How Do Hackathons Foster Creativity? Towards Automated Evaluation of Creativity at Scale

Jeanette Falk*
Computer Science
Aalborg University
Copenhagen, Denmark
jfo@cs.aau.dk

Yiyi Chen*
Department of Computer Science
Aalborg University
Copenhagen, Denmark
yiyic@cs.aau.dk

Janet Rafner
Center for Hybrid Intelligence,
Department of Management
Aarhus University
Aarhus, Denmark
Aarhus Institute for Advanced Studies
Aarhus University
Aarhus, Denmark
janetrafner@mgmt.au.dk

Mike Zhang
Computer Science
Aalborg University
Copenhagen, Denmark
jjz@cs.aau.dk

Johannes Bjerva
Department of Computer Science
Aalborg University
Copenhagen, Denmark
jbjerva@cs.aau.dk

Alexander Nolte
Software Engineering and Technology
Eindhoven University of Technology
Eindhoven, Netherlands
a.u.nolte@tue.nl

Abstract

Hackathons have become popular collaborative events for accelerating the development of creative ideas and prototypes. There are several case studies showcasing creative outcomes across domains such as industry, education, and research. However, there are no large-scale studies on creativity in hackathons which can advance theory on how hackathon formats lead to creative outcomes. We conducted a computational analysis of 193,353 hackathon projects. By operationalizing creativity through usefulness and novelty, we refined our dataset to 10,363 projects, allowing us to analyze how participant characteristics, collaboration patterns, and hackathon setups influence the development of creative projects. The contribution of our paper is twofold: We identified means for organizers to foster creativity in hackathons. We also explore the use of large language models (LLMs) to augment the evaluation of creative outcomes and discuss challenges and opportunities of doing this, which has implications for creativity research at large.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; **Collaborative content creation**; • **Computing methodologies** → *Machine learning*.

Keywords

Hackathons, creativity, human-centered AI, large language models, quantitative methods

*Both authors contributed equally to this paper

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713447

ACM Reference Format:
Jeanette Falk, Yiyi Chen, Janet Rafner, Mike Zhang, Johannes Bjerva, and Alexander Nolte. 2025. How Do Hackathons Foster Creativity? Towards Automated Evaluation of Creativity at Scale. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3706598.3713447

1 Introduction

Reacting *fast* to rapid changes in society and technology development has generally been considered a prerequisite for innovation [79]. Originating in niche software development environments in Silicon Valley during the late 1990's, *hackathons* have become a popular format for accelerating people's creativity and developing creative ideas, prototypes, products, or services in a variety of contexts such as education [82], entrepreneurship [41], corporations [71], scientific communities [38], civic engagement [81] and others [24]. Hackathons are time-bounded, participant-driven design events, often spanning only a few days, where participants form teams and collaborate on projects to address a common theme or challenge, resulting in a perceptible outcome, such as an interactive prototype [23]. They are often celebrated for their potential for creative outcomes, see e.g. [5, 11, 27, 45, 88]. For this reason, hackathons have become an attractive approach for entrepreneurs, companies, educators and researchers to develop creative solutions to problems they face or to complement or enhance their existing innovation processes [24].

Since 2016, the CHI community has shown a growing interest in hackathons. A search for "hackathon" in the CHI proceedings on the ACM Digital Library returns 87 results for full research papers. Most research focusing on creativity in hackathons, however, consists of case studies of one or only a few events [11, 27, 28, 80]. While they can provide valuable insights, case studies are not suitable to advance and validate theory, either on hackathons as a phenomenon in itself or on *how* hackathon formats lead to creative outcomes. Answering fundamental research questions in this frame requires larger-scale quantitative studies in order to test hypotheses

[21, 23]. Current quantitative studies of hackathons focus mainly on understanding project continuation [63, 65] and re-use of code that was created during hackathons [39, 40, 59]. To the best of our knowledge, as of yet, no research has conducted large-scale analysis of hackathon projects from a creativity perspective. Our work addresses this gap by asking the following first research question:

RQ1: How can we define and implement an analysis of creativity in a way that enables large-scale analysis of creativity within hackathon projects?

Large-scale creativity assessment is essential for validating the effectiveness of the different contexts in which hackathons are organized for supporting creative outcomes – such as in educational programs, workplace training, and other interventions – ensuring that the formats are successfully cultivating the creative skills demanded [74]. From this perspective, Rafner and colleagues have suggested that exploring methods which combine scalability "and product-oriented assessments (data with high ecological validity) could greatly enhance the construct validity of creativity assessment instruments" [74]. In our paper, we contribute to exploring such scalable methods with potentially high ecological validity for assessing creative products in the context of hackathons, since the data we analyze resemble real-world data rather than data created in controlled creativity test settings [73]. We thus also address the following second research question:

RQ2: What insights into participants, collaboration patterns, and hackathon setups can we gain from analyzing a large number of creative hackathon projects and how do these insights relate to fostering creativity?

Similar to the contributions by [39, 40, 59, 65], we use Devpost's database of hackathons and hackathon projects for our large-scale analysis of creativity.[1] Similar to Fang, Herbsleb and Vasilescu [25], we operationalize a definition of creativity which we explore as a lens to analyze our dataset. Specifically, we operationalize the concepts of *novelty* and *usefulness* to create a subset of creative projects. We conduct statistical analysis on this subset of creative projects to identify patterns in the interaction of aptitude, process and environment, a contribution which we frame as particularly valuable for researchers and practitioners who organize hackathons and wish to support hackathon participants' creativity and increase the potential for creative outcomes.

Furthermore, to address calls for exploring a Generative AI-augmented expert assessment [10, 58, 73], we explore the compelling case of Large Language Models (LLMs)-as-a-judge to augment our large-scale creativity evaluation method and address the following research question:

RQ3:How might LLMs be used to augment large-scale evaluations of creative hackathon projects?

The contribution of this two-pronged large-scale exploration of creativity in hackathon projects is the following: In addressing RQ1, we arrive at an operationalization of creativity theory which guides the first prong of our two-pronged approach; a statistical analysis of a dataset consisting of 10,363 hackathon projects. We discuss the findings from this statistical analysis as take-aways for organizers, thereby addressing RQ2. As the second prong of the two-pronged approach, we address RQ3 by using LLMs to analyze the creativity of a subset consisting of 21,318 randomly sampled hackathon projects which contained both creative and non-creative projects, as defined by our operationalization (RQ1). We also compared the ratings of LLM judges and human judges on a randomly selected subset of 30 hackathon projects. Drawing from the LLM analysis, we explore how LLMs can function as supplementary judges, supporting human judgment within a human-AI hybrid intelligent system [84], and how they can facilitate large-scale analyses of hackathon creativity. The large-scale exploration of creativity in hackathons contributes towards a greater understanding of how creativity manifests in hackathons, which we argue contributes to moving large-scale creativity evaluation towards real-world relevant scenarios with high validity. Finally, we discuss the challenges, limitations, and opportunities of our exploratory methods of measuring creativity on a large scale which contribute towards aiding researchers interested in quantitative instruments for measuring creativity in general.

2 Related Work

In this section we outline related works researching creativity in hackathons and how creativity can be evaluated.

2.1 Creativity in Hackathons

Hackathons are often highlighted for accelerating participants' creativity and resulting in creative outcomes [11]. Employee-focused internal hackathons have been highlighted as a means for testing "new products and services as well as to generate new ideas" [80] and streamline creativity by bypassing slow decision-making processes and fixed organizational structures [27]. Not surprisingly, creative industries such as media, arts, and culture fields have also adopted and adapted hackathons [46]. Karlsen and Løvlie explains this with these industries' "long tradition of using constraints and games to facilitate creativity in arts" [46].

Previous work has explored how to organize hackathons in order to facilitate creativity. Tenório and colleagues provide suggestions which mainly focus on pre-events such as "workshops, coaching, training, mentoring, and so on" [90] for: (1) supporting *knowledge application*, where the participants apply and combine their individual domain knowledge [67] into new perspectives and solutions for challenges; (2) learning about how to *manage conflicts*; (3) supporting participants' *individual learning*; (4) organizing shorter events to mitigate fatigue [90]. However, these recommendations are based on a small sample, making it difficult to assess their validity and understand to what extent they might hold in different contexts. Lobbe, Bazzaro, and Sagot researched collaborative design tools used to enhance creativity and innovation in a hackathon attended by 1,310 engineering students, resulting in around 160 projects [45]. This larger-scale study focuses on a narrow perspective on how to support creativity contributing mainly with insights on some select tools. In addition, the study focused on a single event with a specific setup and population of participants. Broadening the scope somewhat further in terms of number of hackathons studied, Attalah, Nylund and Brem studied three different hackathons through participant observation to understand "the impact of open innovation and collective intelligence in hackathons" [5], where

[1]https://devpost.com/

they frame the collective creativity taking place in hackathons as playing a key role for fostering collective intelligence. Collective intelligence is a kind of distributed intelligence constantly developed and coordinated by a group of people, for example, a community or society, while collective creativity is a "momentary, collective process that includes interaction in the form of help seeking, help giving, reflective reframing and reinforcing" [34] as cited in [5]. A key finding is that organizers should frame hackathons as a lasting development of collective intelligence, rather than momentary events of collective creativity [5].

These contributions are valuable in their own right to understand creativity in the context of hackathons; however, these cases take place in specific settings, which means that the findings may not be translated to other settings, as hackathons are very diverse [22]. In order to advance and validate theories on creativity in hackathons, we need quantitative methods [21] and large-scale analyses [74] in addition to case studies.

Moving from small-scale and qualitative data collections to larger ones, a couple of literature reviews have aimed at consolidating knowledge on how to organize hackathons including how to support creativity. Kollwitz and Dinter's contribution is a taxonomy of hackathons where they "figured out which dimensions and objectives are discussed in the literature" [51] and which "contributes to a better understanding of the opportunities and characteristics of hackathons" [51]. While they mention innovation and creativity as characteristic of hackathons, they mainly discuss their taxonomy as a way to reduce uncertainties for organizers regarding results, processes and resources, for example by making "detailed specifications regarding the solution space as well as the degree of elaboration in order to channel the creativity of the participants in a desired direction" [51]. Furthermore, their taxonomy "need[s] further evidence to show that those aspects are actually relevant from a practical point of view" [51].

Heller and colleagues conducted a literature review of 87 articles on how to best execute hackathons including elements which are meant to facilitate creativity [36]. The following elements were specifically highlighted for their relation to increased creativity: Pre-registered teams may improve teamwork but may jeopardize creativity as they are usually made up of participants with similar backgrounds [66] whereas diverse teams have shown increased creativity [87]. Heller and colleagues further found that "a very competitive atmosphere with high-value prizes creates extrinsic motivation to stand out, but an atmosphere of collaboration leads to improved creativity, better teamwork, and intrinsic motivation" [36]. Drawing especially on the findings by Lifshitz-Assaf and colleagues, Heller and colleagues emphasized that although research has discussed the detrimental effect of time pressure on creativity (see e.g. [4]), the limited time-frame of hackathons are by design, and instead of following traditional coordination strategies participants should adopt *adaptive coordination processes* [55], similar to Edmondson's distinction between teamwork and *teaming* [20]. For supporting ideation, Heller and colleagues identified having mentors as ideation facilitators for increasing creativity and referred to Wilson's description of good brainstorming facilitators [96]: "preventing participants from offering premature criticism, encouraging the flow of ideas, focusing on quantity rather than the quality of ideas, and promoting tolerance for radical ideas" [36].

However, similar to related works involving case studies of hackathons, Heller and colleagues' literature review included "existing studies [that] are limited to mainly correlative case studies, [and] which do not allow for a proper understanding of the causal processes underlying effective hackathon execution" [36]. They further recommend future research exploring hackathons to conduct "advanced qualitative (such as collecting previous hackathons' post-event surveys and interviews) and quantitative methods (such as meta-analysis) to offer data-driven conclusions on how to best plan and execute hackathons" [36]. Our contribution falls into the latter type of research. Heller and colleagues also suggest that "studies can examine participant-, organization-, or event-level variables of interest by comparing the outcomes of two hackathons that are identical in all characteristics except one (e.g., staff diversity, judges' identity, virtual versus physical versus hybrid, etc.)" [36]. This is, to some extent, what we contribute in this paper: However, instead of comparing only two hackathons where all except one variable are controlled, we utilized a large-scale dataset consisting of almost 200,000 hackathon projects.

2.2 Evaluating Creativity

The need for accurate and valid methods to evaluate creativity was first emphasized during Guilford's 1950 presidential address to the American Psychological Association (APA) and has been a popular topic of study in many disciplines, including the CHI community [93]. More recently, the 2019 special issue of Philosophy, Aesthetics, Creativity and the Arts, featured an editorial statement stressing that "without proper instruments to measure creativity or adequate standards of assessment, the validity of any creativity study is seriously questioned" [7]. However, developing and adhering to rigorous creativity assessments standards proves exceptionally difficult, given the complex, contextual nature of creativity and the many ways in which it can be evaluated.

In psychology, creativity assessments include standardized instruments targeting components of divergent thinking (e.g., originality, flexibility, fluency), as well as expert product evaluations and measures of creative behavior or self-efficacy [74]. One of the major drawbacks of many creativity assessments is that they require extensive human effort for manual evaluation which is labor intensive and often suffers from disagreement among raters [74]. Some automated creativity scoring leverages advances in Natural Language Processing (NLP) to address these limitations [6, 26, 75].

LLM-as-a-judge. Recent NLP research trends highlight a shift towards using LLMs for automated evaluation of text documents, potentially supplementing or replacing human judgments [16, 56, 94, 97]. This approach manifests in competition leaderboards such as [97] or [53], where researchers evaluate models against each other based on human evaluations or pre-defined benchmarks. In general, rating long and detailed survey responses based on pre-designed criteria is challenging even to human experts. To reduce manual effort, researchers can instruct an LLM to rate any given text on, e.g., a Likert-like scale (1-7) or other custom rubrics as, e.g., proposed by Kim and colleagues [49]. This "LLM-as-a-judge" approach has demonstrated strong correlations with human judgments on various natural language tasks [15, 29, 37, 50, 56, 64, 91, 97] and an even better agreement when used in a "jury setting" with multiple models judging the same input [92] and aggregating the ratings.

Automated methods for creativity assessment have mostly been applied to scoring divergent thinking tasks, such as the Alternative Uses Task[33, 68]. These automated methods have focused exclusively on scoring the originality of ideas [1, 8, 9, 35, 70] . However, assessment of solutions from real-world creative problem-solving, often requires both evaluating novelty and usefulness (i.e., plausible and effective) [75].

Luchini and colleagues [58] have taken an important step towards addressing this issue by using LLMs to automatically evaluating standard creative problem solving tasks for both novelty and usefulness. While more complex than data from the Alternative Uses Task, their creative problem solving tasks still result in well structured, reasonably short, homogeneous data. There is a need for methods to handle the variability of real-world data, ensuring LLM evaluations of creativity remain as effective as in controlled settings.

For complex and context-rich datasets such as hackathon outputs, we propose that instead of replacing subjective human evaluations, LLM technologies should assist human judges by providing scalability, analytical insights, and potentially consistency, to enhance human expertise. This approach aligns with recent advancements in hybrid intelligence and human-centered AI, which emphasize synergetic human-AI interactions [60, 62, 73, 84], fostering collaborative partnerships where AI serves as a supportive tool for humans [54, 85].

Our contribution approaches large-scale analysis of our context-rich hackathon project dataset from two angles: First, we explore how to operationalize creativity in order to conduct statistical analysis, and secondly, we explore how LLMs may be used as an additional supporting judge.

3 Method

Similar to our motivation for exploring large-scale evaluation of creativity, Fang, Herbsleb and Vasilescu explored creativity and innovation by analyzing 70,891 projects from the World of Code dataset. They follow the Schumpeterian tradition [83] of "viewing innovation as emerging from the novel recombination of existing bits of knowledge" [25] and analyze their data by operationalizing this at code level, by focusing on unusual combinations of software packages. From this lens, they found that innovative projects tended to have more GitHub star counts – in other words, novelty begets popularity. We expand on this approach by (1) considering the established standard definition of creativity as defined by Plucker, Beghetto and Dow [44], and (2) evaluate this in the context of a new dataset based on a large sample of hackathon projects (further detailed in Section 3.2).

3.1 Creativity Framework

In order to answer the three research questions stated in the introduction, we constructed a framework to guide our analysis and frame the results. This was done iteratively, by continuously exploring our dataset, familiarizing ourselves with it, reviewing related creativity literature, and discussing among ourselves about how creativity is measured in general and how we might analyze hackathon creativity based on our available data. During these discussions, we constructed the framework which was used to scaffold the final analysis of the dataset. To create the framework for our analysis, we took point of departure in Plucker, Beghetto and Dow's standard definition of creativity, which is based on a cross-disciplinary review of how peer-reviewed business, education, psychology, and creativity journal articles evaluated the term creativity:

> "Creativity is the **interaction** among **aptitude**, **process**, and **environment** by which an individual or group produces a **perceptible product** that is both **novel** and **useful** as defined within a social context" [44]

3.1.1 Operationalizing the Standard Definition of Creativity. We approach the definition in a reverse order: First, we defined how *novelty* and *usefulness* may be expressed within the context of the dataset consisting of hackathon projects. While there have been many definitions of creativity in addition to Plucker, Beghetto and Dow's [44], this twin criteria of novelty and usefulness have formed "principal components of numerous definitions of creativity dating back at least 70 years" [42]. This means that for a project to be creative it should be both novel *and* useful, and we thereby expand on the approach taken by Fang, Herbsleb and Vasilescu [25] who only focused on novelty. The precise terminology of novelty and usefulness can vary [42], and in our definitions of them used for the LLM-as-a-judge approach, we aimed at including oft-repeated elements from the creativity research literature.

Novelty. Related to our exploration of LLM-as-a-judge, we took inspiration from Luchini and colleagues' prompt design, where they explored automatic scoring of Creative Problem-Solving with LLMs. In their prompt, novelty refers "to how unique the approach is compared to typical solutions" [58]. This resembles the strategy taken by Fang, Herbsleb and Vasilescu [25]. Hence, we adopted a definition of novelty, which entail elements of uniqueness and originality [42], see Appendix A for the full prompt:

> *How unique and original is the project's concept, approach, or solution? Does it introduce new ideas, methods, or perspectives that are significantly different from existing ones?*

To operationalize this definition for the analysis, we decided to replicate the approach by Fang, Herbsleb and Vasilescu who defined this "as a function of the libraries and packages a project imports [i.e.] projects built on top of more atypical combinations of libraries are considered to be more innovative" [25]. We replicated this approach and identified the hackathon projects which included unusual combinations of packages and libraries. We included the five most popular programming languages used in hackathon projects in our dataset: Python, Node.js, Java, C#, and PHP. While this may, indeed, exclude some potentially creative projects, we decided for this approach partly to repeat the research by Fang, Herbsleb and Vasilescu [25] but also because we hypothesized that this approach includes projects that have a higher chance of getting utilized or continued in the future because the languages they were developed in are so wider spread.

Usefulness. Luchini and colleagues do not use the term "usefulness" in their prompt design [58], however, as there is a lack of established terminology in this area, concepts used in definitions of creativity may therefore vary in their exact wording. Hence, Luchini and colleagues include the terms "completeness" and "effectiveness", which closely resembles usefulness: "Completeness refers to how well the solution addresses multiple issues raised by the problem. Effectiveness examines whether the solution is viable, feasible, practical, or appropriate" [58]. Our adopted definition of usefulness is the following (see also Appendix A):

> *How practical and appropriate is the project in addressing the problem, situation, or challenge its targets? Does it effectively solve a real-world issue or meet a specific need?*

We first needed to distinguish a subset of the dataset which can be identified as containing creative projects. Since it would not be feasible to conduct expert evaluation on the full dataset, we opted for a creativity proxy, by considering winning hackathon projects, which include a *winner*-tag in our dataset. This, of course, excludes hackathon projects which did not win a certain hackathon according to a set of criteria, but for our exploration of how to evaluate creativity at scale we find this a sufficient proxy for the expert evaluation. While we do not know at scale which criteria define winner-projects, considering the wide-spread emphasis on creativity in hackathons as mentioned in the introduction, we hypothesize that the majority of our dataset's winner-projects have been considered creative or have addressed a challenge in a sufficient way which has been perceived as useful.

However, what is considered useful within a hackathon context may not necessarily be considered useful outside of the hackathon context. We therefore hypothesized that how *accessible* a hackathon project is for outsiders can be a proxy for considering usefulness outside of a hackathon context. Operationalizing this hypothesis, we decided to follow an approach inspired by Imam and colleagues who explored antecedents of reuse of code that was created during hackathons [40]. If code is reused this points towards the code being useful outside of the confines of a hackathon. Imam et al. found that the availability of data and documentation in addition to code and the presence of an open-source license significantly contributed to a project being reused. Furthermore, larger data files may indicate that projects are structured with reusability in mind. Following their findings, we include hackathon projects which have GitHub repositories and include data files or folders and an Open Source license for the repository in our final subset of creative perceptible hackathon projects.

Hence, the subset of creative hackathon projects satisfies the above-mentioned criteria of novelty (including atypical combinations of software) *and* usefulness (containing a winner-tag *and* is accessible for outsiders). We discuss the limitations of this delimitation in the Discussion section. Using our operationalized definitions of novelty and usefulness, we identified **619 creative perceptible products in the hackathon context**, see fig. 1. The technical details of this process are described in section 3.2. The next subsection describes some guiding questions for the different angles which we analyzed the dataset from.

The Interaction Between Aptitude, Process and Environment: We finally explored the three **aspects of creativity** in terms of aptitude, process and environment in this subset of the creative perceptible hackathon products. However, we found that the dataset contained rich data which went beyond these three aspects, therefore we extended the framework to better suit our analysis. Some frameworks have been suggested to distinguish between different aspects of creativity such as: The 4P model [78] (People, Process, Product, Press), the 5A model [30] (Actor, Action, Artifact, Audience, Affordances) and the 7C model (Creators, Creating, Collaborations, Contexts, Creations, Consumption, Curricula) [57]. We turned to the framework of the 7 C's of creativity because it has the added aspect of Collaboration, which is an important characteristic part of hackathons [23]. We excluded Creating and Curricula from our analysis for the following reasons: Creating would require more in-situ data which documents the creative design process in terms of, for example, the moment to moment decisions. The dataset does not contain this kind of data. Similarly, Curricula revolves around the overarching context of how creativity is taught and learned in the specific context. This leaves us with the aspects of **Creators, Collaborations, Contexts, Creations**, and **Consumption**.

With this framework, we were able to approach the computational analysis of creativity in hackathons in a structured way and base the analysis on established definitions from creativity research. Turning back to the standard definition of creativity, the findings in section 4 then relate to the **interaction** between the selected aspects of the 7 C model [57]. In the following subsections, we outline research questions for each of the five selected C's.

Creators: Lubart describes creators as referring to "those who engage in the production of original, meaningful content" [57]. To explore the dataset, we asked the following questions which could provide insights about the creators of the creative hackathon projects:

(1) Is the creator a hackathon winner?
(2) Which skill sets have they provided on their profile?
(3) What is the experience of creators? How many hackathons have they participated in? How many years have they participated in hackathons?

Collaborations "is the term used to signify the involvement of significant others in the creative process. This may be an individual creator, such as a writer, who interacts with another person, such as his or her literary agent or critic, it may be a dyad of creators who work together, or a team of people who work on a project [...] The collaboration, in terms of interaction patterns, the nature and complementarity of the collaborators (team diversity) are some specific examples of topics that concern this 'C'." [57] From the dataset, we wanted to explore:

(1) Do creators work together in different hackathons?
(2) What are typical team sizes?
(3) What skills and interests do teams possess?

Contexts "refers to the physical and social world in which creators engage in the creative process [...] The environment provides resources and constraints, it orients behavior. It affords certain actions more than others, facilitating or hindering creative behavior but also providing the field within which new productions will be situated and evaluated" [57]. To explore the hackathon contexts, we looked for the following elements:

(1) Which hackathon themes do creative projects address?
(2) What is the distribution of themes?

(3) What is the level of competition in hackathons, i.e., what is the interaction between the number of projects of hackathon and the number of winners in a hackathon (is there only one first prize or multiple prizes)?
(4) How big are hackathons with creative projects?
(5) Are the hackathons mostly onsite or online?
(6) How many creators participate?

Creations are "the production resulting from the creative process, maybe a tangible or intangible output. It may be a relatively unformed idea, or a full-fledged 'product'. The characteristics of the production, such as its originality compared to previous works, and its 'usefulness', maybe some criteria that the creator and external judges take into account" [57]. To explore the creations, we wanted to analyze the following components:

(1) How does the hackathon projects compare their inspiration source with what they do? What is the *semantic distance* within project descriptions, i.e. "an aspect of originality [or novelty] reflecting remote relationships between concepts" [58].
(2) How elaborate are the project descriptions?
(3) How much do project descriptions align with or differ from the hackathon context, i.e. the posed challenge the project was meant to address?

Consumption "refers to the adoption of creative ideas and productions. Those who encounter a creative product may adopt it more or less quickly, with more or less enthusiasm. Creations are situated within a context of the marketplace of existing ideas, products, or previously known solutions" [57]. As we described above, a hackathon project may be perceived as creative in one way during the hackathon and in another outside the hackathon context. To explore whether creative hackathon projects are adopted outside the hackathon context, we explore the following:

(1) Do the projects have GitHub repositories, and are these reachable and maintained?
(2) What is the life span of a project outside of the hackathon?

While the findings in section 4 do not reflect all these research questions for the C's, they were used to structure the initial analyses of the large-scale dataset and the results of these initial explorations can be found in the appendix, see section B. Researchers who are interested in similar evaluation methods for creativity, may find the tables interesting to see which parameters may not contribute to creative hackathon projects.

3.2 Experimental Setup

In this section, we first describe the series of steps in our pipeline for data collection and pre-processing, before analyzing this data with regards to our operationalizations of usefulness and novelty, and finally predict the creativity of projects with a mixed random effects model. As each stage in the data pre-processing process, relating to the five C's, requires adding incremental filtering criteria, the number of projects that we analyze varies as described in Fig. 1.

3.2.1 Data Collection and Pre-processing. We collected publicly available information on hackathons, projects and participants from the hackathon database Devpost which represents a snapshot of the state of hackathons, projects and participants at the moment

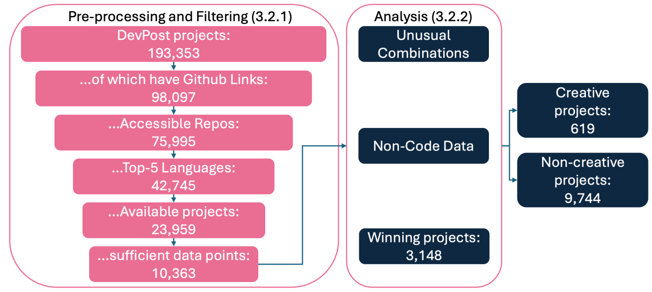


Figure 1: Number of projects remaining after each pre-processing step (left). This process is described in detail in section 3.2.1; Extracted variables for analysis (mid). The analysis is described in detail in section 3.2.2; Resulting prediction of creative projects (right). The analysis results are described in detail in section 3.2.3.

in time when we collected the data. Historical data was not available to us. We begin by curating and processing a dataset with 193,353 projects (ca. 6k hackathons and 314k participants). After preprocessing, we narrowed it down to 23,959 projects with GitHub repositories. Further filtering for repositories with code scripts importing more than two packages and non-code data yielded 10,363 projects for usefulness and novelty analysis, in which 3,148 projects are identified as winning projects. Finally, the projects are identified into 619 identified as creative – which fulfills the categories "Unusual Combinations", "Non-Code Data" and were winning projects – and 9,744 as non-creative (Fig. 1).

As shown in Fig. 2, each project can be submitted to multiple hackathons, and conversely, a single hackathon can host multiple projects. To investigate collaboration patterns in hackathons, we process the dataset to capture the dynamics of collaboration among creators at the project level, considering both their experience in hackathon participation and a variety of individual interests and skill sets. In the following, we explain the pre-processing steps in detail for exploring each aspect of the five C's.

Creators. We crawled data for $310,900$ creators which have participated in the hackathon projects, observing variables including interests, skills, received likes, had followers and participated in 1.5 hackathons and projects on average. As very few participants have followers or likes, we binarize these numerical data points into *have followers* and *have likes* for further analysis. To explore the correlation between creators' individual traits and their winning hackathons, a new variable is calculated as follows:

$$\text{AVG. Weighted Winning} = \frac{1}{n}\sum_{i=1}^{n}\frac{\text{Places}_i}{\text{Participants}_j * \text{Sub. Projects}_i}, \quad (1)$$

where n is the number of hackathons that the creator won, Places_i is the number of places reserved for winners in a hackathon i, and Participants_j is the number of participants who are creators in one project team j, and Sub. Projects_i is the total number of submitted projects in the regarding hackathon i. Table 2 contains the correlations among variables for participants.

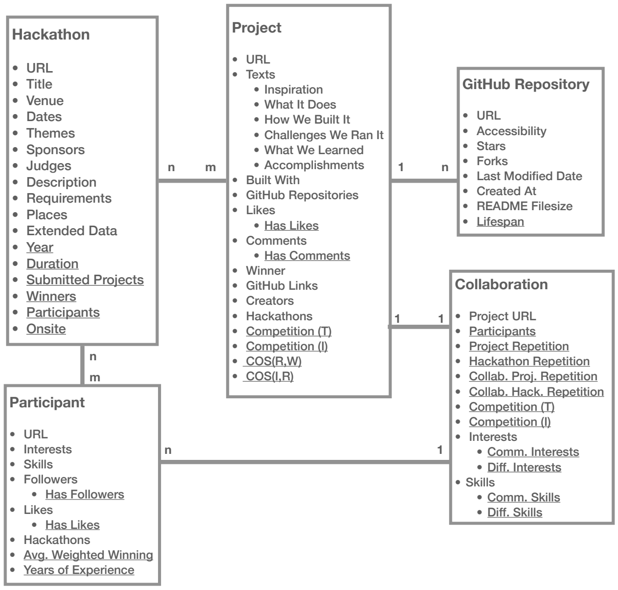


Figure 2: Variables in our dataset. The underlined Variables are processed from the collected data.

Collaborations. Collaboration among creators is a dynamic temporal process. The collaboration among creators in hackathons accumulates through the repeated participation of the same pair of participants, so does the collaboration on the projects. Moreover, individual experience accumulates through repeated participation in projects and hackathons with others. To capture the dynamics of collaboration for each year, we devise an algorithm 1. It initializes dictionaries to track creator-to-creator interactions for projects and hackathons, as well as individual hackathon and project repetition. Then, it iterates through projects by year, updating the interaction counts for each pair of creators who collaborate on a project or participate in the same hackathon. Finally, it computes the interaction metrics for each creator at both the project and the hackathon level. The results are then aggregated over time, recorded as *Collaboration Project Repetition*, *Collaboration Hackathon Repetition*, *Project Repetition* and *Hackathon Repetition*. Furthermore, we extract the intersected, different, and all sets of *interests* and *skills* among participants for each project as variables to explore the dynamics of collaboration. As shown in Table 3, the variety of skills is positively correlated with repeated collaborations for projects and hackathons.

Contexts. A total of 5,458 hackathon URLs remain accessible, and information has been extracted from them for the period 2009-2024 in the dataset. Using the crawled data, we extract the *year* and *duration* by days of the hackathons by parsing the schedule dates using Python package *dateutil*.[2] The number of submitted projects, the number of actual winners and participants for each hackathon are aggregated utilizing project data. Hackathons occur worldwide, and we are particularly interested in whether the format of the hackathons—whether online or onsite—affects creativity. To analyze this, we convert the *venue* variable into a binary variable, *onsite*. Furthermore, to measure the level of competition and explore its relevance in the context of hackathons, we calculate the ratio of reserved winning spots to the number of participating teams, as well as to the number of participants. This provides *competition* metrics at both the team level (T) and the individual level (I). Refer to Table 4 in Appendix B for correlations among variables in hackathons.

Creations. Projects are creations in the context of hackathons. There are originally 193,353 project data in total, with some projects submitted to multiple hackathons. And 187,425 of which contain information descriptions. To explore how hackathon project descriptions, their inspiration and the overarching hackathon description compare, we calculate the semantic distances by leveraging the text descriptions from fields such as *What It Does*, *Requirements*, and *Inspiration*. To achieve this, we use a sentence transformer LaBSE[3] to encode texts into 768-dimensional representation vectors, with maximal sequence length 256. We then calculate cosine similarities between (1) the embeddings of *Inspiration* and *What It Does* and (2) the embeddings of *What It Does* and *Requirements*, recorded as $COS_{R,W}$ and $COS_{I,W}$, respectively.

Consumption. As shown in Fig. 1, among the 193,353 projects we are investigating, 98,097 of them have GitHub links, among which 75,995 GitHub repositories are still accessible, 42,745 of which are written in the top 5 programming languages, and eventually, we were able to crawl 23,959 GitHub repositories with non-empty scripts, 4,037 of which has a specific license, the top three of which are MIT, Apache and GNU licenses (Fig. 7). Moreover, 22,718 GitHub repositories contain non-code data (Fig. 8). As consumption overlaps with the number of retrievable repositories, this coincides with the number of projects we use for the remainder of our analysis.

The Pearson correlation coefficient between the winning project and GitHub accessibility is 0.06 with $p < 0.001$.

3.2.2 Analysis of Usefulness and Novelty. To assess both usefulness and novelty, we adopt a two-pronged approach. Usefulness is evaluated by focusing on hackathon projects that include non-code data and those recognized as winners, reflecting their practical application and broader impact beyond hackathons, extending Fang, Herbsleb and Vasilescu's work [25]. Novelty, on the other hand, is measured by unusual combinations of imported packages within the project code, highlighting innovative approaches to problem-solving.

Usefulness: Non-code data and Winners Continuing the dataset of 23,959 Github repositories in total, with Python, Java, Node.js, C# and PHP in the `built with` tag from the project. In detail, we extract information such as *data size (in Bytes)*, *License*, *GitHub Lifespan (Days)*, *# Data Files*, *# Data Folders*, *Watches* and *Stars*.

Novelty: Unusual combinations In addition, to operationalize novelty of the project, we investigate the unusual combination of the imported packages in the code. To achieve that, we process the same set of GitHub repositories with the following steps: (1) the utilized packages are extracted from each script; (2) compute the number of imported packages for the same project; (3) filter the repositories out when their imported packages are fewer than 2. As

[2]https://pypi.org/project/python-dateutil/

[3]huggingface: sentence-transformers/LaBSE

a result, we have 10,363 GitHub repositories for analyzing novelty leveraging the atypical combinations of imported Python packages.

To conduct our experiment, first, we simulate the actual package imports over time. We represent package imports by projects over time as a matrix where each row is a project, and each column is a package. In this matrix, a value of 1 indicates that a project imports a particular package in that year, while 0 signifies no import. Then for each year, we simulate random package combinations for projects. Each project imports the same number of packages as in the actual data, but the choice of which packages are imported is random. This gives us a counterfactual dataset to compare against. This uses the Monte Carlo simulation to capture the atypicality of package combinations. Then *empirical frequency* is computed, where how often two specific packages are used together in the actual data. At the same time, we compute *simulated frequency*, which calculates how often the two packages appear together by chance after several simulation runs. The atypicality score is derived from the ratio of empirical frequency to the average simulated frequency. The combination is atypical if the empirical frequency is much lower than expected. We compute a z-score for each package with the equation:

$$z_{ijt} = (obs_{ijt} - exp_{ijt})/\sigma_{ijt} \tag{2}$$

where obs_{ijt} represents the empirically observed frequency of packages i and j appearing in the same project in year t, exp_{ijt} is the average number of times that packages i and j appear in the same project in the year j over twenty simulated event sets, and σ_{ijt} is the standard deviation of the co-occurrence frequency of packages i and j in those sets as well. In the end, we use the smoothed version of z-score to measure atypicality, as follows:

$$Z_{ijt} = \begin{cases} \log(z_{ijt}+1) & \text{if } z_{ijt} \geq 0 \\ -\log(-z_{ijt}+1) & \text{if } z_{ijt} < 0 \end{cases} \tag{3}$$

A low Z-score indicates high atypicality and, consequently, novelty, whereas a high Z-score suggests the opposite.

3.2.3 Predicting Creative Projects using Mixed Random Effects Logistic Regression. To investigate novelty and usefulness as proxies for creativity of projects, we use a combination of Z-score and *winner*-tag as the dependent variable in a mixed-effects logistic regression model. This approach allows us to identify which variable contribute most to creative projects. The Z-score is a continuous variable, where a Z-score of zero indicates a high level of novelty, while *winner*-tag is binary. We integrate those two variables into a binary variable using the following equation:

$$\text{Creative-Projects} = \begin{cases} 1, & \text{if } \textit{winner}\text{-tag} = 1 \text{ and } Z\text{-score} = 0 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

To simulate the conditions in which a project qualifies as a creative project, characterized by the presence of a *winner* tag and a valid Z-score, we filter the data to include only hackathons that have at least one designated *winner*-tag. A project may have different outcomes across various hackathons (either winner or not), but it retains the same Z-score for its GitHub repository. In total, we have a dataset of 10,363 projects for further analysis, with 619 identified as creative and 9,744 projects identified as "un-creative". In the following we analyze what distinguishes the 619 creative projects from the rest.

Table 1: Model predicting Creative Projects. p<0.1 (.), p<0.05 (*), p<0.01 (**), p<0.001 (***).

(Intercept)	-3.250649 (***)	0.137272	-23.680	< 2e-16
Hackathon				
#Participants (in Hackathons)	-0.226608 (*)	0.113896	-1.990	0.04664
Onsite	0.067976	0.065501	1.038	0.29937
Duration	-0.078477	0.057713	-1.360	0.17390
Team				
Competition	0.424168 (***)	0.102305	4.146	3.38e-05
# Team Members	0.156271 (*)	0.073603	2.123	0.03374
Different Interests	-0.305535 (*)	0.131234	-2.328	0.01990
All Skills	0.168547	0.154694	1.090	0.27591
All Interests	0.121256	0.128110	0.946	0.34389
Hackathon Repetition	-0.046386	0.097278	-0.477	0.63347
Different Skills	-0.213552	0.173468	-1.231	0.21829
Project				
Has Likes	0.275278 (.)	0.152604	1.804	0.07125
GitHub Watches	1.584174 (*)	0.654169	2.422	0.01545
GitHub Data Size	0.097405 (**)	0.031475	3.095	0.00197
Project Repetition	0.021523	0.120655	0.178	0.85842
COS(I,W)	0.008229	0.056720	0.145	0.88464
Has Comments	0.023240	0.150865	0.154	0.87757
GitHub has License	0.050863	0.043701	1.164	0.24447
# GitHub Data Folders	-0.009275	0.123769	-0.075	0.94027
GitHub Lifespan (Days)	-0.077932	0.048979	-1.591	0.11158
COS(R,W)	-0.095202	0.060317	-1.578	0.11448
# GitHub Files	-0.643382	0.398770	-1.613	0.10665
GitHub Stars	-11.130124	7.239957	-1.537	0.12422
Individual				
Average Won Hackathons	0.918309 (***)	0.158897	5.779	7.50e-09
Hackathon Repetition	0.484675 (***)	0.058513	8.283	< 2e-16
Has Followers	0.246149 (***)	0.055379	4.445	8.80e-06
Average number of Hackathons	-0.769460 (***)	0.192835	-3.990	6.60e-05
Project Repetition	-0.247321 (*)	0.100642	-2.457	0.01399
Competition	-0.220472 (*)	0.105092	-2.098	0.03591
Has Likes	-0.059746	0.056079	-1.065	0.28670
	R^2_m	**0.967**	R^2_c	**0.970**

We process the variables from various aspects of the dataset (containing both creative and un-creative projects) - namely participants, projects, and hackathons - into project-level variables. For example, the number of participated hackathons and the years of experience for participants are averaged for the regarding project. To ensure the comparability among variables, numerical variables are scaled. We then apply a mixed effects logistic regression model using *glmer* function from R-package *lme4*. In this model, variables directly related to hackathons are treated as random effects, where other variables are fixed effects. The model uses *Binomial* family and is optimized with *nloptwarp* control optimizer. We examine the multi-collinearity of the variables in the models, and select the variables which are the most relevant and non-redundant. The results are presented in Table 1, and the heatmap of Spearman's rank correlation coefficients among the selected variables are reported in Fig. 9.

4 How are Hackathons Creative?

In this section, we focus on the findings which tell us something about the factors which seem to have a significant effect for whether a hackathon project is considered creative.

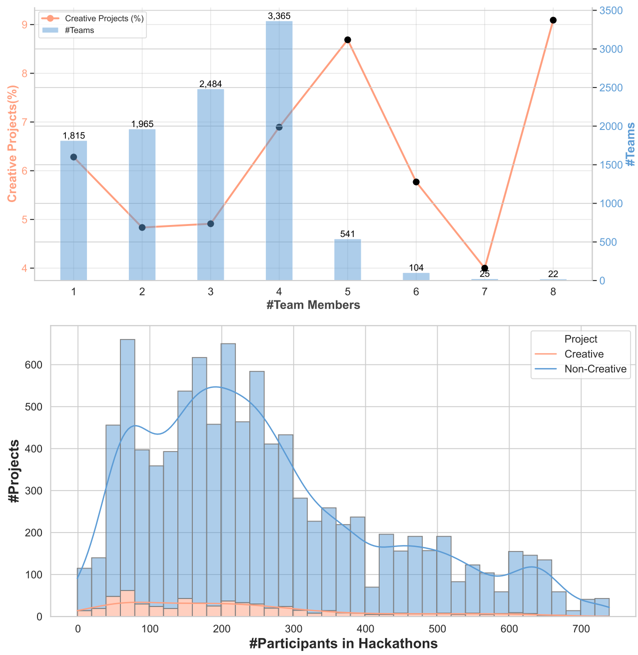


Figure 3: Top: The percentage of Creative Projects of the number of Teams by the number of Team Members. Bottom: The distribution of creative and non-creative projects by the number of participants in hackathons (outliers removed outside the 75% percentile).

We fitted a mixed-effects logistic regression model to the processed data (c.f. Section 3.2) in order to investigate the relationship between a project being considered creative – operationalized as it being among the winning projects of a hackathon and with unusual combinations of imported packages in GitHub repositories - and the aforementioned aspects that relate to individual, team, and hackathon characteristics. We model creativity as a dichotomous variable and consequently used logistic regression for our analysis. We modeled hackathon-related aspects as a random effect because projects that take place at the same hackathon are not independent. The results of the logistic regression are presented in Table 1. We only included variables in the model that were significantly associated with creative projects.

Our findings show effects related to individual, team, project, and event characteristics. Related to event characteristics we found that hackathon size is negatively associated with creativity ($\beta = -0.226608$, $p < 0.05$), suggesting that smaller hackathons tend to be more conducive to creative outcomes. Further analyzing this aspect we created a plot that shows creative projects compared to the size of a hackathon (Fig. 3, right). Due to the long-tail distribution of our dataset – the largest hackathon had 11553 participants – we only plotted the bottom 75% of events. This plot shows that most creative projects were conducted in hackathons that had around 60 to 80 participants.

Related to team characteristics we found that teams facing a lot of competition – as in many teams vying for few prizes – to be a strong predictor of creativity ($\beta = 0.424168$, $p < 0.001$) indicating that perceived competition might foster creativity. In addition, we also found team size as in the number of team members to be positively associated with creativity ($\beta = 0.156271$, $p < 0.05$). Further analyzing this aspect we again created a plot that shows creative projects compared to the number of team members (Fig. 3, left). The plot only shows teams from 1 to 8 members because none of the 42 projects that had more than 8 members in our dataset were recognized as being creative. The plot shows that teams with 4, 5, and 8 members – which could be an outlier due to the small number of teams that had 8 members – exhibit more creative projects.

At the same time, our analysis showed a negative association between the number of interests in a team and creativity ($\beta = -0.305535$, $p < 0.05$).

Summarizing these findings, it appears that larger teams with coherent viewpoints that have to compete for few prizes account for more creative projects in hackathons.

Related to the projects that teams worked on, we found that the number of likes that a project received ($\beta = 0.275278$, $p < 0.1$), the number of of GitHub Watches ($\beta = 1.584174$, $p < 0.05$) and the size of GitHub projects ($\beta = 0.097405$, $p < 0.01$) – as in file size – to be positively associated with creativity. This finding is difficult to interpret, though, since likes might have been used to determine a winning team and a project might have received likes and GitHub watches after a team won a prize. Moreover, the size of a project being positively associated with creativity could be attributed to us utilizing unusual combinations as an indicator for creativity with larger projects having more potential for such unusual combinations. It is, however, also possible that projects with more elaborate artifacts are considered to be more creative in the context of hackathons.

Most individual aspects we considered for our model were associated with creativity. The largest positive individual predictor was the average number of hackathons an individual has won ($\beta = 0.918309$, $p < 0.001$). This appears reasonable, since we utilized winning as one of the criteria to assess creativity and it can be expected that someone who has won an event would know what it takes to win again. Just utilizing the same project again, however, does not appear to be perceived as creative as shown by the negative association between project repetition and creativity ($\beta = -0.247321$, $p < 0.05$). Our findings also show a positive relationship between repeated hackathon participation with the same team and creativity ($\beta = 0.484675$, $p < 0.001$). Conversely, frequent individual participation across multiple hackathons is negatively associated with creativity ($\beta = -0.7695$, $p < 0.001$). These results highlight the value of shared experiences, even if participants do not work on the same project, and suggest that common exposure through repeated events may enhance collective ideation.

We also found a negative association between perceived individual competition and creativity ($\beta = -0.220472$, $p < 0.05$) which appears surprising since competition was the largest positive predictor for creativity on a team level. One possible explanation is that larger teams might aid the confidence of individuals thus making the competition less daunting. Finally, our findings also revealed that an individual having followers on Devpost is positively associated with creativity ($\beta = 0.246149$, $p < 0.001$). This – like the connection between a team receiving likes and creativity as discussed before – is difficult to interpret, though, since it is likely that an individual might have gained followers after they had won.

With respect to the goodness of fitting the model, the high R_m^2 of 0.967 suggests that a significant proportion of the variance is explained by the fixed effects alone. Meanwhile, the R_c^2 of 0.970 reflects the proportion of variance explained by both the fixed and random effects combined. Overall, the model demonstrates a strong fit to the data.

5 Exploring LLM-as-a-judge

We investigate LLM-based methods, inspired by [58] and others, to expand our exploration of creativity evaluation at scale. Specifically, we employ four LLMs to evaluate the usefulness and novelty of project descriptions. The models we use for judging the hackathon descriptions are Llama-3.1-8b-instruct [19], Mistral-7b-instruct-v0.3 [43], Gemma-2-9b-it [89], and Prometheus-7b-v2.0 [49]. We selected these models for two main reasons. First, they are all instruction-following models, meaning users can prompt them to perform specific tasks. Second, their smaller size fits our computational constraints, and avoids incurring an excessive environmental impact. There is no direct difference between the models, apart from the pre-training data or instruction tuning data that the organizations that trained the models do *not* release. Therefore, we also account for diversity of models. In the case of Prometheus-7b-v2.0, this model is specifically trained to score in rubrics and give feedback, which should be well-suited for our experiments. Figure 6 in Appendix A presents the detailed prompt we use. We base our descriptive prompt on [49]. Both [48] and [49] advocate for a detailed prompt that specifies what to evaluate in a text, rather than simply asking to "rate this text 1 to 5".

In Figure 4, we show a barplot of the novelty and usefulness scores from the four LLM-as-a-judges on a subsample of 21,318 hackathon descriptions (randomly from the total 193,353 descriptions) of both creative and non-creative projects. The figure reveals distinct patterns. For example, we notice that Gemma-2-9b-it consistently favors median scores, with peaks at score 3 for both novelty (±80%) and usefulness (±75%). Mistral-7b-instruct-v0.3 and Llama-3.1-8b-instruct show a tendency towards higher ratings, both peaking at score 4 (±65%) for both metrics. Prometheus-7b-v2.0 displays the most uniform distribution across scores 2-5. Usefulness scores generally exceed novelty scores, particularly for higher ratings.

This suggests, according to LLMs, hackathon participants often create practical solutions, even if not always novel. The variation between models highlights evaluation subjectivity and the need for multiple judges. Most entries receive middle to upper-middle scores, indicating room for improvement in both novelty and usefulness. The data shows a correlation between novelty and usefulness scores across all models, suggesting these attributes often coincide in hackathon entries.

5.1 Comparing Humans and LLM-as-a-judge Ratings

As a next step in exploring LLMs-as-a-judge to evaluate creativity, we wanted to investigate the agreement between humans and LLMs evaluating the novelty and usefulness of hackathon projects. We randomly sampled 30 projects and had two human raters—experienced with hackathon organization, participation, and judging—evaluate each project. In turn, these evaluations are then compared to the novelty and usefulness scores from the LLMs 5.1.

We measured the inter-rater reliability with unweighted Cohen's κ [17]. Figures 5 to 5f display these results, with three levels of granularity: (a-b) scores bucketed as 1–3 and 4–5, (c-d) bucketed as 1–2, 3, and 4–5, and (e-f) each score treated as its own bucket. The human raters showed low agreement on novelty, with Cohen's Kappa decreasing as granularity increased (e.g., $k = 0.18$ in a-b vs. $k = 0.04$ in e-f), suggesting subjectivity in evaluating novelty. Conversely, they showed higher consistency for usefulness ($k = 0.53$ in a-b, $k = 0.12$ in e-f), indicating shared criteria for this dimension. This aligns with related work, which indicate that when rating creativity in the form of novelty and usefulness in complex projects (like urban planning), judges put more emphasis on usefulness when rating such projects whereas they put more emphasis on novelty in alternative uses tests [42].

Agreement between humans and LLMs varied. Human 1 agreed more with Llama ($k = 0.42$ in a-b), and both humans showed fair agreement with Gemma. Among LLMs, Prometheus and Llama achieved the highest agreement for novelty ($k = 0.59$ in a-b), while Mistral and Llama had the best agreement for usefulness ($k = 0.46$). Fleiss' Kappa scores further indicated stronger agreement for usefulness than novelty across all settings. The discrepancies among LLMs likely stem from differences in pre-training data, as models are trained on varying data snapshots.

In creativity evaluations, it is common to prioritize rater consistency—how raters rate items relative to each other—over achieving absolute agreement, where all raters provide identical scores [14]. Therefore, it is not surprising that agreement between models and humans is low. Hence, instead of framing LLMs as an equal to human creativity raters, we propose Human-AI collaboration on creativity evaluation.

5.2 Towards AI Collaborative Creativity Evaluations

The same two human raters also had a detailed discussion on the generated output of the four judge models. Empirically, the raters decided on two additional ratings for the output of the LLMs, one is about *representativeness* (R; rating 1–5, i.e., how well the LLM output encapsulated the actual hackathon description?) and *expert usefulness* (EU; rating 1–5, i.e., can an expert or rater make use of this generated output in their decision making?). We bucket the scores 1–2 and 3–5 and compare agreement over these. For Prometheus, there is high agreement among the human raters that the generated output is both representative and useful for an expert to use to integrate in their decision (R; $k = 0.47$, EU; $k = 0.44$). For the other models, we observed low agreement among representativeness, likely due to the generated output being shorter than Prometheus' (R; $0.05 < k < 0.2$). With respect to expert usefulness, we overall see a higher agreement as the generated output can still support a judge in their evaluation of a hackathon project's creativity (EU; $0.20 < k < 0.30$).

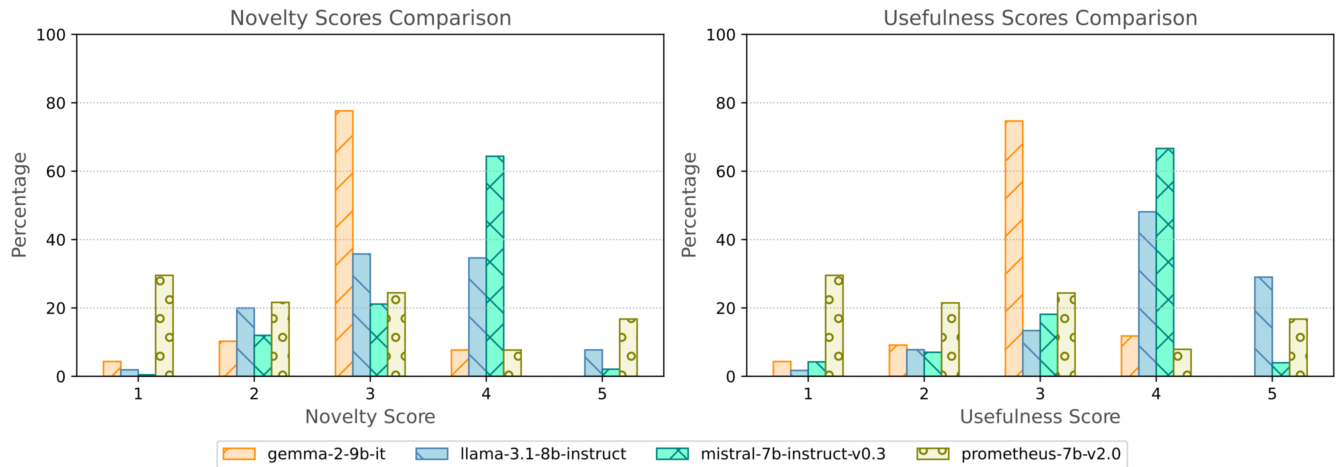


Figure 4: Score distribution of Novelty and Usefulness scores of LLM-as-a-judge on the subset of hackathons.

More broadly, there are several key findings that emerged highlighting the complexities involved in using LLMs for such evaluations. Firstly, assessing creativity requires temporal context; a project that was novel and useful at a certain point in time may lose its novelty as similar solutions emerge over time. This necessitates situating creative assessments within the appropriate time frame and considering changing conditions, such as those brought about by the COVID-19 pandemic, which can affect a project's relevance.

Secondly, the models often provided overly optimistic evaluations with little nuance (i.e., scoring similarly), sometimes based on unfounded assumptions about the projects' functionalities. They occasionally perceived usefulness where human evaluators did not, leading to discrepancies. For instance, Prometheus and Llama disagreed on how much a project's approach differed from existing ones, likely due to variations in their training data and embedded knowledge. This inconsistency poses a challenge for human evaluators, especially novices, who may not have the expertise to verify the models' claims. For example, Llama3.1 generated "*The project description lacks a clear and unique concept, relying heavily on existing ideas (e.g., local hack day inspiration). However, it addresses a practical problem (communication assistance) and presents a feasible solution (a speaking assistant). The use of Python as a development tool is common, but the potential for adding features to the bot suggests some originality in its approach*". This however, is based on an almost empty project description.

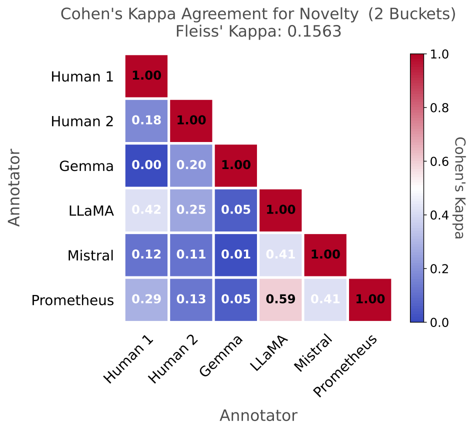
Finally, the models tended to focus on evaluating the prototypes rather than the underlying ideas or their potential for future development. This raises questions about the models' ability to assess the possible potential of creative projects. Moreover, hackathons are celebrated not only for fostering the development of creative ideas but also for offering participants valuable educational and hands-on experiences. In one example from a project description, the creators described how "Neither of us have ever done anything with Machine Learning, so this was something we were proud of."

6 Discussion

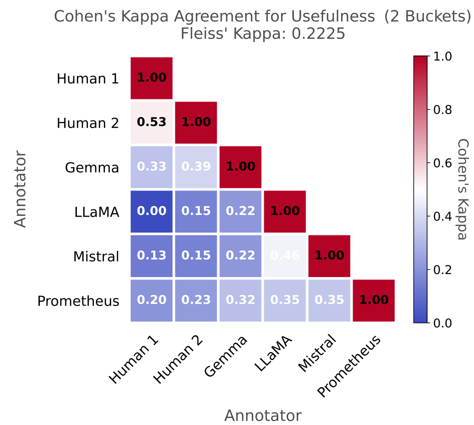
According to Devpost[4], the largest hackathon database, over 1,000 hackathons are held annually, though the actual number is likely higher, as many are not registered [23]. From a creativity research perspective, hackathons offer valuable opportunities to study large-scale, real-world projects addressing problems creatively. This paper is a first exploration of evaluating creativity at scale. Prior work has mainly focused on studying few events that took place in a specific context and that were organized and attended by specific individuals with certain backgrounds, motivations and goals [11, 27, 28, 80] which limits the usefulness of the reported findings to the studied contexts. To expand this body of knowledge we conducted an analysis on the dataset of hackathon projects, repeating and expanding on the approach by Fang, Herbsleb and Vasilescu [25], and subsequently exploring automated evaluation of creativity by using LLMs-as-a-judge, inspired by the approach by [58]. In addition to insights on organizing hackathons to foster creativity (see section 4), we discuss our evaluation method, and suggest directions for future research.

Plucker, Beghetto and Dow emphasized the importance of considering "creativity for whom" and "creativity in what context" [44] when defining creativity. Creativity is not a uniform but operates on different scales. Kaufman and Beghetto's Four C Model of Creativity aims to capture these scales: mini-C (i.e. personal and developmental aspects of creativity inhering in learning processes), Little-C (i.e. everyday creativity which the average person may engage in), Pro-C (i.e. highly accomplished but not yet eminent forms of creative expression), and Big-C (i.e. creative genius or extraordinary creativity) [47]. Even though a hackathon Creation may not have been identified as creative in our analysis, the Creators may have experienced it as very creative to them personally, indicating a mini-C experience of creativity, which we also observed as the last point in section 5.2. These mini-C experiences, including personal learning, are key motivators for hackathon participation [86]. Future evaluation metrics could consider multiple levels of creativity

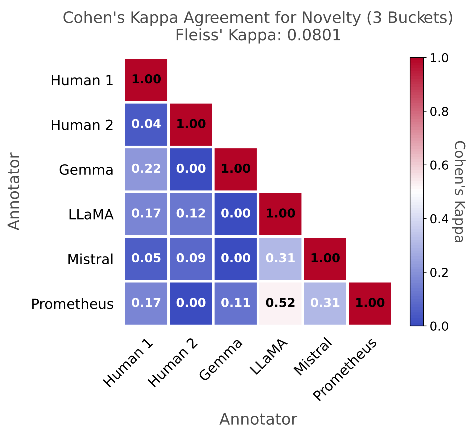
[4]https://devpost.com/hackathons



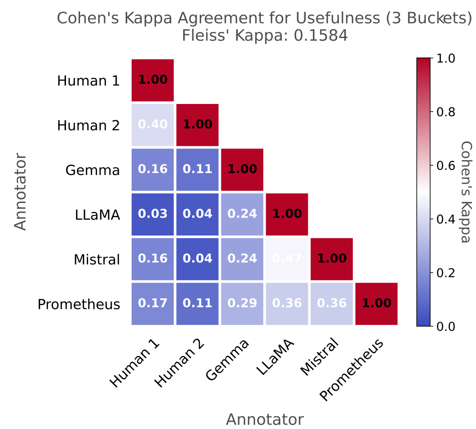
(a) Inter-rater reliability measured by Cohen's Kappa for Novelty. The value among all raters is 0.1563.



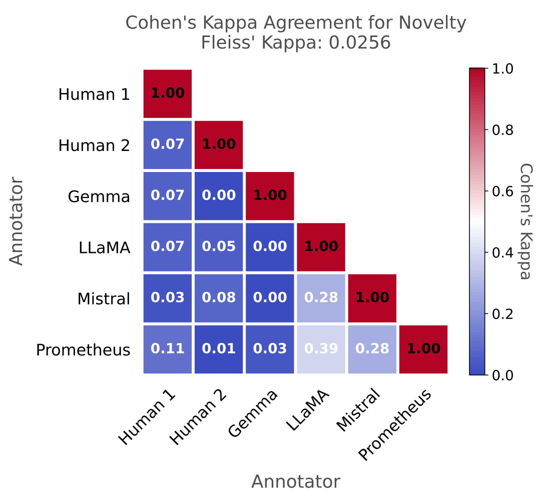
(b) Inter-rater reliability measured by Cohen's Kappa for Usefulness. The value among all raters is 0.2225.



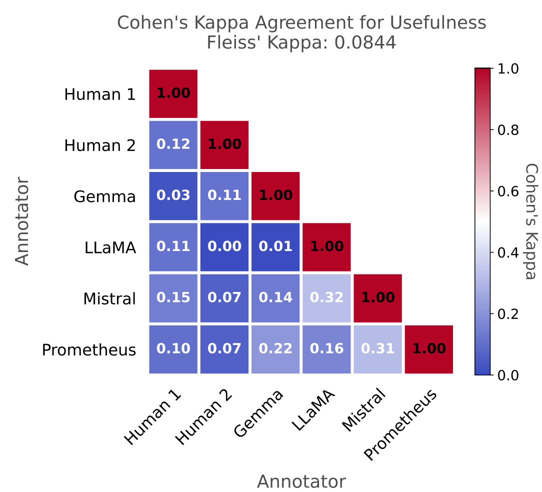
(c) Inter-rater reliability measured by Cohen's Kappa for Novelty. The value among all raters is 0.0801.



(d) Inter-rater reliability measured by Cohen's Kappa for Usefulness. The value among all raters is 0.1584.



(e) Inter-rater reliability measured by Cohen's Kappa for Novelty. The value among all raters is 0.0256.



(f) Inter-rater reliability measured by Cohen's Kappa for Usefulness. The value among all raters is 0.0844.

Figure 5: Heatmaps of inter-rater reliability for both novelty and usefulness in hackathons in different orders of granularity, i.e., how the scores are bucketed. (a-b) contain scores (1-2-3, 4-5), (c-d) contains scores (1-2, 3, 4-5), and (e-f) contains (1-5) in separate buckets.

to account for *tangible* (project-level) and *intangible* (experience and learning) outcomes [22].

6.1 Take-Aways for Organizers

One contribution of this work is to develop our understanding of how individual participants, teams, and event characteristics foster creativity in hackathons (RQ2), leading to concrete recommendations for hackathon organizers. Moreover, we expect our findings to be useful for organizers of hackathons across different contexts, since they were obtained from a large-scale study rather than a study of few specific events.

(1) Creative projects are more likely to emerge in hackathons with around 60 to 80 participants (see fig. 3), potentially due to greater opportunities for team interaction, progress presentations, and feedback. Smaller events also may offer more to resources such as mentors or tools enhancing teams' ability to develop creative solutions. These findings add nuance to Attalah, Nylund and Brem's discussion on how hackathons is a form of collective creativity which can feed into collective intelligence [5]. Our results suggest collective creativity is not universal across *all* hackathons but may depend on event size, with smaller events being more conducive than larger ones. This is related to our finding that **repeated collective participation in different hackathons with the same team is related to creative projects** contrary to repeated individual participation in hackathons.

(2) Larger team sizes are positively associated with creativity, countering prior research suggesting that larger teams inhibit idea generation due to evaluation apprehension [12]. In hackathons, teams must both generate ideas and produce functioning prototypes. This dual requirement may explain why larger teams succeeded despite the risk of apprehension. Our analysis suggests an optimal team size of four to five members may represent this sweet spot (Fig. 3, left), however future research is needed to validate this finding.

(3) Perceived competition can foster creativity on a team level but can at the same time be detrimental on an individual level. Organizers thus need to walk the tight rope of making the event competitive while at the same time providing an environment where individuals feel safe enough to develop risky and creative ideas [32]. One approach could be to provide a few prizes for the most creative projects (fostering perceived team competition), while at the same time making sure that individuals work in larger (see the next point) stable teams which may shield them from perceiving the competition as too daunting.

(4) Teams with less diverse interests are related to creative projects, a counterintuitive finding given that functional diversity (i.e, education, job-relevant knowledge, skills, and abilities) often support creativity [76]. In hackathons, developing a creative idea is only the start, though. Teams also need to develop an artifact and diverse teams might have a harder time to quickly decide which project to work on leaving less time for development. This aligns with observations from Irani's ethnographic study of a hackathon "When there isn't time, you don't want to bring people into the room who are too different from you, who see things differently, or you think might create conflict." [41].

(5) Prior hackathon experience is positively associated with creative projects, potentially because ideas evolve over time. Experienced participants may draw on past hackathons to refine their projects and apply domain knowledge effectively [90]. While prior research highlights how diverse teams with various knowledge bases lead to more creativity [66, 87], our findings suggest that experience itself is a key driver in creativity in hackathons.

Finally, **(6) creative projects receive more likes on Devpost than less creative ones**, suggesting that other participants are skilled at identifying creativity. It is generally agreed upon that creative outcomes are best judged by experts familiar with the relevant domain, as creativity is determined by their independent agreement. Thus if the participants themselves are considered to be experts, our findings fit well with theory [3, 18]. Another explanation for our finding could also be that teams might actively promote their project if they perceive it to be creative, which might manifest itself in the form of likes in our dataset. A similar finding has been reported in the context of hackathon project continuation, where one of the predictors of the long-term survival of projects was reported to be teams promoting their project [65].

These findings may be valuable for organizers to support them in fostering creative projects in their events. It might be advisable to run smaller (60 to 80 participants) rather than larger events that offer opportunities for teams to interact and see each other's work during the event. Moreover, while the focus is often on attracting newcomers, it might be advisable to invite experienced teams with individuals who have participated in multiple hackathons. In relation to this point, it might also be advantageous to encourage people to participate in hackathons in the future. The interactions that they experienced during one event might as well be the spark that helped them develop a creative project during the next one. Finally, our findings show that it might be advisable to trust the wisdom of the hackathon peers when trying to identify creative projects.

6.2 Evaluating Creativity at Scale: Challenges and Opportunities

To address RQ1, we needed to take several decisions to operationalize theory from creativity research into measurable constructs for data science. Although Plucker, Beghetto and Dow called for the creativity research community to develop a unified definition of creativity 20 years ago [44], "the field of creativity research has continued to lack solidarity and cohesiveness" [72]. Needless to say, it is a challenge to then quantify and operationalize creativity constructs to enable a large scale data analysis and, furthermore, develop an automated evaluation. While our contribution is moving large-scale evaluation towards real-world data compared to prior, related work, filtering and analyzing a large-scale dataset is a trade-off between accuracy and feasibility. Our findings are therefore limited to the kind of method which we have constructed in order to answer our research question of how hackathons are creative, and in the following subsections we discuss the challenges and opportunities of our approach.

6.2.1 Operationalizing Novelty and Usefulness. As this research is an initial exploration of a large-scale data analysis and automated evaluation of creativity, we do not claim that our operationalization of the creativity definition, or its constructs of *novelty* and *usefulness* are the only way nor the most accurate way to identify creative hackathon projects from the dataset. Our definition of novelty and usefulness may have excluded some hackathon projects which others may have deemed creative. While we aimed at replicating Fang, Herbsleb and Vasilescu's approach for operationalizing novelty [25] – by finding projects with unusual combinations of software packages and libraries for the five most popular programming languages – it is just one way of capturing novelty. Although we would argue that we thereby should capture a good representation of creative projects, this of course limits the analyzed dataset and potentially excludes some hackathon projects which could be considered creative. Furthermore, this approach also only captures creative software development, whereas creativity in hackathon projects may also happen on the user interface/interaction design side of things, where a very basic and "uncreative" combination of software is still used to develop a very novel and creative interface/interaction design.

To operationalize usefulness in the context of a hackathon, we utilized the "winner"-tag in our dataset, hypothesizing that this could serve as a proxy for an expert evaluating a project as having addressed a challenge or contributing to the theme of an event in a useful and creative way. This operationalization, however, has limitations in that from the dataset we used it is not clear which criteria hackathon organizers utilized to judge projects. Moreover, some of the winners might be chosen by popular vote. It is still reasonable to assume that it is more likely for teams to win at a hackathon that created an artifact which can be considered useful to address a challenge or create an opportunity that did not exist before.

We delimited our analysis of Creations to a single experiment in which we measured the cosine similarities between sentence representations of textual representations including, e.g., the requirements of hackathons. While a more large-scale investigation could have afforded further insights into this particular aspect of hackathons, we leave it to future work to explore, e.g., using LLMs-as-a-judge to enable such large-scale analysis without being overly costly.

Future research could explore approaches for operationalizing creativity constructs further to validate them, or explore other ways of operationalizing novelty and usefulness in meaningful ways. Hackathon platforms such as Devpost could consider creating templates for hackathon descriptions which enable users to reflect on and write how their project is creative in terms of novelty and usefulness, to enable creativity assessment not just for creativity researchers but also for hackathon judges and perhaps recruiters as well.

6.2.2 LLM-as-a(n additional)-Judge. The motivation for exploring LLMs as-a-judge in large-scale creativity evaluation is compelling (RQ3). Because of their training on huge datasets, LLMs possess broad and detailed information about many different topics and domains which they can draw on in their inference phase. When assessing a product for its creativity, an expert would draw on their detailed knowledge about a domain to judge whether a product within that domain is creative—i.e. novel and useful within that context—or not. Recruiting human experts for all the different topics in our hackathon project dataset would be practically infeasible, which underscores the need for exploring automated approaches such as LLMs for this. Hence, a few LLMs should be able to judge a large dataset with an expert view, because they possess detailed information about probably all of the relevant topics in the dataset. This could potentially also circumvent the limitations of attempts to operationalize creativity for statistical analysis, which risk being too constrained as discussed above.

For this reason, Luchini and colleagues have therefore explored LLMs to judge the creativity of creative problem-solving tasks [58]. Our findings from Section 5, which explores LLM as-a-judge, differ somewhat from Luchini and colleagues' approach, despite overall similarities [58]. While they fine-tuned language models to score quality and originality in creative problem-solving tasks, demonstrating high correlation with human scoring, we contend this method has limitations. Their models, trained on task-specific texts and human ratings, naturally correlate well when applied to similar texts. However, this correlation may be more a result of the training process on homogeneous data than a true measure of creativity assessment. Our method addresses this potential bias by using an out-of-the-box LLM, not fine-tuned on the task at hand, to investigate whether its predictions still correlate with human ratings.

Despite their potential, LLMs as judges face notable challenges. Our findings reveal their tendency to provide overly optimistic and uniform evaluations, undermining their ability to offer nuanced critique or differentiate effectively between creative projects. This phenomenon aligns with the broader trend of designing AI systems to align with "human values" by prioritizing servility and safety [13]. While this approach has commercial success, it limits the models' capacity for critical engagement. The tradeoff between "pleasing" and "provoking" interactions highlights a fundamental challenge for AI in balancing reinforcement with critical divergence [77, 84]. For hackathon evaluations, this limitation suggests the need for mechanisms that encourage models to assess more critically and avoid falling into overly affirmative patterns.

One of the most relevant works for our discussion is by Organisciak et al. [68], who not only fine-tuned GPT-style models similar to Luchini and colleagues, but also explored a "few-shot" approach with GPT-4 [2], presenting several task examples in the prompt instead of attempting it "zero-shot" like us (i.e., no task examples). Their findings show that GPT-4 achieves relatively good correlation with human ratings on Alternative Use Tests (AUT) responses (r = .70). However, we argue that hackathon project ratings involve greater subjectivity, as evidenced by our results in Section 5. Nevertheless, following [68]'s approach, we see merit in providing demonstrations to LLMs on rating hackathon project descriptions or fine-tuning them on a set of examples. We hypothesize that this method could potentially improve agreement between humans and LLMs. Additionally, research by Kim and colleagues [48] suggests that providing detailed explanations for each score in a rubric improves scoring robustness, offering another avenue for improvement. An important consideration is the current context length limitation of LLMs, typically 8,192 subwords (chunks of a word) for the model sizes we used. This constraint has implications not only for potentially lengthy hackathon project descriptions but also for applications in other domains such as patent, grant funding and academic paper analysis.

Rather than positioning LLMs as equal to human judges, we propose framing them within the paradigm of hybrid intelligence [61, 62, 84], where LLMs complement and enhance human evaluation and decision-making processes. For instance, in the context of judging creative projects within a single hackathon, LLMs could serve as supplementary judges, providing valuable support to human judges. They might offer synthesized overviews of large datasets, highlight trends, and pose probing questions to foster deeper insights. Beyond judging, a fine-tuned LLM trained on hackathon-specific data could act as an assistant to hackathon organizers, helping uncover statistically significant relationships between factors such as team composition, participant confidence, and hackathon planning effectiveness. Additionally, building on the statistical analysis in Section 4 and the predictive model for creativity introduced in Section 3.2, a direction for future work would be to explore how the model's outcomes align with the creativity judgments provided by LLMs.

6.2.3 Other Venues for Exploring Real-World Large-Scale Data on Creativity. Data on hackathon projects are one way to identify real-world data on creativity. For creativity researchers wishing to explore similar approaches to large-scale analyses of creativity as ours, we see opportunities in a range of other disciplines for this pursuit. Grant writing has been suggested as a form of creative writing: "Creativity often sets apart winnable, funded grant projects from projects that are less impressive to funders in real-world settings" [31]. Continuing the theme of creative writing, papers and patents could be considered here too. In a similar vein, Park, Leahey and Funk conducted a large-scale analysis on papers and patents and calculated their "consolidating or disruptive nature" [69]. Such an analysis could also explore their creativity in terms of novelty and usefulness. Crowdsourcing social innovation platforms, like OpenIDEO[5], are also interesting for exploring large scale creativity evaluations [95].

6.3 Environmental Impact

We acknowledge that conducting a large-scale analysis using LLMs comes with an environmental impact. Experiments were conducted using private infrastructure in Denmark, which has a carbon efficiency of 0.115 $kgCO_2eq/kWh$ in the month of August. A cumulative of 109 GPU hours (summed for all four LLMs) of computation was performed on NVIDIA A40 GPUs, which has a TDP of 300 Watts. Total emissions are estimated to be 3.76 $kgCO_2eq$. Estimations were conducted using the Machine Learning Impact calculator[6] presented in [52].

7 Conclusion

In this paper, we have explored methods for evaluating creativity at scale. We frame hackathon projects as valuable real-world data to evaluate creativity, however, while we have sought to replicate and extend prior research contributions' approaches for doing this, going from process-related creativity evaluations to evaluating creativity in real-world data is challenging. We replicated and extended the prior work in the following ways: (1) we operationalized not only novelty as a creativity construct, but also usefulness as a creativity construct. We filtered the dataset with these two operationalized constructs and discussed the most interesting findings from this subset, which we framed as creative. (2) We explored recent calls to explore LLM-as-a-judge to provide further insights into creativity from the full dataset, and to discuss how LLMs may be framed as an additional judge which can support human raters. We discuss the overall challenges and opportunities for evaluating creativity from large-scale data, which remains a challenge to do. Our findings also challenge some of the previous work which has been done in this area and discuss future research directions.

[5]https://www.openideo.com/

[6]Find the tool here: https://mlco2.github.io/impact.

Acknowledgments

YC and JB are funded by the Carlsberg Foundation, under the Semper Ardens: Accelerate Programme (project nr. CF21-0454). MZ and JB are funded by the Villum Foundation (project nr. VIL57392). We would further like to thank the Aarhus University Research Foundation and Center for Shaping digital citizenship.

References

[1] Selcuk Acar. 2023. Creativity assessment, research, and practice in the age of artificial intelligence. *Creativity Research Journal* (2023), 1–7.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[3] Teresa M. Amabile. 1982. Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology* 43, 5 (1982), 997–1013. https://doi.org/10.1037/0022-3514.43.5.997

[4] Teresa M Amabile, Constance N Hadley, and Steven J Kramer. 2002. Creativity under the gun. *Harvard business review* 80 (2002), 52–63.

[5] Issam Attalah, Petra A. Nylund, and Alexander Brem. 2023. Who captures value from hackathons? Innovation contests with collective intelligence tools bridging creativity and coupled open innovation. *Creativity and Innovation Management* 32, 2 (2023), 266–280. https://doi.org/10.1111/caim.12552 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/caim.12552

[6] Baptiste Barbot. 2019. Measuring creativity change and development. *Psychology of Aesthetics, Creativity, and the Arts* 13, 2 (2019), 203.

[7] Baptiste Barbot and Roni Reiter-Palmon. 2019. Creativity assessment: Pitfalls, solutions, and standards. *Psychology of Aesthetics, Creativity, and the Arts* 13, 2 (2019), 131–132. https://doi.org/10.1037/aca0000251

[8] Roger E Beaty and Dan R Johnson. 2021. Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior research methods* 53, 2 (2021), 757–780.

[9] Roger E Beaty, Dan R Johnson, Daniel C Zeitlen, and Boris Forthmann. 2022. Semantic distance and the alternate uses task: Recommendations for reliable automated assessment of originality. *Creativity Research Journal* 34, 3 (2022), 245–260.

[10] Ronald A. Beghetto. 2019. Large-Scale Assessments, Personalized Learning, and Creativity: Paradoxes and Possibilities. *ECNU Review of Education* 2, 3 (2019), 311–327. https://doi.org/10.1177/2096531119878963 arXiv:https://doi.org/10.1177/2096531119878963

[11] Gerard Briscoe. 2014. Digital Innovation: The Hackathon Phenomenon. https://api.semanticscholar.org/CorpusID:31734217

[12] Kris Byron, Sejin Keem, Tanja Darden, Christina E. Shalley, and Jing Zhou. 2023. Building blocks of idea generation and implementation in teams: A meta-analysis of team design and team creativity and innovation. *Personnel Psychology* 76, 1 (2023), 249–278. https://doi.org/10.1111/peps.12501 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/peps.12501

[13] Alice Cai, Ian Arawjo, and Elena L. Glassman. 2024. Antagonistic AI. *arXiv preprint arXiv:2402.07350* (2024). https://arxiv.org/abs/2402.07350v1

[14] Simon Majed Ceh, Carina Edelmann, Gabriela Hofer, and Mathias Benedek. 2022. Assessing Raters: What Factors Predict Discernment in Novice Creativity Raters? *The Journal of Creative Behavior* 56, 1 (2022), 41–54. https://doi.org/10.1002/jocb.515 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/jocb.515

[15] Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the Use of Large Language Models for Reference-Free Text Quality Evaluation:

An Empirical Study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi (Eds.). Association for Computational Linguistics, Nusa Dua, Bali, 361–374. https://doi.org/10.18653/v1/2023.findings-ijcnlp.32

[16] Cheng-Han Chiang and Hung-yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations?. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 15607–15631. https://doi.org/10.18653/v1/2023.acl-long.870

[17] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[18] Genevieve M Cseh and Karl K Jeffries. 2019. A scattered CAT: A critical evaluation of the consensual assessment technique for creativity research. *Psychology of Aesthetics, Creativity, and the Arts* 13, 2 (2019), 159.

[19] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[20] Amy C Edmondson. 2012. *Teaming: How Organizations Learn, Innovate, and Compete in the Knowledge Economy.* John Wiley & Sons.

[21] Amy C. Edmondson and Stacy E. Mcmanus. 2007. Methodological fit in management field research. *Academy of Management Review* 32, 4 (2007), 1246–1264. https://doi.org/10.5465/amr.2007.26586086 arXiv:https://doi.org/10.5465/amr.2007.26586086

[22] Jeanette Falk, Gopinaath Kannabiran, and Nicolai Brodersen Hansen. 2021. What Do Hackathons Do? Understanding Participation in Hackathons Through Program Theory Analysis. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 147, 16 pages. https://doi.org/10.1145/3411764.3445198

[23] Jeanette Falk, Alexander Nolte, Daniela Huppenkothen, Marion Weinzierl, Kiev Gama, Daniel Spikol, Erik Tollerud, Neil Chue Hong, Ines Knäpper, and Linda Bailey Hayden. 2024. The Future of Hackathon Research and Practice. *IEEE Access* (2024), 1–1. https://doi.org/10.1109/ACCESS.2024.3455092

[24] Jeanette Falk Olesen and Kim Halskov. 2020. 10 Years of Research With and On Hackathons. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (Eindhoven, Netherlands) *(DIS '20)*. Association for Computing Machinery, New York, NY, USA, 1073–1088. https://doi.org/10.1145/3357236.3395543

[25] Hongbo Fang, James Herbsleb, and Bogdan Vasilescu. 2024. Novelty Begets Long-Term Popularity, But Curbs Participation. (2024).

[26] Boris Forthmann, Heinz Holling, Nima Zandi, Anne Gerwig, Pınar Çelik, Martin Storme, and Todd Lubart. 2017. Missing creativity: The effect of cognitive workload on rater (dis-) agreement in subjective divergent-thinking scores. *Thinking Skills and Creativity* 23 (2017), 129–139.

[27] Frank J. Frey and Michael Luks. 2016. The innovation-driven hackathon: one means for accelerating innovation. In *Proceedings of the 21st European Conference on Pattern Languages of Programs* (Kaufbeuren, Germany) *(EuroPlop '16)*. Association for Computing Machinery, New York, NY, USA, Article 10, 11 pages. https://doi.org/10.1145/3011784.3011794

[28] Kiev Gama, George Valença, Pedro Alessio, Rafael Formiga, André Neves, and Nycolas Lacerda. 2023. The developers' design thinking toolbox in hackathons: a study on the recurring design methods in software development marathons. *International Journal of Human–Computer Interaction* 39, 12 (2023), 2269–2291.

[29] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* 120, 30 (2023), e2305016120.

[30] Vlad Petre Glăveanu. 2013. Rewriting the Language of Creativity: The Five A's Framework. *Review of General Psychology* 17, 1 (2013), 69–81. https://doi.org/10.1037/a0029528 arXiv:https://doi.org/10.1037/a0029528

[31] Ellen W. Gorsevski, Kate Magsamen-Conrad, and Lisa Kiyomi Hanasono. 2019. Grant Writing as a Creative Process. *The Oxford Handbook of Methods for Public Scholarship* (2019). https://api.semanticscholar.org/CorpusID:210484339

[32] Daniel P. Gross. 2020. Creativity Under Fire: The Effects of Competition on Creative Production. *The Review of Economics and Statistics* 102, 3 (07 2020), 583–599. https://doi.org/10.1162/rest_a_00831 arXiv:https://direct.mit.edu/rest/article-pdf/102/3/583/1881450/rest_a_00831.pdf

[33] Eran Hadas and Arnon Hershkovitz. 2024. Using large language models to evaluate alternative uses task flexibility score. *Thinking Skills and Creativity* 52 (2024), 101549. https://doi.org/10.1016/j.tsc.2024.101549

[34] Andrew B. Hargadon and Beth A. Bechky. 2006. When Collections of Creatives Become Creative Collectives: A Field Study of Problem Solving at Work. *Organization Science* 17, 4 (2006), 484–500. https://doi.org/10.1287/orsc.1060.0200

[35] David JP Heinen and Dan R Johnson. 2018. Semantic distance: An automated measure of creativity that is novel and appropriate. *Psychology of Aesthetics, Creativity, and the Arts* 12, 2 (2018), 144.

[36] Ben Heller, Atar Amir, Roy Waxman, and Yossi Maaravi. 2023. Hack your organizational innovation: literature review and integrative model for running hackathons. *Journal of Innovation and Entrepreneurship* 12, 1 (03 Mar 2023), 6. https://doi.org/10.1186/s13731-023-00269-0

[37] Fan Huang, Haewoon Kwak, Kunwoo Park, and Jisun An. 2024. ChatGPT Rates Natural Language Explanation Quality like Humans: But on Which Scales?. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 3111–3132. https://aclanthology.org/2024.lrec-main.277

[38] Daniela Huppenkothen, Anthony Arendt, David W Hogg, Karthik Ram, Jacob T VanderPlas, and Ariel Rokem. 2018. Hack weeks as a model for data science education and collaboration. *Proceedings of the National Academy of Sciences* 115, 36 (2018), 8872–8877.

[39] Ahmed Imam and Tapajit Dey. 2021. Tracking Hackathon Code Creation and Reuse. In *2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR)*. 615–617. https://doi.org/10.1109/MSR52588.2021.00085

[40] Ahmed Imam, Tapajit Dey, Alexander Nolte, Audris Mockus, and James D. Herbsleb. 2021. The Secret Life of Hackathon Code Where does it come from and where does it go?. In *2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR)*. 68–79. https://doi.org/10.1109/MSR52588.2021.00020

[41] Lilly Irani. 2015. Hackathons and the making of entrepreneurial citizenship. *Science, Technology, & Human Values* 40, 5 (2015), 799–824.

[42] Alan Pickering James Lloyd-Cox and Joydeep Bhattacharya. 2022. Evaluating Creativity: How Idea Context and Rater Personality Affect Considerations of Novelty and Usefulness. *Creativity Research Journal* 34, 4 (2022), 373–390. https://doi.org/10.1080/10400419.2022.2125721 arXiv:https://doi.org/10.1080/10400419.2022.2125721

[43] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).

[44] Ronald A. Beghetto Jonathan A. Plucker and Gayle T. Dow. 2004. Why Isn't Creativity More Important to Educational Psychologists? Potentials, Pitfalls, and Future Directions in Creativity Research. *Educational Psychologist* 39, 2 (2004), 83–96. https://doi.org/10.1207/s15326985ep3902_1 arXiv:https://doi.org/10.1207/s15326985ep3902_1

[45] Florence Bazzaro Justine Lobbe and Jean-Claude Sagot. 2021. Innovation in collaborative design: an exploratory study in hackathon. *International Journal of Design Creativity and Innovation* 9, 2 (2021), 119–137. https://doi.org/10.1080/21650349.2021.1894989 arXiv:https://doi.org/10.1080/21650349.2021.1894989

[46] Joakim Karlsen and Anders Sundnes Løvlie. 2017. 'You can dance your prototype if you like': independent filmmakers adapting the hackathon. *Digital Creativity* 28, 3 (2017), 224–239. https://doi.org/10.1080/14626268.2017.1351992 arXiv:https://doi.org/10.1080/14626268.2017.1351992

[47] James C. Kaufman and Ronald A. Beghetto. 2009. Beyond Big and Little: The Four C Model of Creativity. *Review of General Psychology* 13, 1 (2009), 1–12. https://doi.org/10.1037/a0013688 arXiv:https://doi.org/10.1037/a0013688

[48] Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

[49] Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535* (2024).

[50] Tom Kocmi and Christian Federmann. 2023. Large Language Models Are State-of-the-Art Evaluators of Translation Quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz (Eds.). European Association for Machine Translation, Tampere, Finland, 193–203. https://aclanthology.org/2023.eamt-1.19

[51] Christoph Kollwitz and Barbara Dinter. 2019. What the Hack? – Towards a Taxonomy of Hackathons. In *Business Process Management*, Thomas Hildebrandt, Boudewijn F. van Dongen, Maximilian Röglinger, and Jan Mendling (Eds.). Springer International Publishing, Cham, 354–369.

[52] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the Carbon Emissions of Machine Learning. *arXiv preprint arXiv:1910.09700* (2019).

[53] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. 2024. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787* (2024).

[54] Fei-Fei Li. 2018. How to Make AI That's Good for People. *International New York Times* (2018), NA–NA.

[55] Hila Lifshitz-Assaf, Sarah Lebovitz, and Lior Zalmanson. 2021. Minimal and Adaptive Coordination: How Hackathons' Projects Accelerate Innovation without

Killing it. *Academy of Management Journal* 64, 3 (2021), 684–715. https://doi.org/10.5465/amj.2017.0712 arXiv:https://doi.org/10.5465/amj.2017.0712

[56] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 2511–2522. https://doi.org/10.18653/v1/2023.emnlp-main.153

[57] Todd Lubart. 2017. The 7 C's of Creativity. *The Journal of Creative Behavior* 51, 4 (2017), 293–296. https://doi.org/10.1002/jocb.190 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/jocb.190

[58] Simone Luchini, Nadine T Maliakkal, Paul V DiStefano, John D Patterson, Roger Beaty, and Roni Reiter-Palmon. 2023. Automatic Scoring of Creative Problem-Solving with Large Language Models: A Comparison of Originality and Quality Ratings. (2023).

[59] Ahmed Samir Imam Mahmoud, Tapajit Dey, Alexander Nolte, Audris Mockus, and James D. Herbsleb. 2022. One-off events? An empirical study of hackathon code creation and reuse. *Empirical Software Engineering* 27, 7 (20 Sep 2022), 167. https://doi.org/10.1007/s10664-022-10201-x

[60] Yaoli Mao, Janet Rafner, Yi Wang, and Jacob Sherson. 2023. A Hybrid Intelligence Approach to Training Generative Design Assistants: Partnership Between Human Experts and AI Enhanced Co-Creative Tools. In *HHAI 2023: Augmenting Human Intellect (Frontiers in Artificial Intelligence and Applications, Vol. 368)*. 108–123. https://doi.org/10.3233/FAIA230078

[61] Yaoli Mao, Janet Rafner, Yi Wang, and Jacob Sherson. 2023. A hybrid intelligence approach to training generative design assistants: Partnership between human experts and AI-enhanced co-creative tools. In *HHAI 2023: Augmenting Human Intellect*. IOS Press, 108–123. https://ebooks.iospress.nl/volumearticle/63326

[62] Yaoli Mao, Janet Rafner, Yi Wang, and Jacob Sherson. 2024. HI-TAM, A Hybrid Intelligence Framework for Training and Adoption of Generative Design Assistants. *Frontiers in Computer Science* 6 (2024). https://doi.org/10.3389/fcomp.2024.1460381

[63] Lukas McIntosh and Caroline D. Hardin. 2021. Do Hackathon Projects Change the World? An Empirical Analysis of GitHub Repositories. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education* (Virtual Event, USA) *(SIGCSE '21)*. Association for Computing Machinery, New York, NY, USA, 879–885. https://doi.org/10.1145/3408877.3432435

[64] Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch (Eds.). Association for Computational Linguistics, Toronto, Canada, 394–403. https://doi.org/10.18653/v1/2023.bea-1.32

[65] Alexander Nolte, Irene-Angelica Chounta, and James D. Herbsleb. 2020. What Happens to All These Hackathon Projects? Identifying Factors to Promote Hackathon Project Continuation. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 145 (oct 2020), 26 pages. https://doi.org/10.1145/3415216

[66] Alexander Nolte, Ei Pa Pa Pe-Than, Abasi-amefon Obot Affia, Chalalai Chaihirunkarn, Anna Filippova, Arun Kalyanasundaram, Maria Angelica Medina Angarita, Erik Trainer, and James D Herbsleb. 2020. How to organize a hackathon–A planning kit. *arXiv preprint arXiv:2008.08025* (2020).

[67] Jeanette Falk Olesen, Nicolai Brodersen Hansen, and Kim Halskov. 2018. Four factors informing design judgement at a hackathon. In *Proceedings of the 30th Australian Conference on Computer-Human Interaction* (Melbourne, Australia) *(OzCHI '18)*. Association for Computing Machinery, New York, NY, USA, 473–483. https://doi.org/10.1145/3292147.3292155

[68] Peter Organisciak, Selcuk Acar, Denis Dumas, and Kelly Berthiaume. 2023. Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity* 49 (2023), 101356. https://doi.org/10.1016/j.tsc.2023.101356

[69] Michael Park, Erin Leahey, and Russell J. Funk. 2023. Papers and patents are becoming less disruptive over time. *Nature* 613, 7942 (01 Jan 2023), 138–144. https://doi.org/10.1038/s41586-022-05543-x

[70] John D Patterson, Hannah M Merseal, Dan R Johnson, Sergio Agnoli, Matthijs Baas, Brendan S Baker, Baptiste Barbot, Mathias Benedek, Khatereh Borhani, Qunlin Chen, et al. 2023. Multilingual semantic distance: Automatic verbal creativity assessment in many languages. *Psychology of Aesthetics, Creativity, and the Arts* 17, 4 (2023), 495.

[71] Ei Pa Pa Pe-Than, Alexander Nolte, Anna Filippova, Christian Bird, Steve Scallen, and James Herbsleb. 2022. Corporate hackathons, how and why? A multiple case study of motivation, projects proposal and selection, goal setting, coordination, and outcomes. *Human–Computer Interaction* 37, 4 (2022), 281–313.

[72] Jeb S. Puryear and Kristen N. Lamb. 2020. Defining Creativity: How Far Have We Come Since Plucker, Beghetto, and Dow? *Creativity Research Journal* 32, 3 (2020), 206–214. https://doi.org/10.1080/10400419.2020.1821552 arXiv:https://doi.org/10.1080/10400419.2020.1821552

[73] Janet Rafner, Roger E. Beaty, James C. Kaufman, Todd Lubart, and Jacob Sherson. 2023. Creativity in the age of generative AI. *Nature Human Behaviour* 7, 11 (01 Nov 2023), 1836–1838. https://doi.org/10.1038/s41562-023-01751-1

[74] Janet Rafner, Michael Mose Biskjær, Blanka Zana, Steven Langsford, Carsten Bergenholtz, Seyedahmad Rahimi, Andrea Carugati, Lior Noy, and Jacob Sherson. 2022. Digital Games for Creativity Assessment: Strengths, Weaknesses and Opportunities. *Creativity Research Journal* 34, 1 (2022), 28–54. https://doi.org/10.1080/10400419.2021.1971447

[75] Roni Reiter-Palmon, Boris Forthmann, and Baptiste Barbot. 2019. Scoring divergent thinking tests: A review and systematic framework. *Psychology of Aesthetics, Creativity, and the Arts* 13, 2 (2019), 144.

[76] Roni Reiter-Palmon, Ben Wigert, and Triparna de Vreede. 2012. Team Creativity and Innovation: The Effect of Group Composition, Social Processes, and Cognition. In *Handbook of Organizational Creativity*, Michael D. Mumford (Ed.). Academic Press, 295–326. https://doi.org/10.1016/B978-0-12-374714-3.00013-6 Available online 5 August 2011.

[77] Jeba Rezwana and Mary Lou Maher. 2023. Designing Creative AI Partners with COFI: A Framework for Modeling Interaction in Human-AI Co-Creative Systems. *ACM Trans. Comput.-Hum. Interact.* 30, 5, Article 67 (Sept. 2023), 28 pages. https://doi.org/10.1145/3519026

[78] Mel Rhodes. 1961. An Analysis of Creativity. *The Phi Delta Kappan* 42, 7 (1961), 305–310. http://www.jstor.org/stable/20342603

[79] Michael Ringel, Andrew Taylor, and Hadi Zablit. 2015. The rising need for innovation speed. *The Boston Consulting Group* (2015).

[80] Bard Rosell, Shiven Kumar, and John Shepherd. 2014. Unleashing innovation through internal hackathons. In *2014 IEEE Innovations in Technology Conference*. 1–8. https://doi.org/10.1109/InnoTek.2014.6877369

[81] Andrew R Schrock. 2016. Civic hacking as data activism and advocacy: A history from publicity to open government data. *New media & society* 18, 4 (2016), 581–599.

[82] Cleo Schulten and Irene-Angelica Chounta. 2024. How do we learn in and from Hackathons? A systematic literature review. *Education and Information Technologies* (2024), 1–32.

[83] Joseph Alois Schumpeter. 1939. Business Cycles: A Theoretical, Historical, and Statistical Analysis of the Capitalist Process. *McQraw Hill* (1939).

[84] Jacob Sherson, Janet Rafner, and Safinaz Büyükgüzel. 2024. Operational Criteria of Hybrid Intelligence for Generative AI Virtual Assistants. In *HHAI 2024: Hybrid Human AI Systems for the Social Good (Frontiers in Artificial Intelligence and Applications, Vol. 386)*. 475–477. https://doi.org/10.3233/FAIA240229

[85] Ben Shneiderman. 2020. Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction* 12, 3 (2020), 109–124.

[86] Anthony Simonofski, Victor Amaral de Sousa, Antoine Clarinval, and Benoît Vanderose. 2020. Participation in Hackathons: A Multi-methods View on Motivators, Demotivators and Citizen Participation. In *Research Challenges in Information Science*, Fabiano Dalpiaz, Jelena Zdravkovic, and Pericles Loucopoulos (Eds.). Springer International Publishing, Cham, 229–246.

[87] Carmit T. Tadmor, Patricia Satterstrom, Sujin Jang, and Jeffrey T. Polzer. 2012. Beyond Individual Creativity: The Superadditive Benefits of Multicultural Experience for Collective Creativity in Culturally Diverse Teams. *Journal of Cross-Cultural Psychology* 43, 3 (2012), 384–392. https://doi.org/10.1177/0022022111435259 arXiv:https://doi.org/10.1177/0022022111435259

[88] Nick Taylor and Loraine Clarke. 2018. Everybody's Hacking: Participation and the Mainstreaming of Hackathons. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173746

[89] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118* (2024).

[90] Nelson N. Tenório, Gisele Caroline Urbano Lourenço, Mariana Oliveira, Steffi Aline Stark Becker, Fabrício Tomaz Bernardelli, Hassan Atifi, and Nada Matta. 2019. Rethinking Strategies of Hackathons to Increase Team's Creativity: Findings of a Qualitative Research. In *International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. https://api.semanticscholar.org/CorpusID:204754469

[91] Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588* (2023).

[92] Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models. *arXiv preprint arXiv:2404.18796* (2024).

[93] Shaun Wallace, Brendan Le, Luis A Leiva, Aman Haq, Ari Kintisch, Gabrielle Bufrem, Linda Chang, and Jeff Huang. 2020. Sketchy: Drawing inspiration from the crowd. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–27.

[94] Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a Good NLG Evaluator? A Preliminary Study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, Yue Dong, Wen Xiao, Lu Wang, Fei Liu, and Giuseppe Carenini (Eds.). Association for Computational Linguistics, Singapore, 1–11. https://doi.org/10.18653/v1/2023.newsum-1.1

[95] Rong Wang and Bin Chen. 2023. A Configurational Approach to Attracting Participation in Crowdsourcing Social Innovation: The Case of Openideo. *Management Communication Quarterly* 37, 2 (2023), 340–367. https://doi.org/10.1177/08933189221108360 arXiv:https://doi.org/10.1177/08933189221108360

[96] Chauncey Wilson. 2013. *Brainstorming and beyond: a user-centered design method.* Newnes.

[97] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* 36 (2024).

A LLM-as-a-judge prompt

In Figure 6, we indicate the full model prompt we use to query the four LLMs we use in our study.

System Prompt
You are a rigorous and efficient evaluation assistant for rating creativity. Your task is to provide scores and feedback systematically. Follow the format strictly: first provide the scores for Novelty and Usefulness, then the rationale for each score in under 100 words. Ensure that feedback is clear, concise, and aligned with the rubric provided, avoiding unnecessary commentary.

User Prompt
Task Description: You are tasked with evaluating a project description based on two criteria: Novelty and Usefulness, using a 5-point Likert scale.
1. Write the scores for Novelty and Usefulness as integers between 1 and 5, strictly referring to the score rubric. 2. Provide concise feedback (within 50 words) justifying each score. 3. Format your output as follows: 'Novelty: [RESULT] (1-5) Usefulness: [RESULT] (1-5) Feedback: [Your feedback here]'. 4. Do not include any other text or explanation.
Evaluation Criteria: Novelty: Evaluate how unique and original the project's concept, approach, or solution is. Consider whether it introduces new ideas, methods, or perspectives that differ significantly from existing ones. Usefulness: Evaluate how practical and appropriate the project is in addressing its targeted problem or challenge. Consider whether it effectively solves a real-world issue or meets a specific need.

Project Description to Evaluate:
{Example Hackathon Description}

Figure 6: Prompt for Evaluating Hackathon Descriptions. We show the prompt we use for all four models in our study. First we give a system prompt, in the form of a "model-should-act-as". Then, in subsequent paragraphs, (1) we give detailed feedback for the task itself, (2) give instructions on how to evaluate the text, (3) the hackathon description, and (4) the rubrics that should be used for the final judgment.

B Correlation Tables

Table 2: Pearson Correlation among Variables for Participants. p<0.05 (*), p<0.01 (**), p<0.001 (***).

	#Hackathons	#Projects	#Interests	#Skills	Years of Experience	Has Followers	Has Likes
#Projects	0.82***						
#Interests	0.08***	0.08***					
#Skills	0.15***	0.14***	0.24***				
Years of Experience	0.48***	0.48***	0.03***	0.14***			
Has Followers	0.26***	0.24***	0.12***	0.18***	0.23***		
Has Likes	0.3***	0.28***	0.11***	0.17***	0.28***	0.32***	
AVG. Weighted Winning	0.06***	0.05***	0.05***	0.04***	0.02***	0.06***	0.09***

Table 3: Pearson Correlation among Variables for Collaborations. p<0.05 (*), p<0.01 (**), p<0.001 (***).

	#Participants	Collab. Proj. Repetition	Collab. Hack. Repetition	Proj. Repetition	Hack. Repetition	Interests	Common Interests	Diff. Interests	Skills	Common Skills	Diff. Skills
Collab. Proj. Repetition	0.49***										
Collab. Hack. Repetition	0.29***	0.86***									
Proj. Repetition	0.19***	0.48***	0.35***								
Hack. Repetition	0.14***	0.58***	0.76***	0.58***							
Interests	0.25***	0.13***	0.08***	0.29***	0.17***						
Common Interests	-0.31***	-0.01	-0.01	0.09***	0.09***	0.42***					
Diff. Interests	0.46***	0.14***	0.09***	0.26***	0.14***	0.84***	-0.14***				
Skills	0.6***	0.33***	0.23***	0.27***	0.18***	0.33***	-0.07***	0.4***			
Common Skills	-0.41***	-0.04***	-0.02***	-0.01*	0.02***	0	0.45***	-0.27***	0.06***		
Diff. Skills	0.73***	0.34***	0.23***	0.25***	0.16***	0.31***	-0.24***	0.48***	0.92***	-0.33***	
Winner	0.09***	0.03***	0.05***	0.09***	0.1***	0.08***	-0.02***	0.1***	0.12***	-0.03***	0.13***

Table 4: Pearson Correlation among Variables in Hackathons. p<0.05 (*), p<0.01 (**), p<0.001 (***).

	# sub. Projects	# act. Participants	# Judges	# Sponsors	Onsite	# Days	# Places	Competition (T)	Competition (I)
# act. Participants	0.92***								
# Judges	0.11***	0.08***							
# Sponsors	0.08***	0.08***	0.27***						
Onsite	-0.17***	-0.11***	-0.08***	-0.07***					
# Days	0.04**	0	0.05***	-0.07***	-0.34***				
# Places	0.28***	0.21***	0.23***	0.23***	0.04**	-0.1***			
Competition (T)	-0.22***	-0.17***	-0.07***	0.01	0.13***	-0.07***	0.02		
Competition (I)	-0.17***	-0.14***	-0.08***	-0.03*	0.05***	0	-0.04**	0.85***	
# Winners	0.7***	0.62***	0.1***	0.14***	-0.25***	0.05***	0.27***	-0.18***	-0.15***

C Figures

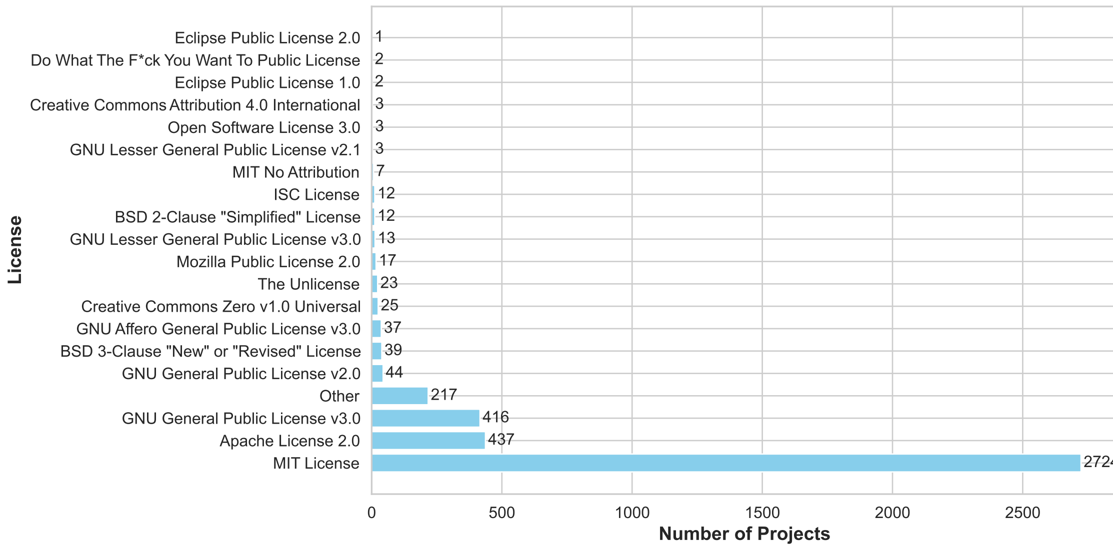


Figure 7: The Distribution of Licenses used in the Projects.

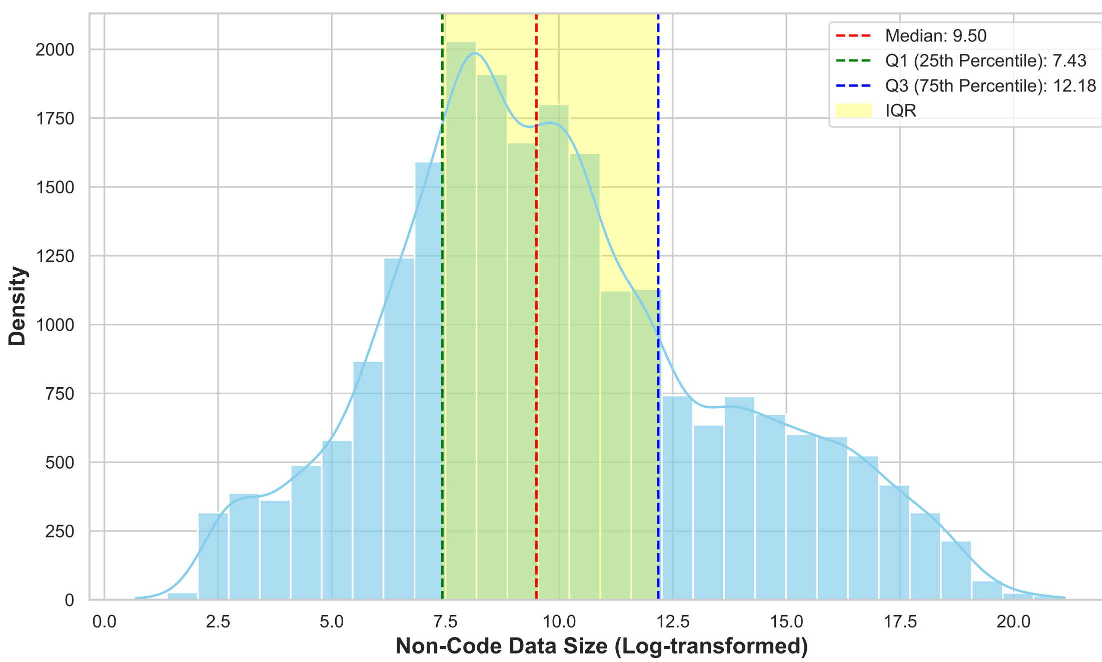


Figure 8: The Distribution of Non-Code Data Sizes in Projects with highlighted Interquartile Range (25% to 75%). 50% of the data sizes are between 1807 to 268336 Bytes.

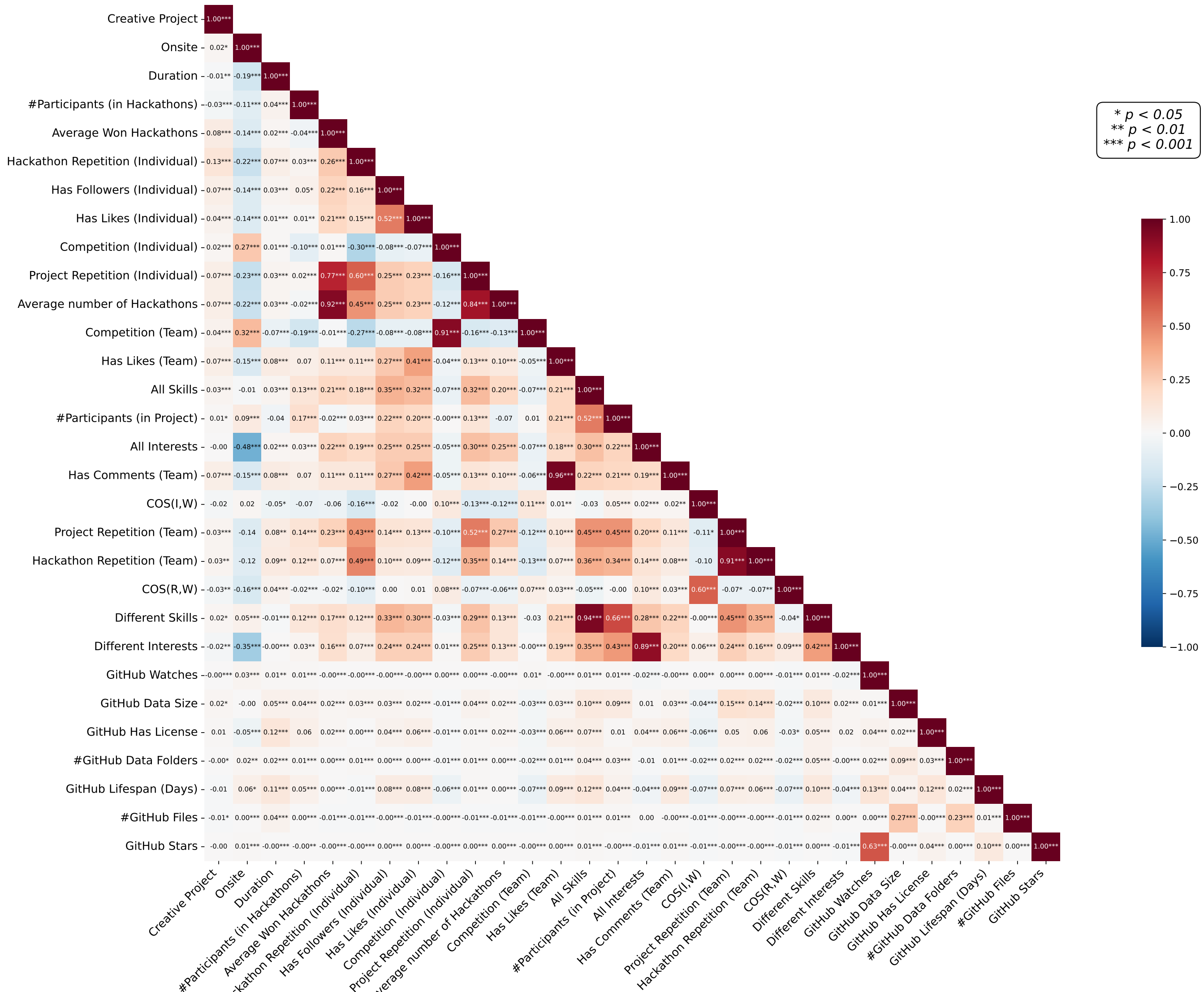


Figure 9: Spearman's Rank Correlation Coefficients among Variables.

D Algorithm

Algorithm 1: Project and Hackathon Collaboration Algorithm

Input: Dataset containing projects, hackathons, and participants

Output: creator2creator-project, creator2creator-hackathon, collaboration repetition and repetition metrics

```
Initialize:
creator2creator_project ← ∅
creator2creator_hackathon ← ∅
collaboration_repetition_project ← ∅
collaboration_repetition_hackathon ← ∅
repetition_project ← ∅
repetition_hackathon ← ∅
foreach project p in dataset['projects'] do
    creators C_p ← p['creators']
    foreach pair of creators (c_i, c_j) in C_p do
        if (c_i, c_j) or (c_j, c_i) exists in creator2creator_project then
            creator2creator_project[(c_i, c_j)] ← creator2creator_project[(c_i, c_j)] + 1
        end
        else
            creator2creator_project[(c_i, c_j)] ← 1
        end
    end
    foreach creator c in C_p do
        collaboration_repetition_project[c] ← collaboration_repetition_project[c] + 1
        repetition_project[c] ← repetition_project[c] + collaboration_repetition_project[c]
    end
end
foreach hackathon h in dataset['hackathons'] do
    creators C_h ← h['creators']
    foreach pair of creators (c_i, c_j) in C_h do
        if (c_i, c_j) or (c_j, c_i) exists in creator2creator_hackathon then
            creator2creator_hackathon[(c_i, c_j)] ← creator2creator_hackathon[(c_i, c_j)] + 1
        end
        else
            creator2creator_hackathon[(c_i, c_j)] ← 1
        end
    end
    foreach creator c in C_h do
        collaboration_repetition_hackathon[c] ← collaboration_repetition_hackathon[c] + 1
        repetition_hackathon[c] ← repetition_hackathon[c] + collaboration_repetition_hackathon[c]
    end
end
return collaboration_repetition_project, collaboration_repetition_hackathon,
repetition_project, repetition_hackathon
```