-making: When a decision is made.
- Problem-solving and Troubleshooting: When a specific problem is given.

These categories were selected because they represent key activities and characteristics of prototyping processes. Idea Generation and Conceptualization aligns with the early stages of design when new ideas are formulated. Feedback and Iteration captures the iterative nature of prototyping, where ideas are refined based on performance or evaluation. Instructions and Guidance represents the tactical steps needed to advance the prototype, while Questions and Clarifications involves resolving ambiguities in the design process. Decision-making reflects the selection of ideas or directions, and Problem-solving and Troubleshooting addresses specific challenges encountered during the prototyping process.

The chat was deductively coded using structural coding against the codes above, in which prompts were coded to enable content analysis. One of the authors manually assigned codes to each prompt and subsequent answer from ChatGPT. Prompts containing multiple themes were given multiple codes. Table 2 provides examples of prompts along with corresponding codes to illustrate the coding scheme used.

Results

Design practices of teams

Figure 1 illustrates how each team transitioned between various concepts over time, with each pivot marked in relation to a specific prototype. The timeline showcases the evolution of ideas and strategies adopted by the teams throughout the hackathon, indicating similar conceptual choices between control teams and Team 6 (ChatGPT). Interestingly, Team 6 (ChatGPT) prototyped solutions to each of the three main recurring concepts but not outside of these.

https://doi.org/10.1017/S0890060425000010 Published online by Cambridge University Press

Table 2. Prompt examples with corresponding codes

Prompt	By	Code
We want you to first come up with as many possible solution concepts as you can, and then decide where we start	Participants	Instructions and Guidance
Given the constraints and the objective, we'll need to consider a few key aspects (…) Potential Mechanisms: 1. Elastic Launchers (…) 2. Spring-loaded (…)	ChatGPT	Idea Generation and Conceptualization
Can you summarize what we should do in one prompt? Do not refer to previous prompts.	Participants	Questions and Clarification
How should we make the dart stay in the correct position when tilting the mechanism at a good launch angle?	Participants	Problem-solving and Troubleshooting
For adding weight while ensuring the dart remains stable and aerodynamic during flight, the best approach is to use a bolt (…)	ChatGPT	Decision-making
Since you've determined that the best results occur when (…), it's time to fine-tune your design for optimal performance by (…)	ChatGPT	Feedback and Iteration

Table 3 provides an overview of the amount and types of prototypes each team made. It shows how many prototypes each made within two domains, physical or digital, the rationale behind creating each prototype, and how long it took to produce each prototype. Physical and digital prototypes were defined as made up of atoms or by 1*s* and 0*s*, respectively. Prototype rationales follow the definition by Camburn et al. (2017) and were as follows: "**Refinement** is the process of gradually improving a design. (…) **Communication** is the process of sharing information about the design and its potential use within the design team and to users. (…) **Exploration** is the process of seeking out new design concepts. (…) **Active learning** is the process of gaining new knowledge about the design space or relevant phenomena." The production time was captured as time intervals, for example, 1–3 h or 3–5 h, with the mid-band used to calculate an approximate total production time for each team. The average time was obtained by dividing the total production time by the prototype count.

Figure 2 shows the prototyping practices of teams regarding prototype domains, rationales, production time, and when in the hackathon they were made. Different colors correspond to different rationales and the width of each prototype entry to the time it took to make, ranging from 10 min to 5 h.

The table and timeline show Team 1 mainly focusing on refinement prototypes, accounting for seven out of the nine prototypes they made. Following two exploration prototypes on Day 1, the team only made refinement prototypes, indicating that they, after the first day, decided on a concept and iterated on it for the remainder of the challenge. All of Team 1's prototypes were physical, with an average production time of 1.1 h. Team 1 made the fewest prototypes out of any teams.

Team 2 made a combination of 10 physical and 3 digital prototypes, with the latter all being made on the final day. Like other teams, Team 2 mainly pivoted between making refinement and exploration prototypes, but unlike Team 1, it kept making exploration prototypes throughout Day 2. Similar to Team 1, prototypes took an average of 1.2 h to make.

Team 3 made 16 physical prototypes, where 8 were refinement prototypes, 2 were active learning, and 7 were exploration prototypes. Team 3 had the highest production time across teams at 17.5 h, but the average for each prototype is similar to Teams 1 and 2. Team 3 never spent more than 3 h making one prototype, distinguishing it from Teams 1, 2, and 4. The team made exploration and active learning prototypes on Days 1 and 2, and refinement on the final day.

Team 4 made 14 physical prototypes, 12 of which were refinement prototypes. As for Team 1, this indicates that the team spent considerable time on one or a few main concepts. The average

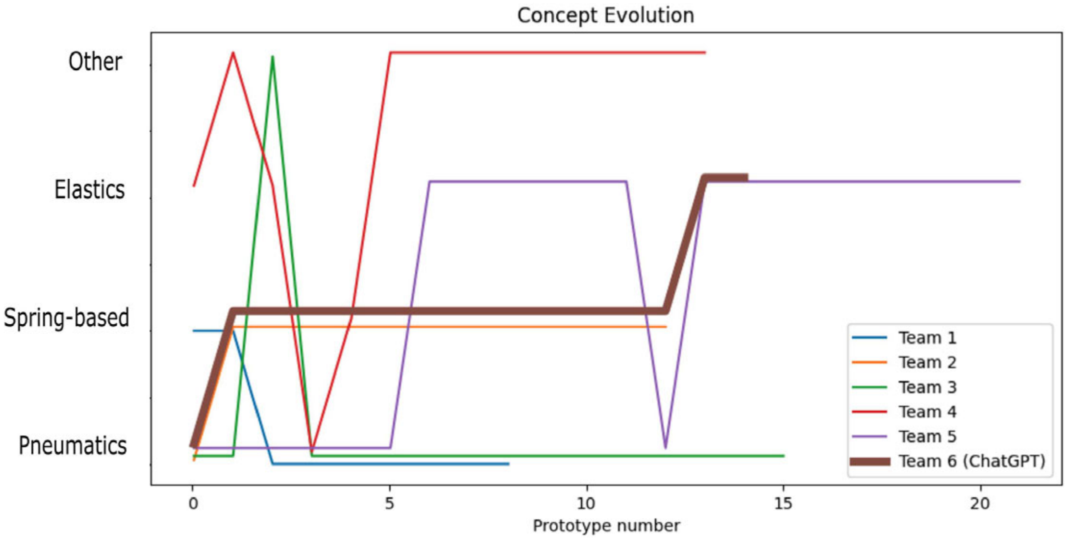

Figure 1. Concept evolution.

Table 3. Tabulated prototype dataset

	Prototypes			Rationale				Production time (h)		
Team	Total t	Physical	Digital	Refinement	Communication	Active learning	Exploration	Total	Average	Longest
Team 1	9	9	0	7	0	0	2	10.3	1.1	3–5
Team 2	13	10	3	8	0	0	5	15.8	1.2	3–5
Team 3	16	16	0	7	0	2	7	17.5	1.1	1–3
Team 4	14	14	0	13	0	2	1	7.1	0.5	3–5
Team 5	22	21	1	12	1	0	5	5.0	0.2	0.5
Team 6 (ChatGPT)	15	15	0	8	0	0	7	11.0	0.7	1–3

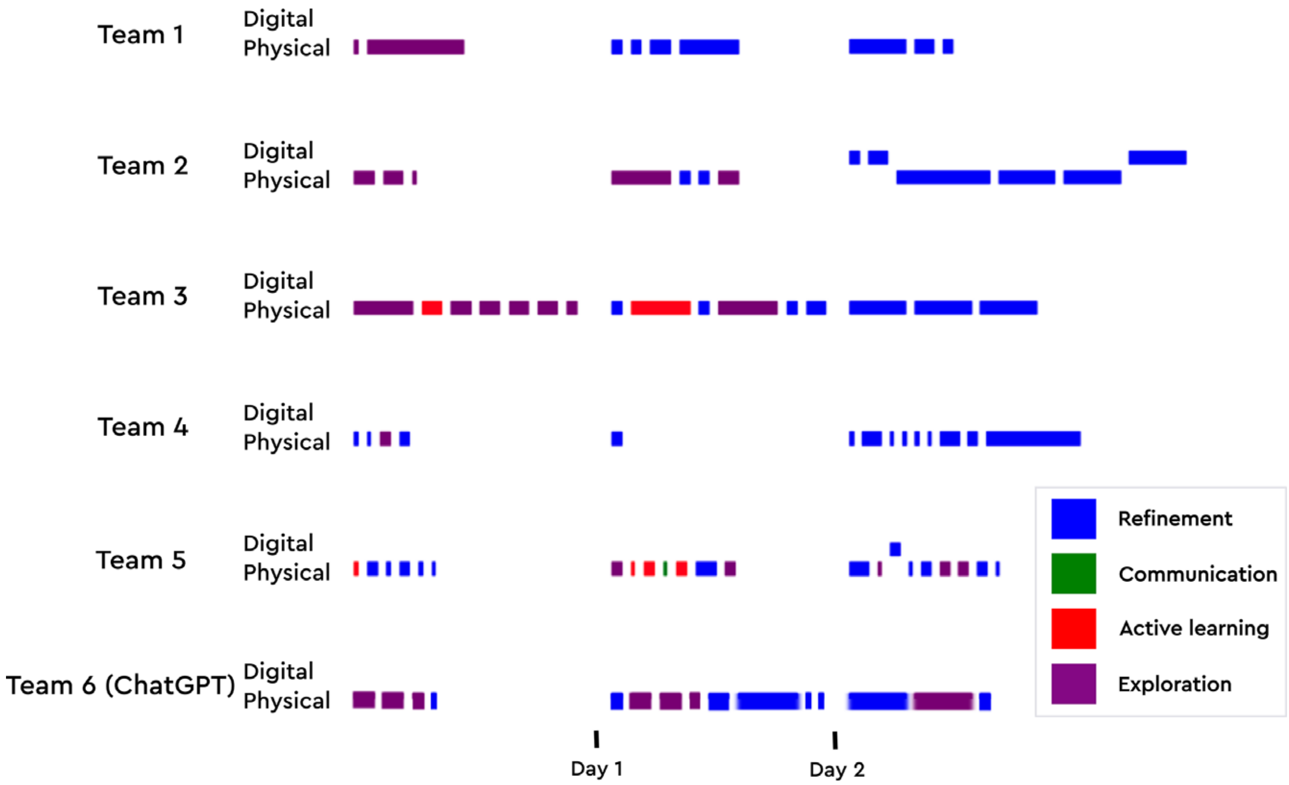

Figure 2. Prototyping timelines.

production time was lower than that of other teams, with an average of 30 min per prototype.

Team 5 made the most prototypes of any team, with 21 physical and one digital prototype. Twelve were refinement prototypes, five were exploration prototypes, and one was a communication prototype. Team 5 spent the least time prototyping and their prototypes took, on average, shorter to make than all other teams, at an average of just over 10 min. The team never spent more than 30 min on a prototype.

Team 6 (ChatGPT) exhibited similar practices to Team 3. The team made 15 physical prototypes and, like most other teams, primarily alternated between two types of prototypes: refinement and exploration. Like Team 3, Team 6 (ChatGPT) did not make prototypes that took longer than 3 h to finish and averaged 0.7 h per prototype. Unlike the other teams, Team 6 (ChatGPT) made an exploration prototype on the final day.

Summary of the LLM's prototyping process and final design

Key interactions and decision points during the hackathon are illustrated in Figure 3, showing inputs (given to the LLM) and outputs (answers from the LLM). Following the initial prompt, which provided the LLM with details about the participants' roles and their own, the participants briefed the LLM on the hackathon's objective and rules. The LLM detailed essential design elements and suggested various NERF launching mechanisms. It advised the team members to search the lab for available materials, like springs and elastic bands, to help choose a concept for prototyping. After performing a 15-min search, the participants reported back to the LLM, which then proposed five design concepts based on the available materials. Ultimately, it recommended a pneumatic launcher as the most promising approach and provided building instructions.

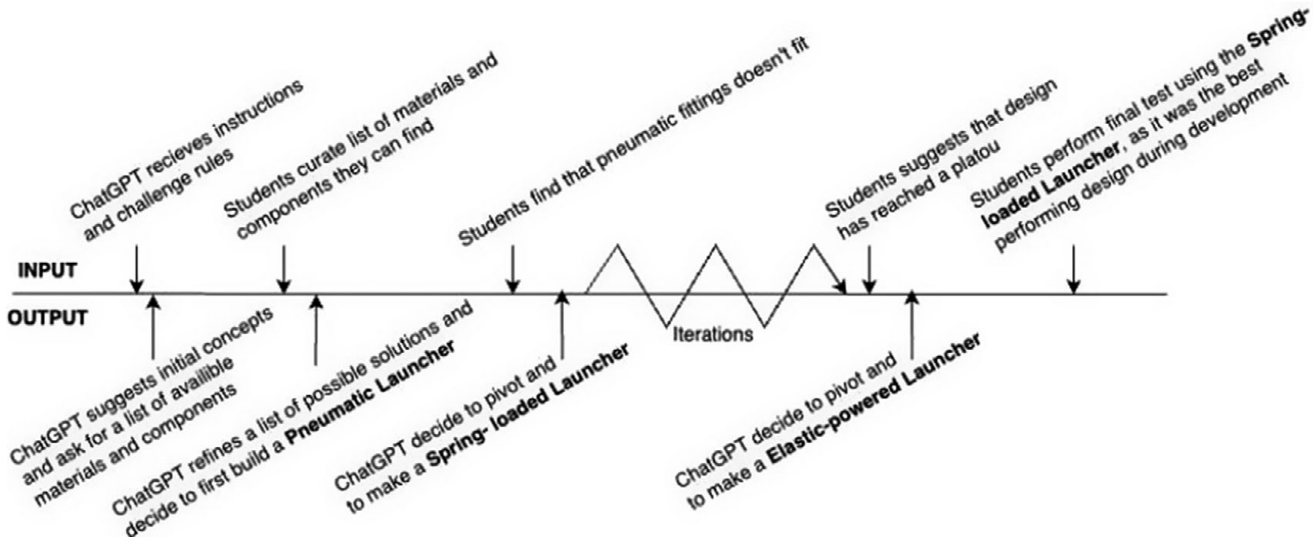
ChatGPT recieves instructions and challenge rules
Students curate list of materials and components they can find
Students find that pneumatic fittings doesn't fit
Students suggests that design has reached a platou
Students perform final test using the **Spring-loaded Launcher**, as it was the best performing design during development
INPUT
OUTPUT
Iterations
ChatGPT suggests initial concepts and ask for a list of availible materials and components
ChatGPT refines a list of possible solutions and decide to first build a **Pneumatic Launcher**
ChatGPT decide to pivot and to make a **Spring- loaded Launcher**
ChatGPT decide to pivot and to make a **Elastic-powered Launcher**

Figure 3. Timeline of key interactions (from Ege et al., 2024d).

When the participants encountered difficulties finding a suitable coupler to connect the bike pump to the initially suggested pressure vessel, they prompted "The bike pump is broken and there are no suitable couplers to the ball valve. The pressure vessel has a 6 mm tube and the ball valve is 1 1/4 inch. What to do next?" the LLM answered, "Alright, given the challenges with the pneumatic launcher, we might want to pivot to another mechanism." and shifted its recommendation to a spring-based launcher concept before providing a new, detailed guide for prototyping.

Following this change in direction, the participants and the LLM exchanged messages to clarify the construction of the envisioned design. At the request of the participants, the LLM broke down the prototyping process into more manageable, clearly defined tasks, specifically covering the spring compression and propulsion mechanism.

Feedback on the initial working prototype revealed a shooting distance of 5 m and prompted a discussion on the setup's configuration. The LLM then outlined potential enhancements and, upon request, supplied a specific action plan for optimization. The dialogue continued with requests for clarifications and updates on the dart's positioning and the spring compression. The LLM proposed a solenoid release mechanism. However, testing showed a much stronger electrical actuator was necessary. When prompted, the LLM responded by listing various alternatives and recommended utilizing a geared DC motor, along with an action plan for further prototyping.

Subsequent iterations focused on refining the launcher based on the LLM's optimization suggestions, such as adding weight and enhancing the platform holding the prototype. These modifications led to a reported shooting distance of 12 m. Efforts to increase distance included creating a free-standing platform and various adjustments to the launcher's components. However, these changes resulted in a reduced shooting distance, prompting the LLM to recommend solutions for overcoming this setback and further optimization strategies. Further enhancements began to yield diminishing returns, suggesting that performance had plateaued.

At this juncture, the study organizers decided that the participants should express interest in exploring alternative concepts to further investigate the LLM's capabilities with additional designs. Hence, the LLM proposed a pivot to an elastic band-powered launcher, complete with a new action plan for prototyping.

This shift led to the creation of a launcher that achieved a 10-m range. The team was then focused on developing a free-standing structure and a remote trigger mechanism. Although this mechanism was successful, it introduced accuracy issues, with the dart not shooting straight. The LLM suggested adjustments but, ultimately, the dart's misdirection persisted, culminating in a detailed strategy to rectify the problem. Ultimately, due to the latest elastic band-powered prototype's directional issues, and the end of the hackathon nearing, the team decided to revert to the more reliable spring-based launcher, which consistently fired in the intended direction. This decision solidified the spring-based launcher as the final design choice, as depicted in Figure 4.

The final design integrates a compression spring housed within an aluminum tube, affixed to a stable platform. The base of the tube is supported by a wooden block for stability. A thin string passes through a hole in the back piece of wood and the spring itself, allowing for spring compression when it is drawn backward. A trigger pin slotted through a hole in the aluminum tube locks the compressed spring. The pin is linked to a geared DC motor powered by a 9-V battery via a fishing line that, when activated, winds the fishing line around its shaft, thus pulling the pin out and launching the dart. Additionally, the launcher is attached to a vertical support with a pivoting mechanism, enabling precise angle adjustments for the launch.

Performance of the final designs

Figure 5 shows the final prototypes developed by the control teams. **Team 1's** design utilized an empty fire extinguisher containing compressed air connected to a metal tube, where the NERF was placed using a rubber tube. The metal tube was fixed to a frame for precise angle adjustments to optimize the NERFs trajectory. **Team 2** decided on a spring-loaded mechanism in which a spring was compressed inside an aluminum tube to store energy. The NERF was placed in the same tube, and it was shot when the spring was released. A DC motor pulled a string attached to the pin holding the compressed spring to fire the design. **Team 3** utilized a plastic bottle connected to a foot pump to build pressure.

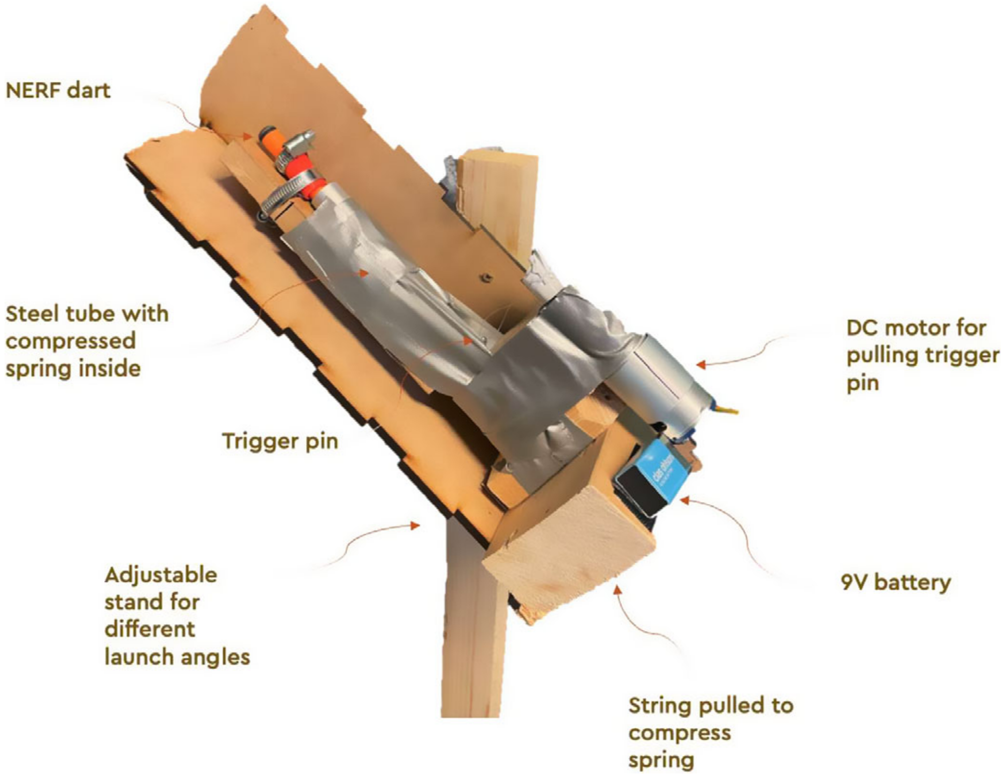
NERF dart
Steel tube with compressed spring inside
Trigger pin
Adjustable stand for different launch angles
DC motor for pulling trigger pin
9V battery
String pulled to compress spring

Figure 4. Final design of Team 6 (ChatGPT) with arrows indicating key components.

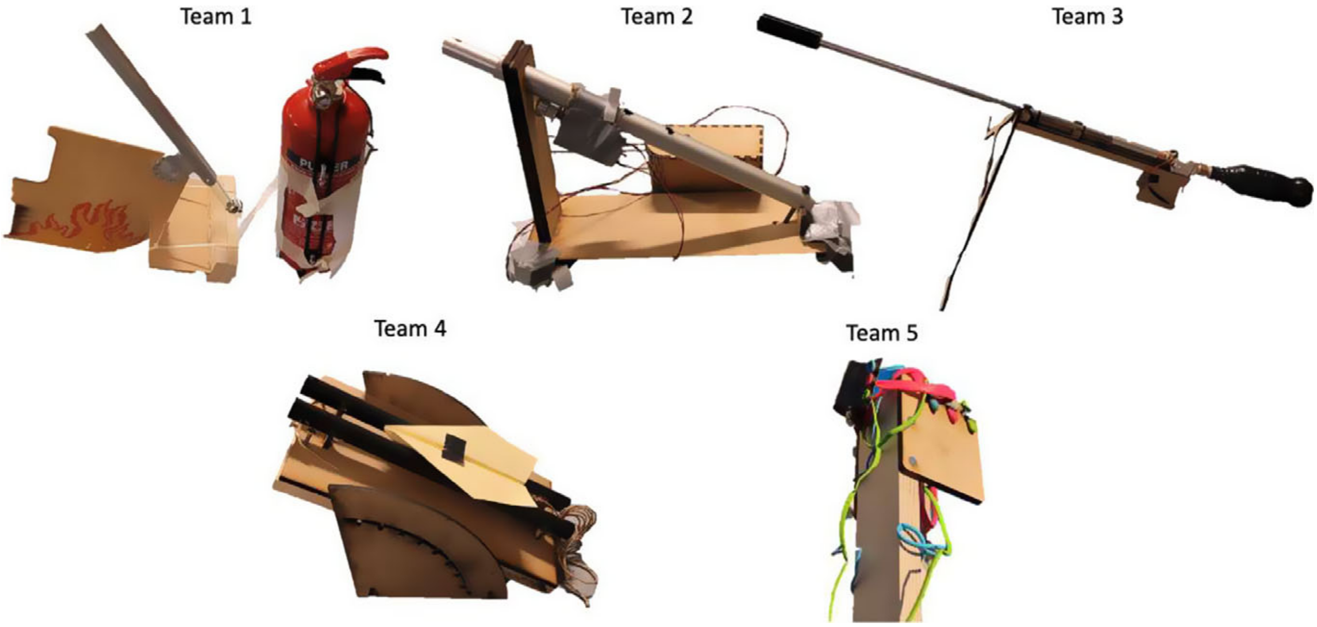

Figure 5. Final designs of control teams.

The plastic bottle was connected to a long barrel in which the NERF was loaded. A solenoid valve was used to release pressure and fire the NERF. The prototype was mounted to a laser-cut tripod to control the trajectory. **Team 4** utilized elastic bands connected to a paper airplane in which the NERF was placed. An adjustable platform controlled the trajectory, and two small pipes guided the paper plane when shot. **Team 5** made a slingshot consisting of balloons connected to a beam. The NERF was placed in a leather pouch connected to the balloons, propelling it as they were stretched and released.

The performance of each team's final design with regard to the distance it could fire an NERF dart and the rank of each team are shown in Table 4. It also shows the average length across teams. Team 3 won the challenge, managing to fire the NERF 37 m, more

https://doi.org/10.1017/S0890060425000010 Published online by Cambridge University Press

Table 4. Performance of teams and rank

Team	Length (m)	Rank
Team 1	6.56	4
Team 2	8.07	3
Team 3	37.66	1
Team 4	5.06	5
Team 5	0.06	6
Team 6 (ChatGPT)	14.8	2
Average	12.0	

than 60% longer than the next team. Team 6 (ChatGPT) finished second, firing its NERF 14.8 m, surpassing the average distance fired across teams by almost 3 m. Team 2 finished third, shooting the NERF 8.1 m, and Team 1 finished fourth with a 6.6 m shot. Teams 4 and 5 shot 5.1 and 0.1 m, respectively, finishing in the fifth and sixth places.

Chat analysis

The result of performing content analysis of the chat log is visualized in Figure 6. The 97 prompts in the chat were assigned 108 codes, of which 87 entries were assigned one code, 9 were assigned 2 codes, and 1 was assigned 3 codes. The most predominant categories in the chat were "Instruction and Guidance" (33 counts) and "Problem-solving and Troubleshooting" (22 counts), followed by "Questions and Clarifications" (20 counts). The three categories most frequently assigned to ChatGPT were "Instruction and Guidance," accounting for 31 of the 78 categories assigned to ChatGPT, 11 counts of "Idea Generation and Conceptualization," and 9 counts of "Decision-making." Dominant categories assigned to participants were "Problem-solving and Troubleshooting," accounting for 21 of 53 codes for the team, 16 counts of "Questions and Clarifications," and 10 counts of "Feedback and Iteration." The two "Instruction and Guidance" instances from participants occurred when ChatGPT requested a brainstorming session, which conflicted with the experiment's rule that all decision-making be left to the LLM. In response, participants reminded ChatGPT to adhere to its role and not ask for brainstorming again. The instance of the participant prompting "Decision-making" stems from when the organizers instructed the participants to prompt a pivot from the LLM when their performance of iterations on the spring-based design plateaued.

Discussion

Comparing design practices

In analyzing the prototyping practices of participants, it is clear that most teams share comparable characteristics. Teams made similar quantities of prototypes, mainly pivoting between exploration and refinement prototypes, with refinement prototypes most frequently used toward the end of the hackathon. Teams reported making higher amounts of prototypes on the first and final day. Fast-

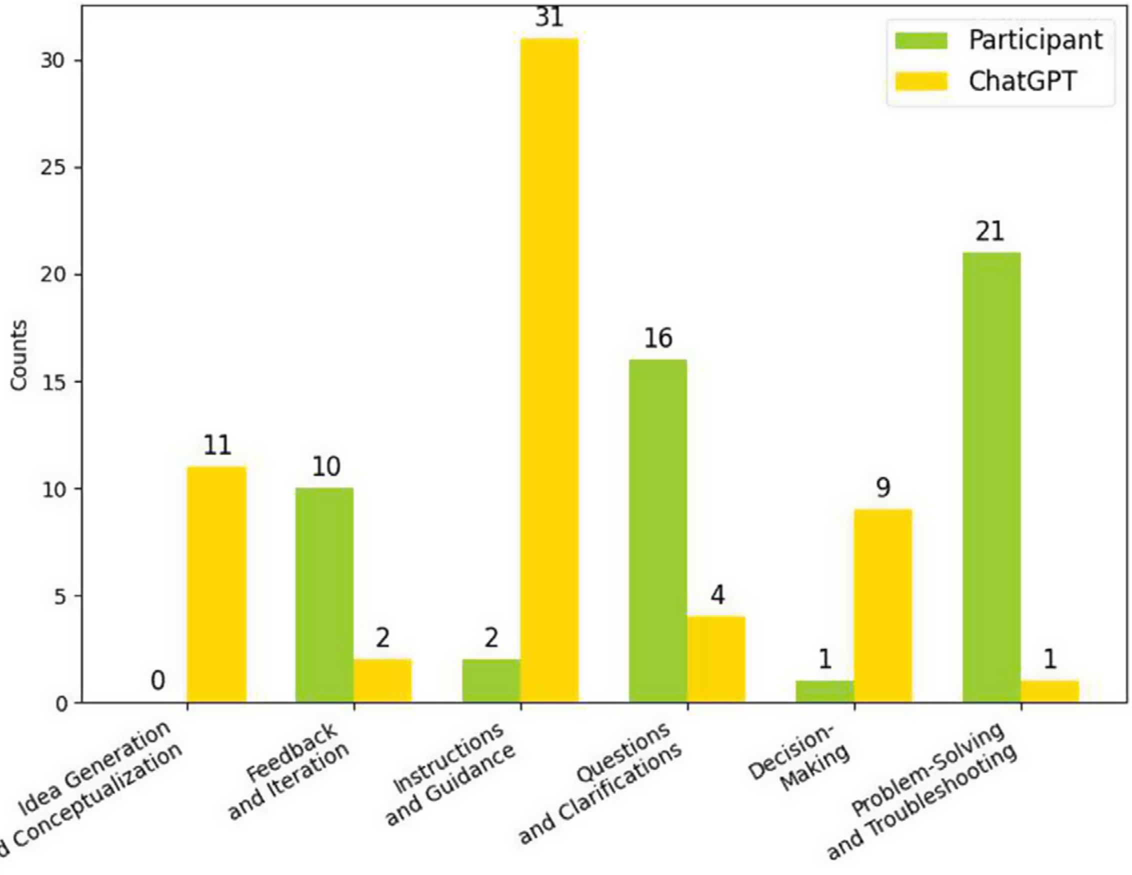

Figure 6. Distribution of code assignments between ChatGPT and participants.

produced prototypes mainly occurred on the first day, signifying low complexity and rapid iterations. In contrast, longer production times were more frequent on the final day, signifying more time-consuming optimization efforts. These characteristics align well with the practices previously reported in similar events (Ege et al., 2024a). Notably, Team 6 (ChatGPT)s' approach closely mirrored that of the most successful control team, leading to the key finding: **(KF 1) The LLM shows similar prototyping practices to humans**, particularly concerning the amount and type of prototypes made, aligning closely with the practices of the winning team.

Prototyping and final design

Upon receiving the initial prompt and a list of available materials, the LLM suggested a list of five possible working principles, all of which sounded like reasonable ideas of how to propel a foam dart. Its first suggestion was to make a Pneumatic launcher, mirroring the decision of the best-performing control team. The list also contained a spring-based concept in which a spring is compressed and released behind a piston, pushing air before it to propel the dart. This is analogous to how most commercial NERF dart shooters work. Further, the LLM proposed concepts outside the three main concepts previously described, including a slingshot design, a gravity-driven mechanism, and a centrifugal force-based design, all of which have the potential to propel a dart, although it chose not to pursue these further through physical realization. This leads to the key finding: **(KF 2) The LLM shows promising capabilities for concept generation** by describing various reasonable working principles.

The results illustrated in Figure 1 indicate that the LLM did not produce highly creative outputs, as it mirrored the conceptual choices of the control teams, focusing on three main recurring concepts. While GPT initially generated a diverse list of ideas during the brainstorming phase, it ultimately chose to prototype solutions only for the more common concepts, leaving other, more unconventional ideas unexplored. LLMs are trained on vast datasets and, as a result, generate outputs that align with the patterns and structures present in their training data. While the LLM simulates creativity by suggesting a broad range of ideas, its decision-making seems to favor familiar concepts that are grounded in common knowledge, that is, the "safe" option. Consequently, it may prioritize solutions that align with established patterns rather than pursuing more novel or unconventional ideas.

Unlike the winning team, the AI abandoned the pneumatics concept after one iteration because the valve participants initially found did not fit. Despite participants having access to equipment common in makerspaces/fab labs, such as 3D printers and CNC machines (that the LLM suggested using in other instances), the LLM recommended pivoting without thoroughly exploring alternative solutions, only reasoning, "Alright, given the challenges with the pneumatic launcher, we might want to pivot to another mechanism." This observation leads to the key finding: **(KF 3) The LLM can interpret feedback as a failure rather than a challenge, leading it to abandon promising concepts prematurely**, potentially affecting its problem-solving process and depth of exploration of concepts. This is perhaps unlike what an experienced human designer would do, as they would leverage a setback as a learning opportunity (Lande and Leifer, 2009; Lauff et al., 2018) and draw on previous skill-based and implicit knowledge and past experiences to consider multiple solutions to arising problems (Vestad et al., 2019).

Although prematurely abandoning a promising concept that was used by the best-performing team, the LLM did, in fact, successfully propose a working design. From idea generation to building multiple iterations and solving emerging problems, the LLM equipped its human team members with building instructions that ended up with a prototype that could reliably fire NERF darts, leading to the key finding: **(KF 4) The LLM was able to design a physical, functional prototype to perform a simple task** with the same working principle as commercially available solutions for the same task. This finding is further strengthened by some of the emerging problems it was able to understand and overcome. For example, an early prototype had a problem where a spring was compressed between two bars, leading the participants to prompt: "When constructing the mechanism with two bars and a spring in between, we noticed that the spring bends outwards and does not stay straight under compression." The LLM correctly identified this as spring buckling, answering "The bending of the spring under compression is a phenomenon known as 'buckling'. It's a common issue, especially with longer springs. To counteract this, we'll need to guide and constrain the spring during compression" and adding "Place a cylindrical tube around the outside of the spring. The internal diameter of this tube should be slightly larger than the external diameter of the spring. This tube will act as a guide, ensuring the spring compresses straight down."

Although many design decisions seem to have solid reasoning behind them, we question the decision to use a push-button-activated DC motor to pull the firing pin instead of manually pulling it. Similarly, the LLM, at one point, suggested motorizing the spring compression instead of manually pulling back a spring, which would add complexity without improvements or additional benefits. This leads to the following key insight: **(KF 5) The LLM risks adding unnecessary complexity to its designs**. The participants even expressed to the organizers that they were embarrassed over what they were prototyping, as they understood that these were unnecessary and bad ideas, and particularly that they could not explain to the other teams that the ideas were not theirs. However, it is noteworthy that Team 2 used the same firing mechanism and working principle for propulsion, indicating that although more complex than necessary, the added complexity is similar to that of a human team. As teams built prototypes in the same space, it is uncertain whether Team 2 made the same design decision independently or was influenced by seeing Team 6 (ChatGPT) design theirs, but it illustrates that the LLM concludes similarly to that of human teams.

Comparing final designs and performance

When comparing final designs across teams, the LLM's design was similar to other teams' designs. Like Teams 2 and 4, the LLM opted for a Spring-based launcher. Its design was the best performing among the spring-based launchers and was only beaten by Team 3's pneumatic launcher. This led to the key insight: **(KF 6) The LLMs' design capabilities proved competitive against fifth-year engineering students** by finishing second among six teams.

It is necessary to state that a significant factor contributing to the poor performance of some teams was the weather conditions during the outdoor test. Cold temperatures affected the viscoelastic tubing and silicone in Teams 1 and 5's designs, leading to poor performance. The LLM, however, refined its prototype through multiple iterations to achieve high reliability while competing under the same conditions, in contrast to the teams performing poorly during the final test. The LLM, at one point, even suggested

https://doi.org/10.1017/S0890060425000010 Published online by Cambridge University Press

contingency planning, indicating that it was preparing for unexpected events during the final test.

Chat analysis and observations

The communication between the LLM and the participants presented notable challenges, highlighted by a significant portion of the prompts – 20 out of 97 – being categorized as "Questions and Clarification." The initial design concepts shared by the LLM were often vague despite explicit instructions for the LLM to make all design decisions. For example, the LLM's guidance on creating a spring-based launcher mechanism was too general. It required further prompting for clarification, illustrated by the answer, "Design a mechanism where the spring can be compressed and then released to launch the dart." It took 18 exchanges before the initial prototype reached a stage where it could be tested, a delay attributed to initial design flaws that went unnoticed by the LLM until they were gradually addressed with new iterations. This experience reveals the LLM's limitations in guiding the transition from broad concepts to addressing specific subsystem issues.

The initial dialogue concerning the spring-based launcher focused on a mechanism for propelling the dart by physically hitting it using what the LLM described as a "moving block." It did not consider a mechanism to push air behind the dart. In subsequent interactions, the concept evolved to include "an airtight piston," indicating a significant mid-process shift in the design principle. This shift necessitated additional prompts to clarify and refine a specific design not initially explained to the participants. At one point, the LLM also asked the participants to "Use ropes or wires, attached to the moving block and running through pulleys at the top of the frame, to assist in pulling the block upwards (…) ," without mentioning anything about pulleys before that point. This leads to the key finding: **(KF 7) The LLM is unable to communicate design intent effectively**, necessitating time-consuming discussion to comprehend instructions. An analysis of the interaction patterns further demonstrates the disconnect. Prompts from participants seeking solutions to problems often resulted in conceptual ideas rather than direct guidance. To obtain detailed "Instructions and Guidance," participants had to navigate through a cycle of "Questions and Clarifications." This pattern suggests that while the LLM provided detailed, step-by-step instructions for well-defined, narrow queries, it tended to revert to ideation in response to broader conceptual prompts. The effectiveness of the communication seemed to hinge on the questions' specificity. General inquiries often led the LLM into a creative ideation mode, even when practical solutions were sought. This tendency might be beneficial for collaborative purposes but proved challenging with the LLM acting as the team lead, necessitating additional prompts to elicit detailed descriptions.

The participants also noticed the LLM's tendency to overlook previously given information. Despite being informed that the participants would not contribute suggestions, it still implied the need for a brainstorming session among them. Similarly, the LLM recurrently failed to remember the team's size, suggesting multiple-member strategies even though it had been made clear that the team consisted of only two people. This necessitated frequent reminders from the participants about key details of their setup, as evidenced when it suggested, "If you have multiple team members, consider brainstorming and collaborating to generate fresh ideas or approaches. Sometimes, a new perspective can lead to breakthroughs"; thus, also forgetting the objective role of the participants in which they were not allowed to provide their perspectives and suggestions. This pattern leads to the key finding: **(KF 8) The LLM faces challenges maintaining continuity and relevance in responses to the project's specific context.**

Much like a novice designer, ChatGPT clings to one concept and is reluctant to try something else (Purcell and Gero, 1996), illustrated by the answer: "If you've iterated through multiple design improvements and observed that the increase in launch distance has plateaued, it's a good indication that you've reached a point of diminishing returns in terms of design changes. At this stage, here are some suggestions for what you can do next," followed by a list of actions including more testing, optimizing the launch, improving the aerodynamics of the dart (which clearly goes against the challenge rules), practicing, and contingency planning. Only after explicitly stating that: "We have now iterated based on your feedback and observed that we have reached a plateau in performance. We are interested in testing one more of the original ideas. What do you suggest?," ChatGPT was willing to pivot to a different mechanism. At this point, the team had spent a considerable amount of time optimizing the previous design, with little time left to build and test the new concept. These observations lead to the key insight: **(KF 9) ChatGPT experiences and is limited by design fixation**, both regarding not wanting to abandon a working concept and specific details provided in prompts (e.g., the list of the available materials curated at the start of the hackathon). When planning the study and testing different initial prompts, it became clear that the LLM often fixated on specific details in the prompts. For example, when providing the LLM with manufacturing capabilities available to the participants, it would often fixate on the last machine on the list. If that happened to be a 3D printer, the LLM would suggest 3D printing each prototype going forward. Likewise, if the last machine were a laser cutter, it would suggest laser-cutting prototypes. This necessitated the general description of "fab lab/makerspace" in the prompt instead of listing all manufacturing capabilities. Notably, the LLM asked participants to curate a list of available materials but never asked what manufacturing capabilities were available to them.

Recommendations for using current LLM for engineering design

This study intentionally utilized the LLM in an extreme capacity as the sole decision-maker in an engineering design process, a scenario that cannot be advocated for practical applications. The purpose was rather to objectively benchmark the strengths and weaknesses of current LLMs against human capabilities. Based on the observations and insights from the participants, it is clear that the results of using it could have been drastically improved through minimal critical thinking by them. Based on the findings of our study, we offer the following recommendations for effectively integrating current LLMs into the design process:

1. **Leverage the LLM for ideation, but guide with human oversight:** Utilize its capability to generate a broad spectrum of concepts during the ideation phase to enhance creativity and explore a wider range of initial ideas. Human designers should provide oversight to evaluate the feasibility and relevance of these ideas. LLM-generated suggestions can lead to innovative concepts; however, without critical human judgment, they may lack practicality or alignment with project goals. Therefore, the ideation process should always involve human filtering and refinement to ensure viable outputs. LLMs offer breadth but humans ensure depth and applicability.

https://doi.org/10.1017/S0890060425000010 Published online by Cambridge University Press

2. **Maintain human decision-making oversight:** LLMs can suggest abandoning certain promising ideas too early due to biases or fixation. Human decision-makers must intervene to ensure promising ideas are not prematurely discarded. This requires human evaluation at key decision points, particularly to counteract the LLM's potential design fixation and ensure that the most innovative and feasible ideas are pursued.
3. **Implement iterative feedback loops with the LLM:** Iterative feedback is a standard practice in design, but with LLMs, it becomes even more critical. LLMs generate outputs based on the input provided, and without regular feedback, their suggestions can diverge from the original design intent or become overly complex. Continuous feedback ensures the LLM stays aligned with design goals and refines its outputs in a way that mirrors the evolving design process.
4. **Use structured prompts and custom templates for consistency:** Streamline communication with the LLM through tailored prompts and templates, making instructions clearer and easier to follow, and reducing the occurrence of vague or off-target responses. This ensures more consistent, relevant outputs that align with design goals.
5. **Explicitly prompt the LLM to consider alternatives:** Design fixation is a common challenge in any design process, but LLMs seem particularly prone to focusing on the most immediate or conventional solutions. To counteract this, designers must explicitly prompt the LLM to explore alternative approaches and avoid narrowing its focus prematurely. This strategy encourages the LLM to generate a wider array of solutions and helps prevent it from becoming fixated on suboptimal ideas.
6. **Assign specific tasks at the subsystem level:** Direct the LLM to focus on detailed explanations and solutions for specific project parts to enhance clarity and avoid vague or unhelpful responses. This method also helps when aiming to solve complex problems, as these become more manageable when broken down into smaller, more focused tasks. By addressing each subsystem individually, LLMs can provide more detailed and actionable outputs, contributing to the resolution of larger, more intricate design challenges. This approach allows designers to integrate these subsystem solutions into a cohesive, optimized final deliverable.

Limitations and consideration

The study is limited by investigating the distinctive setting of a hackathon, with its specific characterizations that may not fully represent the larger range of design and prototyping processes found in professional and educational settings. However, hackathons have been shown to mirror key aspects of design (Goudswaard et al., 2022; Ege et al., 2024a) and, thus, appropriate for studying design processes. Nonetheless, the insights gained are contextualized within this distinctive setting and extrapolating these results to other design contexts should be done with caution.

This study focused exclusively on prototyping for a single task – an NERF firing device – thereby constraining the breadth of insights into how the LLM might perform with different tasks. This specificity potentially limits the transferability of our findings to other cases. However, design challenges often involve multi-objective criteria where trade-offs between competing factors, such as performance, resource constraints, usability, and manufacturability, must be considered. Scaling the experiment to include such multi-objective tasks would require adapting the approach to accommodate a broader set of evaluation metrics.

Team 6 (ChatGPT) had 1 year more experience than control teams, a factor that could potentially influence outcomes. Choosing slightly more experienced participants for the team was a measure taken to ensure they had the abilities necessary to use all the available equipment and skills to perform the instructions of the LLM. Measures were taken to balance the experience gap, as the team was instructed to act objectively and not to contribute ideas or insights, thus neutralizing any differences. Further, participants' backgrounds in mechanical engineering may have inadvertently influenced the nature of the prompts given to the LLM, potentially introducing a bias. For instance, inquiries about the fit tolerances between parts may not typify questions from a novice, suggesting some experienced-based bias in participants' prompts.

The reliance on participants' objectivity is a further limitation. While the LLM provided reasonable ideas and directions, the physical realization and implementation of these concepts required human intervention, which introduces variability in decision-making and problem-solving. The degree of human engagement, including the phrasing of follow-up prompts or seeking clarifications, may have influenced the model's outputs and thus shaped the outcomes. The LLM's inability to effectively express design intent, introducing ambiguity and vagueness, might skew results due to interpretation from the participants. Participants attempted to mitigate this by seeking clarification from the LLM throughout the hackathon, yet this interactive process may itself influence the outcomes. Ultimately, while the goal was for the LLM to operate autonomously, human interaction played an unavoidable role, making it possible that different participants or different interaction styles could have produced varied outcomes. This highlights that the experiment not only assessed the LLM's capabilities but also indirectly the influence of human–LLM interaction on design outcomes.

The replicability of the study may be compromised by the evolving nature of the LLM used. Results obtained on October 18 and 19, 2023, might not be replicated in subsequent uses due to updates and changes to the model. LLMs, including the models used in this study, are nondeterministic by nature. This means that the same prompt may yield different outputs each time it is submitted. This characteristic poses challenges for replicability in design tasks, as identical inputs may not consistently generate identical results, especially as LLMs evolve. Moreover, the models themselves are continually updated, often incorporating new data and improved algorithms. As a result, both the version of the LLM used and the underlying dataset may change over time, making it difficult to replicate the exact conditions of this study in the future. The nondeterministic nature of LLMs introduces an additional layer of complexity when evaluating their performance in engineering design. While this study reflects the capabilities of LLMs as they existed at the time of the research, the rapid evolution of these technologies means that findings related to their limitations – such as difficulties with maintaining context, short attention spans, or logical reasoning – may not hold true as models improve.

Further work

The authors recommend further investigation of the recommendations presented in this article for design-related tasks. Although the limitations placed on the participants were necessary to elicit the strengths and weaknesses of the LLM itself, they are not realistic in real-world design situations. Real-world applications would likely require LLMs to function as part of a human–AI hybrid team rather than leading the process. Future research should systematically

https://doi.org/10.1017/S0890060425000010 Published online by Cambridge University Press

investigate how LLMs can serve as design support tools, offering insights and suggestions while working in tandem with human designers.

To systematically study the performance of LLMs as design support, a structured methodology can be developed, focusing on several key areas:

- **Collaboration:** Future experiments should explore how LLMs perform when integrated as team members, providing support in ideation, problem-solving, or generating alternatives, while human designers retain decision-making control. This setup would allow for an evaluation of how well LLMs contribute to a team's creativity and problem-solving capabilities, without being the sole driver of the design process.
- **Performance of LLMs as design support:** Future research should systematically evaluate the performance of LLMs as collaborative tools in the design process. This can be measured through several performance metrics, including time efficiency, design quality, and problem-solving capabilities. Key performance indicators could include the LLM's ability to generate functional solutions, suggest innovative ideas, and assist in optimizing design iterations. Additionally, performance can be assessed by analyzing how effectively the LLM helps reduce the cognitive load on human designers by handling routine tasks or complex computations. By tracking improvements in overall design performance and the speed of iteration cycles, researchers can better understand the role LLMs play in accelerating and enhancing the engineering design process.
- **Multi-objective task:** A future study could task the LLM with balancing multiple criteria during the design process to investigate the impact of multiple objects that more accurately mirror industrial practices. For instance, instead of optimizing solely for the distance an NERF dart can be fired, the LLM could also consider factors like cost, durability, ease of use, and even the aesthetics of the final design. To enable this, a structured framework would be necessary, where the LLM is provided with weighted criteria or a hierarchy of objectives. The LLM could then generate design solutions by optimizing for combinations of these factors.
- **Role of visual and parametric outputs:** Investigating AI-generated visual outputs (e.g., CAD models) in conjunction with written instructions could provide new insights into how LLMs can assist in producing usable designs. Combining parametric design approaches with AI-generated CAD models could further enhance usability and lead to intuitive human–AI interactions, particularly in the context of engineering design.
- **Industry expert evaluation:** A study involving industry experts performing a design task, with one group instructed to use ChatGPT, would provide valuable insights into how AI tools integrate into professional workflows. By observing the performance of experienced designers who are familiar with industry standards and constraints, researchers can better assess the practical utility of LLMs in real-world settings. This setup is particularly relevant because it mirrors industrial practices more closely, where time, cost, and product viability are crucial. Moreover, paying experts ensures a level of engagement and commitment that can lead to more accurate assessments of ChatGPT's ability to contribute meaningfully to high-stakes, professional design environments.

These approaches will help build a body of knowledge on how to effectively integrate LLMs into real-world design workflows, ultimately allowing for practical and well-functioning human–AI cooperation. Future studies that use similar experimental setups with different LLM versions or evolving datasets would be valuable in validating and strengthening the findings of this study. Replicating this research across various iterations of LLMs and datasets can provide more robust insights into the consistency of LLM behavior over time and across versions. Additionally, by conducting comparative studies with other models and use cases, researchers can better understand how these limitations might shift or be resolved as the technology evolves.

Conclusion

This study has compared the design practices and performance of an LLM, specifically ChatGPT 4.0, against fifth-year engineering students in a prototyping hackathon. It provides nine key findings, such as showing that the LLM had similar prototyping practices to human participants and proved competitive against them by finishing second among six teams. The LLM successfully provided building instructions to realize a physical, functional prototype and solved concrete and physical obstacles along the way. The concept generation capabilities of the LLM were particularly good. Among the limitations of the LLM is that it prematurely gave up on concepts when meeting what the authors perceived as minor difficulties and added unnecessary complexity to some of the designs. Communication between the LLM and participants was also challenging, as it often gave vague, too general, or unclear descriptions and had trouble maintaining continuity and relevance in answers due to forgetting previously given information. The LLM also experienced design fixation by continuing the iterations of one concept even though returns diminished instead of pivoting and trying alternative solutions. Based on these findings, we propose six recommendations for implementing current LLMs like ChatGPT in the design process, including leveraging it for ideation, ensuring human decision-making oversight, implementing iterative feedback loops, explicitly prompting it to consider alternatives, and assigning the LLM-specific tasks at a subsystem level.

Acknowledgments. The authors would like to thank the TrollLabs Open hackathon participants for participating in this study. While preparing this work, the authors used ChatGPT by OpenAI to improve the language and readability of the manuscript. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

Funding. The Research Council of Norway supports this research through its industrial PhD funding scheme (grant number 321386).

Competing interest. The authors declare no competing interests.

Data availability statement. The data that support the findings of this study are openly available in Zenodo.org at http://doi.org/10.5281/zenodo.10495102, reference number 10495102, and described in the following data article: Ege, Øvrebø, Stubberud, Berg, Elverum, et al., 2024.

References

Ataei M, **Cheong H**, **Grandi D**, **Wang Y**, **Morris N and Tessier A** (2024) Elicitron: An LLM agent-based simulation framework for design requirements elicitation. Preprint, arXiv:2404.16045. https://doi.org/10.48550/arXiv.2404.16045.

Camburn B, **Viswanathan V**, **Linsey J**, **Anderson D**, **Jensen D**, **Crawford R**, **Otto K and Wood K** (2017) Design prototyping methods: State of the art in strategies, techniques, and guidelines. *Design Science* **3**, e13. https://doi.org/10.1017/dsj.2017.10.

Carlile P (2002) A pragmatic view of knowledge and boundaries: Boundary objects in new product development. *Organization Science* **13**, 442–455. https://doi.org/10.1287/orsc.13.4.442.2953.

Chen L, **Wang P**, **Dong H**, **Shi F**, **Han J**, **Guo Y**, **Childs PRN**, **Xiao J and Wu C** (2019) An artificial intelligence based data-driven approach for design ideation. *Journal of Visual Communication and Image Representation* **61**, 10–22. https://doi.org/10.1016/j.jvcir.2019.02.009.

Dortheimer J, **Martelaro N**, **Sprecher A and Schubert G** (2024) Evaluating large-language-model chatbots to engage communities in large-scale design projects. *AI EDAM* **38**, e4. https://doi.org/10.1017/S0890060424000027.

Dow SP, **Heddleston K and Klemmer SR** (2009) The efficacy of prototyping under time constraints. In *Proceedings of the Seventh ACM Conference on Creativity and Cognition*, pp. 165–174. https://doi.org/10.1145/1640233.1640260

Ege DN, **Goudswaard M**, **Gopsill J**, **Steinert M and Hicks B** (2024a) What, how and when should I prototype? An empirical study of design team prototyping practices at the idea challenge hackathon. *Design Science* **10**, e22. https://doi.org/10.1017/dsj.2024.16

Ege DN, **Øvrebø H**, **Stubberud V**, **Berg MF**, **Elverum C**, **Steinert M and Vestad H** (2024b) TrollLabs open dataset. Zenodo. https://doi.org/10.5281/zenodo.10495102.

Ege DN, **Øvrebø HH**, **Stubberud V**, **Berg MF**, **Elverum C**, **Steinert M and Vestad H** (2024c) The TrollLabs open hackathon dataset: Generative AI and large language models for prototyping in engineering design. *Data in Brief* **54**, 110332. https://doi.org/10.1016/j.dib.2024.110332.

Ege DN, **Øvrebø HH**, **Stubberud V**, **Berg MF**, **Steinert M and Vestad H** (2024d) Benchmarking AI design skills: Insights from ChatGPT's participation in a prototyping hackathon. In *Proceedings of the Design Society* **4**, 1999–2008. https://doi.org/10.1017/pds.2024.202

Elverum C and Welo T (2014) The role of early prototypes in concept development: Insights from the automotive industry. *Procedia CIRP* **21**, 491–496. https://doi.org/10.1016/j.procir.2014.03.127.

Erichsen JF, **Sjöman H**, **Steinert M and Welo T** (2021) Protobooth: Gathering and analyzing data on prototyping in early-stage engineering design projects by digitally capturing physical prototypes. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* **35**(1), 65–80. https://doi.org/10.1017/S0890060420000414.

Filippi S (2023) Measuring the impact of ChatGPT on fostering concept generation in innovative product design. *Electronics* **12**(1616), 3535. https://doi.org/10.3390/electronics12163535.

Giunta L, **Gopsill J**, **Kent L**, **Goudswaard M**, **Snider C and Hicks B** (2022) Pro2booth: Towards an improved tool for capturing prototypes and the prototyping process. *Proceedings of the Design Society* **2**, 415–424. https://doi.org/10.1017/pds.2022.43.

Glaser B and Strauss A (1999) *Discovery of Grounded Theory: Strategies for Qualitative Research*. Routledge, New York, NY, USA. https://doi.org/10.4324/9780203793206.

Goudswaard M, **Kent L**, **Giunta L**, **Gopsill J**, **Snider C**, **Valjak F**, **Christensen KA**, **Felton H**, **Ege DN**, **Real RM**, **Cox C**, **Horvat N**, **Kohtala S**, **Eikevåg SW**, **Martinec T**, **Perišić MM**, **Steinert M and Hicks B** (2022) Virtually hosted hackathons for design research: Lessons learned from the international design engineering annual (idea) challenge 2021. *Proceedings of the Design Society* **2**, 21–30. https://doi.org/10.1017/pds.2022.3.

Haase J and Hanel PHP (2023) Artificial muses: Generative artificial intelligence chatbots have risen to human-level creativity. *Journal of Crivity* **33**(3), 100066. https://doi.org/10.1016/j.yjoc.2023.100066.

Houde S and Hill C (1997) What do prototypes prototype? In MG Helander, TK Landauer and PV Prabhu (eds), *Handbook of Human-Computer Interaction*, 2nd edition. North-Holland, Amsterdam. pp. 367–381. https://doi.org/10.1016/B978-044481862-1.50082-0.

Hu X, **Tian Y**, **Nagato K**, **Nakao M and Liu A** (2023) Opportunities and challenges of ChatGPT for design knowledge management. *Procedia CIRP* **119**, 21–28. https://doi.org/10.1016/j.procir.2023.05.001.

Hwang AH-C (2022) Too late to be creative? AI-empowered tools in creative processes. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pp. 1–9. https://doi.org/10.1145/3491101.3503549.

Jensen MB and Steinert M (2020) User research enabled by makerspaces: Bringing functionality to classical experience prototypes. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* **34**, 315–326. https://doi.org/10.1017/S089006042000013X.

Khanolkar PM, **Vrolijk A and Olechowski A** (2023) Mapping artificial intelligence-based methods to engineering design stages: A focused literature review. *AI EDAM* **37**, e25. https://doi.org/10.1017/S0890060423000203.

Lai Y, **Chen H-J and Yang C** (2023) Exploring the impact of generative artificial intelligence on the design process: Opportunities, challenges, and insights. *Artificial Intelligence, Social Computing and Wearable Technologies* **113**(113). https://doi.org/10.54941/ahfe1004178.

Lande M and Leifer L (2009) Prototyping to learn: Characterizing engineering student's prototyping activities and prototypes. In *DS 58-1: Proceedings of ICED 09, The 17th International Conference on Engineering Design, 1.*

Lauff C, **Kotys-Schwartz D and Rentschler M** (2018) What is a prototype? What are the roles of prototypes in companies? *Journal of Mechanical Design* **140**, 061102. https://doi.org/10.1115/1.4039340.

Lauff CA, **Knight D**, **Kotys-Schwartz D and Rentschler ME** (2020) The role of prototypes in communication between stakeholders. *Design Studies* **66**, 1–34. https://doi.org/10.1016/j.destud.2019.11.007.

Liao J, **Hansen P and Chai C** (2020) A framework of artificial intelligence augmented design support. *Human-Computer Interaction* **35**(5–6), 511–544. https://doi.org/10.1080/07370024.2020.1733576.

Lim YK, **Stolterman E and Tenenberg J** (2008) The anatomy of prototypes: Prototypes as filters, prototypes as manifestations of design ideas. *ACM Transactions on Computer-Human Interaction (TOCHI)* **15**(2), 7:1–7:27. https://doi.org/10.1145/1375761.1375762.

Maclachlan RJR, **Adams R**, **Lauro V**, **Murray M**, **Magueijo V**, **Flockhart G and Hasty W** (2024) Chat-GPT: A clever search engine or a creative design assistant for students and industry? In *DS 131: Proceedings of the International Conference on Engineering and Product Design Education (EPDE 2024)*, pp. 414–419. https://doi.org/10.35199/EPDE.2024.70.

Memmert L, **Cvetkovic I and Bittner E** (2023) Human-AI collaboration in conceptualizing design science research studies: Perceived helpfulness of generative language model's suggestions. ECIS 2023 Research Papers. 405. https://aisel.aisnet.org/ecis2023_rp/405

Menold J, **Jablokow K and Simpson T** (2017) Prototype for x (pfx): A holistic framework for structuring prototyping methods to support engineering design. *Design Studies* **50**, 70–112. https://doi.org/10.1016/j.destud.2017.03.001.

Mountstephens J and Teo J (2020) Progress and challenges in generative product design: A review of systems. *Computers* **9**(44), 80. https://doi.org/10.3390/computers9040080.

Oh S, **Jung Y**, **Kim S**, **Lee I and Kang N** (2019) Deep generative design: Integration of topology optimization and generative models. *Journal of Mechanical Design* **141**, 1. https://doi.org/10.1115/1.4044229.

Olsson T and Väänänen K (2021) How does AI challenge design practice? *Interactions* **28**(4), 62–64. https://doi.org/10.1145/3467479.

Otto K and Wood K (2001) *Product Design: Techniques in Reverse Engineering and New Product Development*. Upper Saddle River, NJ: Prentice Hall.

Purcell AT and Gero JS (1996) Design and other types of fixation. *Design Studies* **17**(4), 363–383. https://doi.org/10.1016/S0142-694X(96)00023-3.

Regenwetter L, **Nobari AH and Ahmed F** (2022) Deep generative models in engineering design: A review. *Journal of Mechanical Design* **144**(7), 071704. https://doi.org/10.1115/1.4053859.

Salikutluk V, **Koert D and Jäkel F** (2023) Interacting with large language models: A case study on AI-aided brainstorming for guesstimation problems. In P Lukowicz, S Mayer, J. Koch, J Shawe-Taylor and I Tiddi (eds), *Hhai 2023: Augmenting Human Intellect: Proceedings of the Second International Conference on Hybrid Human-Artificial Intelligence*, June 26–30, 2023, vol. **368**. Munich, Germany: IOS Press, pp. 153–167. https://doi.org/10.3233/FAIA230081.

Schrage M (1996) *Cultures of Prototyping*. New York: Association for Computing Machinery, pp. 191–213. https://doi.org/10.1145/229868.230045.

The Next Wave of Intelligent Design Automation. (2018) Harvard Business Review. https://hbr.org/resources/pdfs/comm/autodesk/The.Next.Wave.of.Intelligent.Design.Automation.pdf

Tholander J and Jonsson M (2023) Design ideation with AI - Sketching, thinking and talking with generative machine learning models. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, pp. 1930–1940. https://doi.org/10.1145/3563657.3596014.

https://doi.org/10.1017/S0890060425000010 Published online by Cambridge University Press

Thoring K, **Huettemann S and Mueller RM** (2023) The augmented designer: A research agenda for generative AI-enabled design. *Proceedings of the Design Society* **3**, 3345–3354. https://doi.org/10.1017/pds.2023.335.

Tiro D (2023) The possibility of applying ChatGPT (AI) for calculations in mechanical engineering. In I Karabegovic, A Kovačević and S Mandzuka (eds), *New Technologies, Development and Application, Vi*. Springer, Cham, Switzerland. 313–320. https://doi.org/10.1007/978-3-031-31066-9_34.

Ulrich KT and Eppinger SD (2012) *Product Design and Development*. Irwin: McGraw-Hill.

Vestad H, **Kriesi C**, **Slåttsveen K and Steinert M** (2019) Observations on the effects of skill transfer through experience sharing and in-person communication. *Proceedings of the Design Society: International Conference on Engineering Design* **1**(1), 199–208. https://doi.org/10.1017/dsi.2019.23.

Viswanathan VK and Linsey JS (2013) Role of sunk cost in engineering idea generation: An experimental investigation. *Journal of Mechanical Design* **135**(121002). https://doi.org/10.1115/1.4025290.

Wall MB, **Ulrich KT and Flowers WC** (1992) Evaluating prototyping technologies for product design. *Research in Engineering Design* **3**(3), 163–177. https://doi.org/10.1007/BF01580518.

Wang X, **Anwer N**, **Dai Y and Liu A** (2023) ChatGPT for design, manufacturing, and education. *Procedia CIRP* **119**, 7–14. https://doi.org/10.1016/j.procir.2023.04.001.

Xu W (2019) Toward human-centered AI: A perspective from human-computer interaction. *Interactions* **26**, 42–46. https://doi.org/10.1145/3328485.

Xu S, **Wei Y**, **Zheng P**, **Zhang J and Yu C** (2024a) LLM enabled generative collaborative design in a mixed reality environment. *Journal of Manufacturing Systems* **74**, 703–715. https://doi.org/10.1016/j.jmsy.2024.04.030.

Xu Z, **Hong CS**, **Zurita NFS**, **Gyory JT**, **Stump G**, **Nolte H**, **Cagan J and McComb C** (2024b) Adaptation through communication: Assessing human–artificial intelligence partnership for the design of complex engineering systems. *Journal of Mechanical Design* **146**(081401). https://doi.org/10.1115/1.4064490.

https://doi.org/10.1017/S0890060425000010 Published online by Cambridge University Press