AI-Augmented Brainwriting: Investigating the use of LLMs in group ideation

Orit Shaer*
oshaer@wellesley.edu
Wellesley College
Wellesley, MA, USA

Angelora Cooper
acooper5@wellesley.edu
Wellesley College
Wellesley, MA, USA

Osnat Mokryn
omokryn@is.haifa.ac.il
University of Haifa
Haifa, Israel

Andrew L. Kun
andrew.kun@unh.edu
University of New Hampshire
Durham, NH, USA

Hagit Ben Shoshan
hagits@gmail.com
University of Haifa
Haifa, Israel

ABSTRACT

The growing availability of generative AI technologies such as large language models (LLMs) has significant implications for creative work. This paper explores twofold aspects of integrating LLMs into the creative process – the divergence stage of idea generation, and the convergence stage of evaluation and selection of ideas. We devised a collaborative group-AI Brainwriting ideation framework, which incorporated an LLM as an enhancement into the group ideation process, and evaluated the idea generation process and the resulted solution space. To assess the potential of using LLMs in the idea evaluation process, we design an evaluation engine and compared it to idea ratings assigned by three expert and six novice evaluators. Our findings suggest that integrating LLM in Brainwriting could enhance both the ideation process and its outcome. We also provide evidence that LLMs can support idea evaluation. We conclude by discussing implications for HCI education and practice.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; **Collaborative interaction**.

KEYWORDS

LLM, Brainwriting, human-AI collaboration

ACM Reference Format:
Orit Shaer, Angelora Cooper, Osnat Mokryn, Andrew L. Kun, and Hagit Ben Shoshan. 2024. AI-Augmented Brainwriting: Investigating the use of LLMs in group ideation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3613904.3642414

1 INTRODUCTION

The increasing availability of generative AI technologies [72] such as large language models (LLMs) [23, 44, 53] and image generators [4, 45, 52] has significant implications for creative work [27, 51]. Given their wide adoption [49], it is critical to investigate the merits and limitations of integrating such tools into the creative process through new forms of co-creation.

Recent work has begun to explore how co-creation with generative AI could be used for interaction design [63] and what co-creation practices might look like for problem solving [62], ideation [37, 73, 78], prototyping, making, and programming [35, 59, 76, 82]. Emerging theories about posthumanism, post-human, and more-than-human interaction design [19, 30, 79, 80] provide further context for human-AI co-creation activities by highlighting possibilities to distribute agency in design among humans and non-humans.

The overarching research question we are interested in is how LLMs can contribute to enhancing the human creative thought process through new forms of co-creation for groups. In this paper, we take a step toward exploring this question by focusing on the use of LLMs in a specific type of a creative ideation process for groups: *Brainwriting* [83]. Brainwriting derives from brainstorming [54], which is a structured technique for group ideation. During a successful group brainstorming session, participants draw on each other's ideas and pre-existing knowledge to combine ideas in new ways [26]. Despite the perception that groups are more productive at brainstorming, a greater number of ideas and better quality ideas are often found in individual brainstorming [12]. This is because individuals working alone tend to consider many different potential solutions, while group members working together often consider fewer alternative solutions due to peer judgment, free riding, and production blocking [31].

Brainwriting [83] is an alternative or a complement to face-to-face group brainstorming, which aims to address these shortcomings. It begins with asking participants to write down their ideas in response to a prompt before sharing their ideas with others. After writing ideas in a parallel process, participants review others' ideas and add new ones. The number of ideas generated from Brainwriting often exceeds face-to-face brainstorming because of the more inclusive parallel process [83]. With the capability of LLMs to generate new content, several commercial products have integrated LLMs support for Brainwriting in their products (e.g. [47, 61]).

This paper explores twofold aspects of integrating LLMs into a group Brainwriting ideation process – the divergence stage of idea generation, and the convergence stage of evaluation and selection of ideas. Specifically, our investigation focuses on the following research questions:

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642414

RQ1: Does the use of an LLM during the divergence stage of collaborative group Brainwriting enhance the idea generation process and its outcome?

RQ2: How can LLMs assist to evaluate ideas during the convergence stage of a collaborative group Brainwriting process?

To explore these questions we devised a collaborative group-AI Brainwriting ideation framework, which incorporated an LLM as an enhancement into the group ideation process. We evaluated the use of the framework during the divergence stage for idea generation and the resulting solution space (RQ1) by integrating it into an advanced undergraduate course on tangible interaction design. The course seeks to expose students to novel human-AI co-creation processes within tangible interaction design [64], and to prepare them to engage with emerging LLM-based interaction design methods [63]. We conducted the evaluation with 16 students using both qualitative and quantitative methods.

To assess the potential of using LLMs in the divergence stage of group Brainwriting for idea evaluation (RQ2), we designed an LLM evaluation engine, which rates ideas based on three criteria: *Relevance* – the extent to which the idea is connected to the problem statement, *Innovation* – how original and creative the idea is, and *Insightfulness* – the extent to which the idea reflects a profound and nuanced understanding of the problem statement. We then compared the ratings produced by the LLM evaluation engine to ratings assigned by three expert and six novice evaluators.

This paper contributes to the HCI field by expanding the pedagogical frameworks and offering new AI-augmented tools for educators and novice designers, as well as by providing empirical insights into the challenges and opportunities of incorporating AI into collaborative ideation. Specific contributions include: 1) a collaborative group-AI Brainwriting ideation framework which enhances both divergent and convergent stages; 2) an LLM idea evaluation engine, which rates idea quality based on relevance, innovation, and insightfulness; 3) empirical insights into how the Brainwriting participants who are novice designers engage with and perceive the process of group-AI Brainwriting; 4) evidence that integrating the use of LLM into Brainwriting could enhance both the ideation process and its outcome; 5) evidence that LLMs can assist users in idea evaluation; 6) finally, we discuss merits and limitations of integrating LLMs into a collaborative brainwriting ideation process for both HCI education and practice.

In the following we describe the designed framework, our methods and findings. We begin with related work.

2 RELATED WORK

2.1 Structured Approaches to Ideation

Structured approaches to generating, refining, and evaluating ideas play a crucial role in creative processes across domains. Collaborative ideation approaches include techniques such as brainstorming [54], Brainwriting [83], and Six Thinking Hats [9]. Research indicates that collaborative approaches for ideation could lead to more creative solutions because when people are exposed to different perspectives, they might be inspired to explore new connections through diverse ideas [15, 28, 41, 66].

To leverage diversity of ideas, several online platforms for large-scale ideation allow users to share their ideas and to explore ideas shared by others. However, in order to expose users to those ideas that are creative and potentially inspiring, such systems need to implement methods to select and present creative and diverse ideas [66]. HCI and CSCW research have demonstrated various crowd-based and algorithmic approaches for addressing this challenge [66].

In this paper, rather than focusing on large-scale ideation, we explore ways to enhance small groups (3-4 people) ideation through the use of LLMs. Brainstorming [54] is one of the most widely adopted techniques for generating creative ideas within groups [11]. However, there are several known barriers which limit the effectiveness of group brainstorming in producing a high number of high quality creative ideas [68], including peer judgment, group thinking, free riding, and production blocking - when group members wait for their turn before sharing an idea [12]. It is also shown that group members tend to overestimate their group productivity and creativity [56].

Brainwriting [83], is an alternative or complementary method to face-to-face group brainstorming, which aims to address these shortcomings through a parallel rather than sequential process. While there are several variations of the process [29], generally, in a Brainwriting session, participants are asked to write down their ideas in response to a prompt before sharing their ideas with others. After writing ideas in a parallel process, after participants work silently on writing their ideas, participants review others' ideas and then add new ones by either individually writing additional ideas or through discussion and collaboration. The number of quality ideas generated from Brainwriting sessions often exceeds face-to-face brainstorming because the process mitigates the barriers posed from brainstorming through a more inclusive parallel process [57], however it is important to consider context and adjust the process for the specific group characteristics [29]. In recent years, online visual workspaces such as Miro [46], ConceptBoard [6] and Mural [50] offer support and template for remote and co-located Brainwriting processes. With the increasing capability of LLMs to generate new content, such services have integrated LLMs functionality as part of their products. However, there is little knowledge about the merits and limitations of integrating LLMs into ideation processes. Shin et al. led a CHI 2023 workshop to explore the integration of AI in human-human collaborative ideation [65]. Our goal is to add to the emerging body of knowledge on collaborative group-AI ideation.

2.2 Human-AI Co-Creation

Co-creation, where humans and machines work together to create new artifacts or solve a problem, is not new. The origin of computer-aided design (CAD) could be traced back to the pioneering Sketchpad system [71], which was created by Ivan Sutherland as part of his 1963 doctoral dissertation. The system, among other breakthrough innovations in computer graphics, human-computer interaction, and object-oriented programming, demonstrated that a user and a computer could "converse rapidly through the medium of line drawings" [70]. Modern CAD practices, which include generative design, have been used by designers to explore and expand their design space [20, 43].

With the emerging availability of generative AI models and tools, recent work has begun to explore how co-creation with AI models,

which are not domain-specific, could be used for interaction design and what co-creation practices with generative AI tools might look like for ideation [27, 37, 73, 78], persona creation [22] prototyping, making, and programming [2, 35, 59].

Most relevant to this case study is a small scale study conducted by Tholander and Jonsson [73] with experienced designers, which examines how large language models and generative AI can support creative design and ideation. Their findings highlight both opportunities and challenges in integrating and using GPT-3 and Dall-E by experienced designers. The work we present in this case study, extends previous work by shedding light on how students who are *novice designers*, interact with and perceive the results of ideas co-created with LLMs.

These examples of co-creation could be contextualized within emerging theories about post-humanism, post-human, and more-than-human interaction design [19, 30, 79, 80]. These theories consider alternatives to human-centered design, challenging the assumption of the "human at the center of thought and action" [80] by arguing that agency is distributed among humans, non-humans, and the environment. In response to these theories, van Dijk cautions that post-human design could obscure the important fact that non-humans agents such as AI technology are trained upon and imports traditional, humanist forms of logic and language, which in turn might taint post-human design with their humanist roots and biases [77].

2.3 Approaches for Evaluating Ideas

Dean and colleagues provide a framework for evaluating ideas [10]. The framework has four dimensions — *novelty*, *workability* (also called *feasibility*), *relevance*, and *specificity*. The framework allows a systematic evaluation of the *quality* of ideas across studies, using common definitions.

In addition to evaluating the quality of individual ideas, there are also important reasons to evaluate the *quantity* of ideas, which an ideation process generates. This is because people are more likely to find good ideas when choosing from many ideas rather than when only a few are available - in the case of ideation, more is better [34]. For example, there is evidence that having access to more AI-generated ideas improves story-writing [13]. The selection of winning ideas - those ideas that really make a difference - means that when ideas generated by an individual or a team are evaluated, the *average* quality of these ideas is less interesting - after all, as Girotra et al. argue, having a few (or even one) great idea is much better than having many average ideas [21]. Setting such a high importance on high quality ideas is especially reasonable for cases where there is a single ideation event.

While the above approaches to idea evaluation are most often associated with humans evaluating ideas, there is also an opportunity to use AI to evaluate ideas. This approach holds the promise of increased speed of idea evaluation, as well as the opportunity to develop human-AI collaborative teams where the AI could support the creative efforts of humans by providing feedback. Thus, researchers have already explored the use of AI to creativity in drawing [8], and in this work we explore using LLM to evaluate the written ideas generated by teams comprising of humans and another LLM. Domonik shows that AI evaluation could also improve human ideation by reducing evaluation apprehension - the situation where a human will withhold an idea for fear of being evaluated negatively [67].

3 COLLABORATIVE GROUP-AI BRAINWRITING FRAMEWORK DESIGN

Our investigation focuses on designing and evaluating a framework for Group-AI Brainwriting. The collaborative Group-AI design we were aiming for is one of enhancement, in which during the *divergence phase*, the group prompts the AI only *after* a first phase of Brainwriting. Paulus and Yang [57] suggested a two-phase process for the ideation process, where in the second phase participants recall ideas from the first phase, thus promoting attention and cognitive stimulus. Borrowing from their observation, we design the collaborative group-AI ideation process as a multi-phase process. In the divergence stage, group members first generate their own ideas and add them to a shared online whiteboard. Then, group members review and interact with their collective ideas while prompting an LLM for new ideas that will enhance their initial set of ideas.

In the convergence stage, group members evaluate the ideas through discussion and narrow the list of ideas to a few selected chosen ideas, which they enhance through the use of an LLM. Our investigation seeks to examine the feasibility of expanding the use of LLMs in this stage to assist group members to evaluate their ideas. We devised and evaluated a method for an LLM-based evaluation engine (using GPT-4).

Figure 1 illustrates our proposed collaborative group-AI Brainwriting framework. Following, we describe the elements of this framework.

3.1 Brainwriting Divergence stage

3.1.1 Phase 1: Brainwriting using Conceptboard. We modified the Brainwriting process [83] so that group members sit together as a team around a shared table, but write their ideas individually, in parallel, on an online whiteboard called Conceptboard [6]. The Conceptboard template we use is based on the Conceptboards remote Brainwriting template [5]. The problem statement for the Brainwriting session is written at the top of the board. Participants are instructed to each select a color on the board, set a timer for 3 minutes as a group, and use that time so that each group member write at least three ideas relevant to the problem statement and place them on the board using colored coded sticky notes. Then participants are asked to repeat this process until each group member wrote at least six ideas. Figure 2(a) shows the instructions given to participants. Figure 2(b) shows the modified Conceptboard template we used for the Brainwriting activity, populated with ideas generated by one of the student teams in our study. Each group worked on a separate Conceptboard.

3.1.2 Phase 2: Enhancing Ideas with an LLM. In here, each group is required to use an LLM (OpenAI Playground GPT-3) to generate additional ideas. Participants are encouraged to iterate on their LLM prompts and are exposed, prior to the Brainwriting session, to overview materials on prompt engineering. The generated ideas are copied and pasted into sticky notes on the board. We modified the original Brainwriting template offered by Conceptboard to

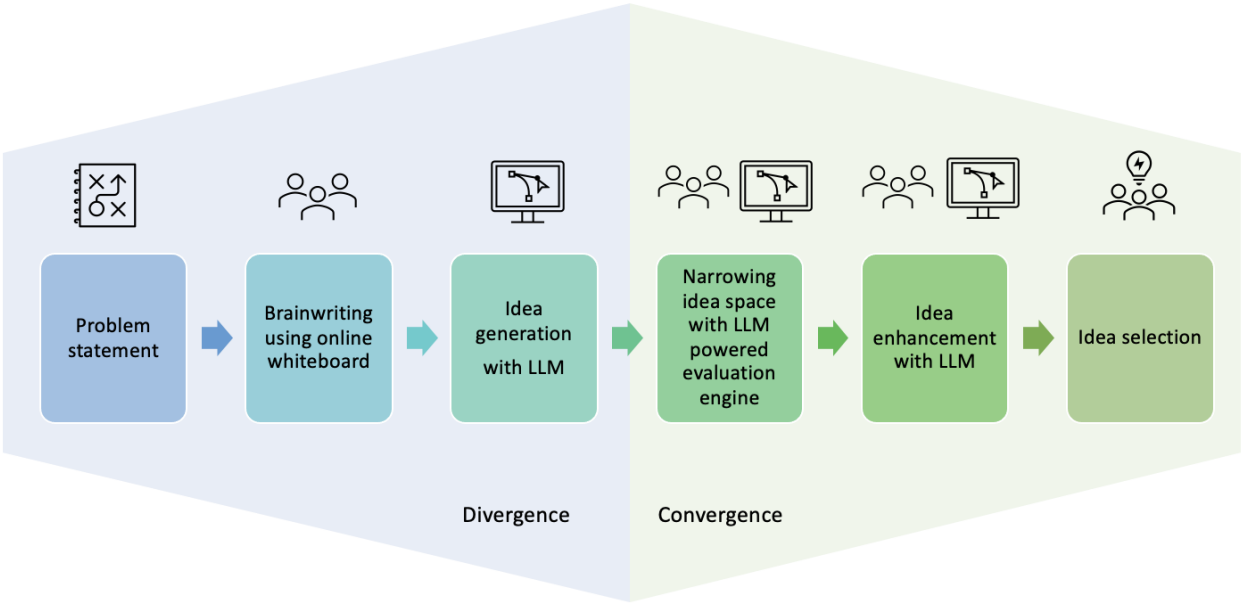


Figure 1: Collaborative Group-AI Brainwriting Process

reflect this new framework for Brainwriting with LLM for the enhancement of ideas.

At this stage the groups were instructed to review all initial ideas, discuss them, and develop together, with the help of GPT-3, new ideas that add to or build upon the existing preliminary ideas. These ideas are added to an area on the board dedicated to collaborative ideas.

For this stage of the experiment, we selected GPT-3 due to its free availability, which allowed students the opportunity to access and experiment with it in various contexts.

3.2 Brainwriting convergence stage

3.2.1 Phase 3: Selecting and developing ideas through discussion. Participants are instructed to select through discussion the best ideas and copy and paste them to a dedicated area on the board. Then they continue to develop these ideas with the help of an LLM.

3.3 Can LLMs Help with Convergence? Developing and implementing an LLM Powered Evaluation Engine

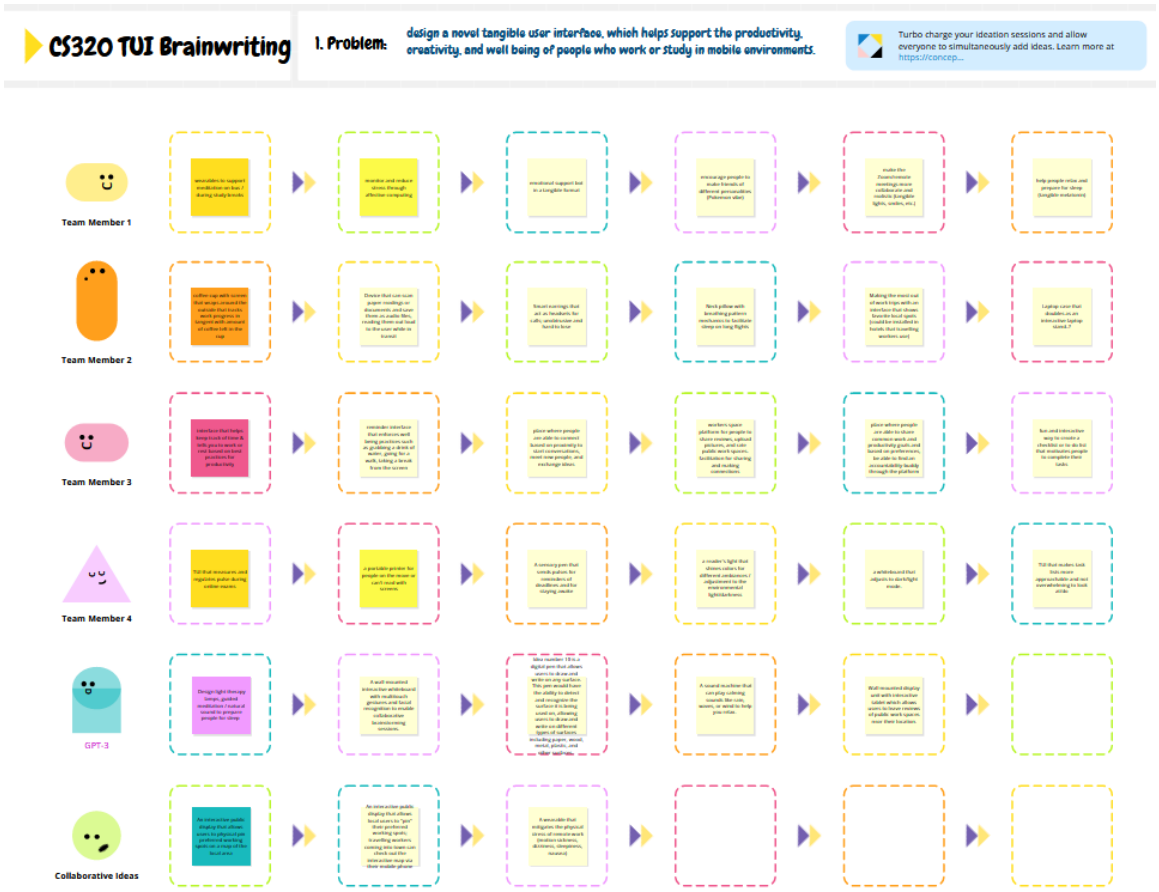
Our goal is to examine the feasibility of using LLMs to assist users in the convergence stage by highlighting the most promising ideas from the overall pool and identifying which ideas do not merit further consideration. For this stage we created an LLM evaluation engine. (The LLM-based evaluation was performed after the conclusion of the Brainwriting exercise and was not used to support the Brainwriting process.)

Our evaluation engine builds on the approach of Dean et al [10] for evaluating the quality of ideas and uses the dimensions of *novelty* (which we call *innovation*) and *relevance* to evaluate ideas. We chose not to use the dimensions of *workability* and *specificity*, because we envision this tool to be used in early stage ideation, in which neither of these dimensions play a large role; both can (and should) be addressed in subsequent stages of ideation. We also introduce an additional dimension that we call *insightfulness*, which is based on the work of Dyer et al. on the origin of innovative ventures [16]. We define an *insightful* idea as one that reflects a profound and nuanced understanding of the problem statement.

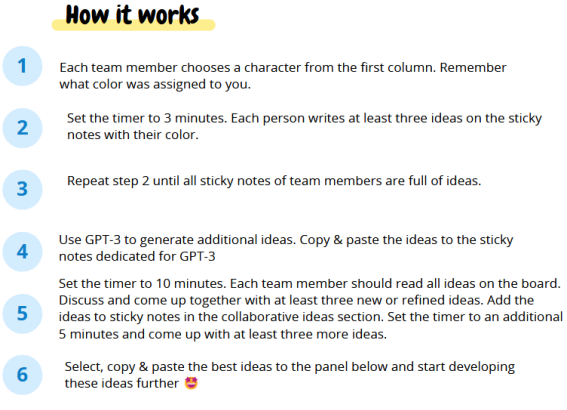
Several additional aspects need to be considered in the design of an LLM evaluation engine. First, there should be no ambiguity in the definition and interpretation of the used scales and evaluation criteria. Users would expect such an engine to communicate its evaluations and using shared definitions and agreed-upon scales. Hence, we define the following requirements:

- **Well-known Scale:** The engine would use a well-known scale, often used by humans. We chose the use of a Likert scale, with a [1 . . . 5] evaluation range [1].
- **Well-defined Criteria:** The engine would be prompted to evaluate ideas according to a well defined set of criteria, which is often used by humans to identify quality, innovative, and creative ideas. We chose to use two criteria from Dean et al.'s evaluation framework [10]: *relevance* and *innovation*. In addition, we chose a third criterion, *insightfulness*, based on Dyer et al.'s research on the origin of innovative ventures [16]. Each of these criteria required a clear definition.
- **Scale x Criteria Definition:** Each scale value for each criterion should be well defined, and in detail.

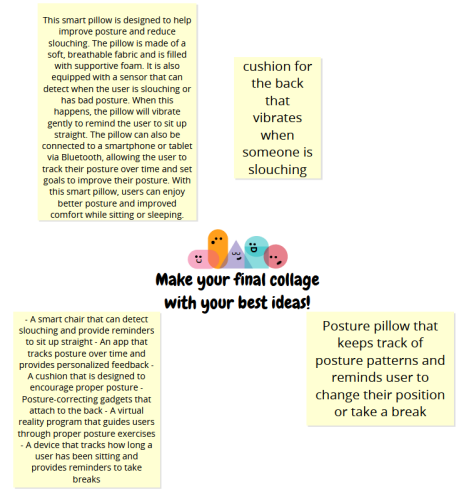
Creating a per scale value and criterion definition. We used the following procedure for developing clear, differentiated, descriptive scale value for each criterion:



(a) Central area of the Conceptboard, containing the ideas produced during the Brainwriting session



(b) Outline of the process



(c) A final set of ideas, chosen from the ideas in the central area

Figure 2: The three main areas of Conceptboards used by teams during the Brainwriting session.

(1) We first developed initial descriptive paragraphs for each criterion - *Relevance, Innovation, Insightfulness*, based on definitions in existing literature, and created descriptive anchors for each scale value.

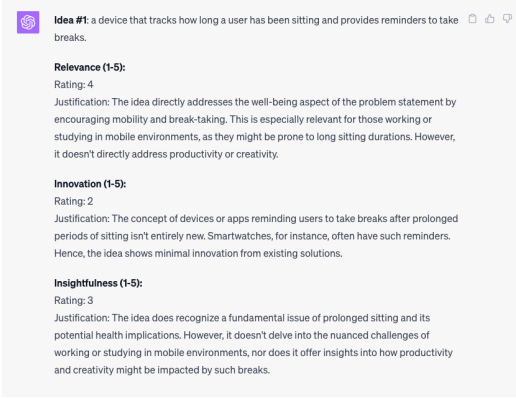


Figure 3: Idea evaluation with GPT-4 using the proposed scales for relevance, innovation, and insightfulness

(2) Three raters who are expert reviewers (researchers in HCI), working independently, rated a small sample of ideas using the initial definitions and anchors.
(3) We met with the researchers as a group to discuss their sample ratings, focusing on areas of disagreement, and came to a shared agreement on the general definition of each criterion and what each of its scale value anchors meant.
(4) Using these new definitions, we prompted GPT-4 to score a sample of ideas and to provide an explanation and justification for its assigned rating per criterion per idea. We then chose evaluation adjective and descriptive nouns from each explanation, and used these in a refined definition for a revised prompt. The definitions given in the prompt are: *Relevance*: To what extent does the idea reflect how well the idea is connected with or appropriate for the objectives, requirements, or challenges of the problem statement? *Innovation*: To what extent does the idea reflect how original and creative the idea is, breaking away from conventional or existing solutions to the problem statement? and *Insightfulness*: To what extent does the idea reflect a profound and nuanced understanding of the problem statement?
We repeated the process approximately three times for each scale until the anchors for each value were sufficiently differentiated. Supplementary Information Figure 1 shows the prompt with the explanations for the various ratings per each criterion given to the GPT-4 evaluation engine. Figure 3 depicts an idea evaluation using the GPT-4 evaluation engine.

3.3.1 Implementation. For this phase we chose GPT-4. At the time we conducted this experiment (June 2023), it has been available only for subscribers, and the researchers purchased a subscription. GPT-4 was chosen for the convergence phase over the free GPT-3 version due to its more advanced reasoning capabilities. We used the OpenAI API to write a Python program that uses the prompt to rate a set of ideas read in from a text file. The program outputs a CSV file with three ratings for each idea (for Relevance, Innovation, and Insightfulness), and a text file that contains GPT-4's justifications for those ratings. The user can indicate the number of times to repeat the process; each repetition will open a new GPT-4 context and produce a new set of ratings.

4 USER STUDY: COLLABORATIVE GROUP-AI BRAINWRITING PROCESS

We conducted a user study on the two stages of the collaborative Brainwriting process, the divergence stage and the convergence phase. In the divergence stage, we integrated the use of GPT-3 into a Brainwriting session of an advanced undergraduate course on foundations of tangible interaction [75]. During a 70 minute session students followed the Brainwriting process described above. They first generated ideas independently, then worked with their team members to co-create ideas with GPT-3, and finally, chose ideas as a team to further develop through collaboration with GPT-3.

In the convergence stage, participants evaluated the quality of the ideas they generated throughout the session in terms of *relevance, innovation, and insightfulness* and chose a small final set of ideas.

Following, we describe each part of the study in detail.

4.1 Divergence: The Collaborative Brainwriting session

In February 2023, we conducted a 70-minutes Brainwriting session with 16 college students (0 men, ages 18-23) who were enrolled in an advanced undergraduate course on tangible interaction design. Considering the challenges interaction designers face when working with AI as a design material [14, 32, 69, 81, 84], this course aims to integrate co-creation and critical engagement with generative AI into its learning goals. Integrating the AI-augmented Brainwriting session into the course activities was thereby aligned with the course learning goals, among these: LG1) Apply a collaborative iterative process, which includes co-creation with AI and ML models for designing innovative tangible and embodied interfaces; LG2) Assess the capabilities and limits of prevalent AI technologies within the context of tangible interaction design; LG3) Implement a functional prototypes of a novel tangible or embodied interface using various technologies for data processing, sensing, and actuation. Develop AI intuition through experimental and creative exploration of AI technology for prototyping. The complete list of learning goals and course materials are available in the course website [link will be added in the camera ready version].

The students were divided into 5 project teams of 3-4 students each. The goal for the session was for students to start developing project ideas for a semester-long group project, which required them to *"design a novel tangible user interface, which helps support the productivity, creativity, and well-being of people who work or study in mobile environments."* Prior to the in-class Brainwriting session students were asked to read about Brainwriting [83] and about ChatGPT [58, 60].

After writing down their individual ideas on their team ConceptBoard, students used the OpenAI Playground GPT-3 to generate additional ideas using repetitive prompts. We reminded students that ideating with GPT-3 might require multiple interactions in which they will need to refine their prompts and provided them with some examples for prompts used to generate similar tangible user interfaces (TUI) ideas. After adding the GPT-3 ideas to the board, we asked them to review, discuss, select, copy & paste the

best ideas to a side panel and start developing these ideas further with the help of GPT-3.

Table 1 shows the number of ideas generated by each team. The average word count of each Human-Generated idea is 16.5; the average word count of each GPT-3-Generated idea is 20.9. In addition to submitting a link to their Conceptboard, students were asked to submit all their GPT-3 prompts.

Table 1: The number of ideas created per team: Human-Generated, GPT-3-Generated, Collaboratively-Generated, and total.

	Human	GPT-3	Collaborative	Total # of ideas
Team 1	20	4	2	26
Team 2	18	11	11	40
Team 3	17	2	0	19
Team 4	24	6	6	36
Team 5	18	6	3	27

4.2 Convergence: Ideas Evaluation and Selection

At the end of the session, the students were asked to rate the ideas: their own, GPT-3's and the collaborative ideas, as a means to narrow down the idea pool and engage in a selection process. The ideas were rated on a Likert scale along the three chosen evaluation criteria of relevance, innovation and insightfulness. Table 2 shows the results of their self-ratings evaluation. The results show that students assign high levels of relevance, innovation, and insightfulness with mean scores of 4.75, 4.45, and 4.45, respectively to the ideas generated in their session. The distribution of scores exhibited a notable skewness, with 60% of the questions attaining the maximum possible rating of 5 out of 5.

Table 2: Average Self Ratings and Standard Deviations for Each Evaluation Criterion

	Relevance		Innovation		Insightful	
Generated by	Avg	Std	Avg	Std	Avg	Std
Human	4.81	0.40	4.31	0.70	4.37	0.61
GPT-3	4.56	0.51	4.25	0.68	4.18	0.65
Collab	4.87	0.34	4.81	0.40	4.81	0.40

After the session, each team chose an idea for their semester-long project. Table 3 depicts the final ideas, and the source of the idea (human generated, LLM-generated, or combined).

Finally, we asked students about their experience Brainwriting with GPT-3 both immediately after the session, as well as again at the end of the semester.

5 FRAMEWORK EVALUATION

The evaluation of the proposed collaborative group-AI Brainwriting framework consists of two parts. In the first, we explore through the use of qualitative and quantitative methods whether the use of LLMs in the divergence stage of group Brainwriting enhances the ideation process and its outcome (RQ1). To evaluate the *quality* of the ideas, in addition to the participating students' self evaluation and to the ratings generated by the GPT-4 evaluation engine, three independent expert reviewers (HCI researchers) and six novice designers (HCI students) rated the quality of ideas on the same dimensions. Since the quality of ideas selected in the converge stage is impacted by the divergence of ideas generated [7], we evaluated divergence by examining the semantic distribution of ideas generated by humans and by GPT-3. We also identify the unique terms used in the different solution spaces. We then explore, in the second part of the evaluation, how LLMs can be used to assist in idea evaluation during the convergence stage (RQ2).

Here we describe the data and methods used in the evaluation of the proposed framework, followed by results organized by research question.

5.1 Data and Methods

We collected the following data: ideas generated by each team during the Brainwriting session; prompts used to interact with GPT-3; student responses to reflection questions; and novice designer ratings, expert ratings, and GPT-4 ratings.

We recruited 6 novice designers (students who completed an HCI course and were not enrolled in the same course in which we conducted the user study), as well as four expert reviewers who are active HCI researchers. Both novice and expert reviewers were asked to rate the set of ideas using the same three criteria definitions and scale value anchors given to the GPT-4 evaluation engine. The ideas given to the reviewers were arranged in a random order and there was no identifying information regarding the source of the idea (human or GPT-3). One expert reviewer provided evaluations for only a subset of ideas produced by student groups. In this document we report on data from the three expert reviewers who evaluated all of the ideas produced by students.

We used thematic analysis [3] to analyze the prompts used to interact with GPT-3 and the student reflection open responses. We first identified common keywords and tags among the responses, then aggregated these in order to extract broad themes and categories.

To examine the divergence of the ideas dataset (aggregated content of all 5 Conceptboards) we used the following methods and tools. We first used the NLP toolkit spaCy to extract nouns and adjectives from the dataset. Also, we used spaCy and Gensim for topic modeling. We further use the Domain-based Latent Personal Analysis (LPA) method [48]. LPA identifies the terms that most separate a document from a corpus. Using an Information-Theory approach, it creates a *signature* for each document, comprised of the terms that differ most in frequency in the document from their frequency in the corpus. These terms are corpus popular terms that are rare or missing in the document, and corpus rare terms that are frequent in the document. To create the signatures, each document is converted to a normalized term frequency vector, and the vectors are aggregated to create a corpus vector representation. LPA creates the signature per document by computing the symmetric per-element Kullback–Leibler Divergence (KLD) [39], also called relative entropy, between each document and the corpus. The relative entropy from distribution C to distribution D over sample

space X is:

$$KLD(C||D) = -\sum_{x \in X} D(x) \log_2 \left(\frac{C(x)}{D(x)} \right) \quad (1)$$

LPA uses the symmetric KLD ($KLD(D||C) + KLD(C||D)$) and pad document vectors with ϵ-values for missing corpus terms. The corpus contains for each term that appeared in at least one of the documents its relative frequency. Here, as there are only two documents, one containing terms used by humans ($V_{\mathcal{H}}$) and the other the terms used by GPT-3 ($V_{\mathcal{G}}$), we perform the following. Each vector is expanded to contain all the terms in the $\cup(V_{\mathcal{H}}, V_{\mathcal{G}})$ set, and missing terms are denoted as having zero frequency. The weight of each corpus term is computed as the average between the normalized term frequency in $V_{\mathcal{H}}$ and $V_{\mathcal{G}}$. Using Equation 1 LPA finds for each document the terms that contributed most to the Relative Entropy of the terms that contributed most to the divergence of each of the normalized frequency vectors $V_{\mathcal{H}}, V_{\mathcal{G}}$ from the corpus. Term weights are assigned according to this contribution, with a corresponding sign. A positive sign indicates a rare corpus term that is overused in the document, and a negative sign indicates a corpus popular term that is underused or not used at all (missing) at the document. The set of terms with the highest absolute weight comprises the document's signature, each with its corresponding sign.

Finally, statistical analysis was conducted using SPSS and Python. SPSS was used for hypothesis testing of agreement. GPT-4 was used for Semantic Analysis, and Python was used for descriptive analysis and LPA.

5.2 Results RQ1: Does the use of an LLM during the divergence stage of collaborative group Brainwriting enhance the idea generation process and its outcome?

To answer RQ1 we examined both (a) student perceptions about the ideation process and (b) the outcome of the ideation process - the set of selected project ideas and their origin in terms of Human- and/or GPT-3-Generated ideas. We then examined (c) the divergence of ideas through semantic analysis, and (d) the solution space explored with and without GPT-3 using LPA. Finally, we analyzed the (e) prompts used by students to interact with GPT-3. In the following, we describe the results.

5.2.1 Students' Reflections. Since the user study was conducted within an educational setting, our evaluation of students' perceptions of this Group-AI Brainwriting framework also involved assessing their learning and critical engagement with AI. In a separate paper [currently under review for a different conference], we contextualized the use of this framework within a broader integration of generative AI into a tangible interaction course, and discussed students' reflections and learning. Here, we summarize student perceptions of the Group-AI Brainwriting process. Specifically, we analyze student responses to a question we asked immediately after the ideation session (Q1): *"In what ways did using GPT-3 contribute to or hinder the ideation session?"* We also analyze their response to a question asked at the end of the semester (Q2): *"Thinking back to your original ideation with GPT-3: to what extent do you feel like your collaboration with text-generative AI influenced the direction of your project?"*

Q1: In what ways did using GPT-3 contribute to or hinder the ideation session?

All students responded to this question (n=16). Overall, we identified seven recurring themes: 3 themes describe positive contributions of GPT-3 to the ideation process, and 4 themes describe shortcomings of GPT-3. 50% of students (8 out of 16) highlighted that GPT-3 offered them a unique or expanded viewpoint on the issue and its possible solutions. For example, one student shared that GPT-3 provided "ideas we had not offered or thought to offer on our own [...] we were focused originally on one niche interpretation of the problem, and ultimately [...] we got a more diverse set of possible products." 44% of students (7 out of 16) felt that GPT-3 significantly assisted them in generating ideas, in the words of one student: "adding in new ideas that we had not considered previously." Some students pointed out that their team(s) selected an idea for their concluding project that was initially suggested by GPT-3, one saying that "the model ultimately contributed the base idea we expanded upon with our own ideas to create the project pitch." A smaller proportion of students (2 out of 16) mentioned that GPT-3 assisted them in articulating and communicating their own ideas. For example, one student wrote that "[GPT-3] helped us communicate our ideas better since it would reword our prompt."

31% of students (5 out of 16) pointed out that GPT-3 tends to be redundant and lacked creativity. For example, one student mentioned that "it didn't come up with anything we didn't." Another student described their experience as though the AI was experiencing a "creative block," they received similar results no matter how they reworded their prompt. 25% of students (4 out of 16) reported challenges with crafting prompts and had to employ a trial and error approach to formulate prompts that produced high-quality responses from GPT-3. For example, one student shared that "there was a steep learning curve in understanding how to correctly prompt the model that hindered initial ideation." One student expressed frustrations with the experience, describing how "it was pretty hard to get GPT-3 to output things the way we wanted it unless we used very specific language," but added that their use of the tool was still helpful "as a way to kickstart our ideation and take the momentum into our own creativity." Some students (2 out of 16) highlighted issues with the output being unrelated to the prompt or noted a lack of 'common sense' in understanding their request. One student shared their frustration with how GPT-3 would continually output "ideas that already existed," such as Apple Watches.

Q2: Thinking back to your original ideation with GPT-3: to what extent do you feel like your collaboration with text-generative AI influenced the direction of your project?

In response to this question, which was asked at the end of the semester, 50% of the students (8 out of 16) indicated that using GPT-3 contributed to reshaping and enhancing their project by elaborating on their concepts, proposing new characteristics, and tackling particular challenges. In the words of one student: "The AI helped us reframe and refine our problem statements and questions so that may have been beneficial since we had to learn how to communicate with the AI. That alone made us more aware of the direction of our project since we had to refine the question on the spot in order for us to work with the AI. The AI also worked as a

jumping off point for the team members to think of more organic, creative ideas." Another student shared "I think that AI gave us many ideas that we could incorporate into [our project]. I think that collaboration with AI didn't necessarily generate an idea. However, with our specific idea in mind, we were able to utilize AI to think of more creative features."

25% (4 out of 16) said that GPT-3 had an impact on the direction of their project. For example, one student wrote "It influenced the direction somewhat greatly - we already had the idea to make something that users could gather around, and use /after/ they had relocated while working remotely [...] but GPT-3 gave us the idea to make the [project] more community-oriented." Another student shared "ChatGPT helped expand our brainstorming process and brought us ideas we hadn't thought of before, so we combined many into one as we decided on our project idea." A few students described GPT-3 as a partner, assisting with particular tasks: "GPT-3 helped us with more specific information such as "how to alleviate motion sickness" and "what heartbeat threshold indicated an onset of motion sickness" that we did not inherently know. It was therefore helpful as a fourth teammate, but it could not replace any of us. So a nice companion, but not a substitute."

5.2.2 Ideation outcomes. The outcome of the human-LLM ideation process was a set of chosen ideas - each team chose one idea to explore in a semester long project. Table 3 shows the chosen idea of each team, and describes the conception of each idea in terms of its human and/or GPT-3 origin. Overall, 3 out 5 chosen ideas were developed through merging a Human-Generated idea and a GPT-3-Generated idea. One idea was developed through merging multiple Human-Generated ideas and multiple GPT-3-Generated ideas. Finally, one out of the 5 ideas is based solely on a GPT-3-Generated idea.

5.2.3 Exploring the Human and LLM solution spaces. To explore the divergence of ideas and the solution space explored with and without LLMs, we evaluated the semantic distribution of ideas generated by humans and by GPT-3, and the terms used in the different solution spaces using LPA.

Evaluating the semantics of different idea spaces allows us to explore potential conceptual differences between the human and AI idea spaces. If these concept spaces, as determined by our methods, show substantial overlap, it would suggest that in this experiment, the AI did not significantly augment the human creative thought process from a conceptual aspect. For the evaluation, we compared a semantic clustering over the terms used in these spaces, and then evaluated the differences in the terminology. A difference in terminology can be semantic or more substantial. A substantial difference, characterized by overused, underused, or entirely absent terms in a solution space, offers deeper insights into the variances that may exist between human and AI-generated ideas.

Semantic clustering analysis. To discuss the semantic clustering analysis, the following terminology is used. The set of Human-Generated ideas as $\mathcal{H}$, and the set of GPT-3-Generated idea as $\mathcal{G}$. The semantic analysis was done by generating semantic clusters of the ideas in both sets, $\mathcal{H}$ and $\mathcal{G}$. The semantic analysis of $\mathcal{H}$ yielded 20 clusters, and of $\mathcal{G}$ 21 clusters. There were 12 similar clusters that contained shared terms. For example, in both sets the cluster Digital Devices & Hardware contained the terms <*computer, monitor, laptop, smartphone (and phone), tablet*>, and the cluster Health & Wellness contained the terms <*sleep, meditation, stress, nausea, heartbeat, pulse*>. The semantic clustering of $\mathcal{H}$ contained the following unique clusters and terms: Vehicle-related terms <*bus, commute, train*>, Personal clothing <*jacket, sweater*>, Food and beverages <*dining, water*>, Learning & information <*study, academy, library*>, and Games & entertainment <*Pokemon, music, leisure*>. The semantic clustering of $\mathcal{G}$ contained Screen and display elements <*background, settings*>, Interactivity and controls <*buttons, dials, gestures*>, Specific measurements <*cm, diameter, intensity*>, Visual & design elements <*shapes, signs*>, and Specific work-related terms <*brainstorming, distractions*>. The full list of clusters and their corresponding terms can be found in the Supplementary Information.

Overall, while many of the semantic clusters were similar, the differences seem to relate to the level of detailing of the concepts. The concepts found only in $\mathcal{H}$ tended to be more abstract or alluded to objects in a generalized manner, while concepts found in $\mathcal{G}$ were more concrete or pertained to specific details of objects or their description, such as their measurements.

LPA of the terminology used in the two solution spaces. Here, we examine the differences in noun terms used within the solution spaces of LLMs and human-generated ideas. Variations in noun term usage can reveal conceptual or thematic differences, highlight the level of detail and depth in the ideas, indicate their specificity and breadth, and may also suggest the context to which the idea pertains. LPA identifies the main differences between the two corresponding noun terms distributions.

LPA analysis of the terms used either by humans or GPT-3 reveals a difference. Figure 4 shows the results of the analysis. The ten most prevalent terms used in ideas by either humans or GPT-3 (normalized to account for the different number of ideas in each group) were *user, device, light, people, sound, surface, task, wrist, pillow, day*, depicted in Figure 4a. However, there were some notable differences, as can be seen from GPT-3's LPA signature, depicted in Figure 4b. For example, while ideas created by humans referred to *people*, GPT-3 kept using the term *users*. The term *device* was prevalent in GPT-3's ideas, while hardly used by humans. Other GPT-3's prevalent terms were *surface, light, posture, wrist* that were hardly used by humans. On the other hand, GPT-3 did not refer to terms that were commonly used in human ideas, such as *wearable, screen, work, time, space, interface, day, app*.

5.2.4 Prompt analysis. To get further insight into the differences between Human-Generated and GPT-3-Generated ideas, we analyzed the prompts used by students to generate new ideas and iterate on existing ones, and identified a few distinct approaches. Typically, students used one of two approaches to initiate their interaction with GPT-3: 1) broad-area prompts, or 2) solution-specific prompts.

Broad-area prompts involved giving GPT-3 an open-ended request for ideas related to the problem statement. For example, one team began their interaction with the prompt, "Tell me a list of ideas for tangible interfaces that support productivity and creativity that doesn't exist yet ". Solution-specific prompts entailed asking for a solution for a concrete problem. For example, "Tell me ways to stabilize a portable desk when on a bus";

Table 3: Brainwriting outcomes - a set of chosen ideas. For each team we describe the idea chosen for a project proposal and the enhancement type which contributed to its development.

	Chosen idea	Enhancement	Description
Team 1	An interactive public display that allows local users to "pin" their preferred working spots; travelling workers coming into town can check out the interactive map via their mobile phone	Combined human & LLM	Inspired by the combination of a Human-Generated idea, a platform for rating work spaces, with a GPT-3 Generated idea, an interactive public display
Team 2	Posture pillow that keeps track of posture patterns and reminds user to change their position or take a break	LLM	Inspired by a GPT-3-Generated idea for a smart pillow that can detect posture
Team 3	Portable desk for commuter students with stability and motion sickness-alleviating features and a built-in wifi hotspot	Combined human & LLM	Not submitted with original workshop idea set, but submitted with project proposal as a combination of a Human-Generated idea ("portable desk that is stable on bumpy rides") and a GPT-3-Generated idea ("installing a wireless router or access point inside a portable desk")
Team 4	A plushie/stress ball keychain that users can hold onto; releases aromatherapy and also communicates with user holding another one to either feel their heartbeat or the same squeezing sensation	Combined human & LLM	Inspired by combining a number of Human-Generated and GPT-3-Generated ideas having to do with aromatherapy for stress and paired devices that transmit the users' pulse. Unlike the other teams, this team combined several ideas together
Team 5	Sleeping eye mask that changes temperature based on where you are in your journey and vibrates to wake you before your stop	Combined human & LLM	Inspired by the combination of a Human-Generated idea, a wearable to notify the user when their public transport stop is near, with a GPT-3-Generated idea, a temperature-controlled sleep mask

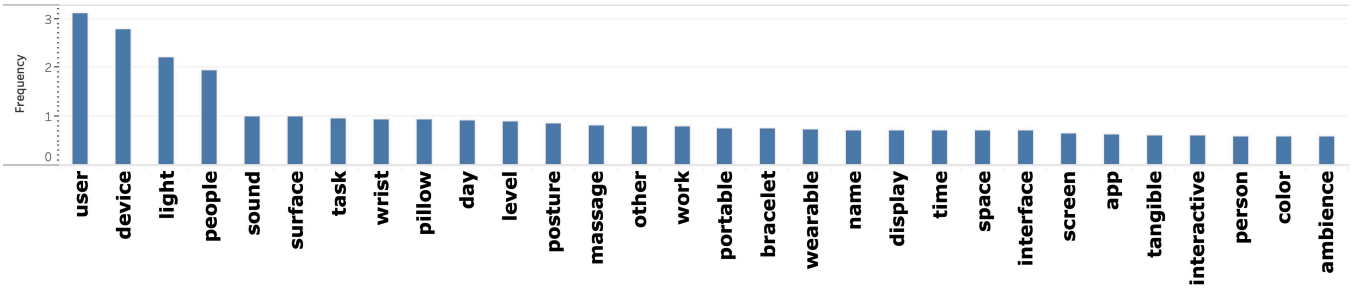
When students decided to focus on a particular idea, they applied two different approaches to expand on their idea: 1) usage-focused follow-up prompts, and 2) detail-focused follow-up prompts. A usage-focused prompt asked GPT-3 to expand on the ways and context users would use their proposed solution. For example, one team asked "How can this device be utilized without Wendy having to change the settings?" A detail-focused prompt, on the other hand, asked GPT-3 to expand on the features and capabilities of a specific idea. For example, "Tell me a list of functionalities that a smart light can do to make you more productive and creative."

Student teams combined these approaches during the ideation session.

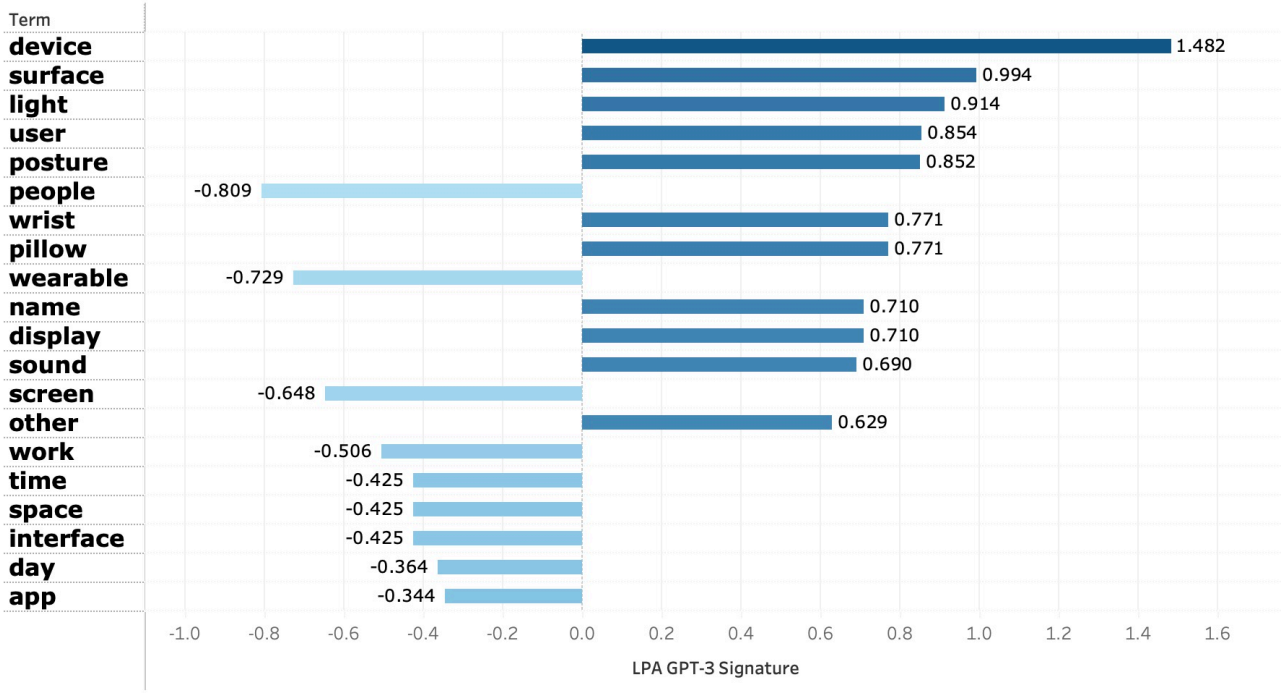
5.2.5 Summary of findings for RQ1. After the session, 50% of students perceived GPT-3 as helpful because it provided a unique or expanded perspective on the problem statement and its possible solutions. 44% shared that it significantly assisted them in generating new ideas. At the end of the semester, 50% of the students mentioned that GPT-3 contributed to reshaping and enhancing their project by elaborating on their concepts, proposing new characteristics, and tackling particular challenges. 31% of students pointed out that GPT-3 tends to be redundant and lacked creativity.

The ideas chosen by each group for their final project were mostly created by combining an idea generated by team members and an idea suggested or enhanced by the LLM. In one case (Team 2), the chosen idea was directly inspired by an idea generated by GPT-3.

Semantic clustering analysis of Human- and GPT-3-Generated ideas indicates that humans tended to allude to abstract concepts and refer to objects in a general way, while the ideas generated by GPT-3 were more concrete. The solution space, denoted by the different vocabulary used in ideas generated by humans and GPT-3, is consistent with these findings. For example, the term "device" appears almost exclusively in GPT-3-Generated ideas, which often



(a) The top used terms in the joint vocabulary created by LPA for the terms used by humans and GPT-3 in their ideas.



(b) GPT-3 LPA signature, denoting terms that are overused in ideas compared to humans, and terms that are underused or missing, when compared with humans. The corresponding weight of each term denotes the difference in the percentage of times it is used in GPT-3's ideas compared to the average usage across the ideas generated by either humans or GPT-3.

Figure 4: Identifying biases in LLM-generated ideas. (a) introduces the top terms used in all ideas generated either by humans or by GPT-3, as calculated using the Latent Personal Analysis (LPA) method. (b) depicts GPT-3's LPA signature, denoting its unique use of terms when compared to the shared vocabulary, either underused or overused.

also reference their "users". In Human-Generated ideas, the reference is to "people", and the term "wearable" appears only in human ideas. Humans tend also to refer more to "space" and "time", while GPT-3 referred more to "surface" and "light".

The prompt analysis reveals that students combined approaches when interacting with GPT-3, typically starting with a broad request for ideas, then requesting solutions for a concrete problem, or asking for additional details regrading the usage, features, and/or capabilities of a specific idea. These results explain, to some extent, the higher level of details we found in GPT-3-Generated ideas.

5.3 RQ2: How can LLMs assist to evaluate ideas during the convergence stage of a collaborative group Brainwriting process?

We assess here the feasibility of using an LLM to assist in idea evaluation in the convergence phase. These idea evaluations were not part of the User Study and were conducted after the student deadline for choosing the final ideas. To evaluate how LLMs can help in the convergence phase, in which all ideas are evaluated and a few are selected, we assess here: (a) whether LLMs' evaluations are consistent, and (b) how they compare with evaluations made

by experts and novices. Our goal here is to assess whether LLMs can be used to filter out ideas reliably.

All ideas created during the Brainwriting process: Human-Generated, GPT-3-Generated, and Collaboratively-Generated, were evaluated by 3 Experts, 6 Novices, and the GPT-4 evaluation engine. All evaluations used the same 1 to 5 Likert Scale for Relevance, Innovation, and Insightfulness. Both Novice and Expert reviewers were given the same criteria definition and scale value anchors given to the GPT-4 evaluation engine. The ideas given to the reviewers were arranged in a random order and there was no identifying information regarding the source of the idea (human or GPT-3). The GPT-4 engine was prompted to repeat each evaluation 30 times (29 rounds were completed successfully), each evaluation conducted in a new context.

5.3.1 Consistency of the GPT-4 evaluation engine. First, we assess the internal consistency of the 29 GPT-4 evaluations for the ideas on the three criteria of Relevance, Innovation, and insightfulness. To evaluate consistency we treat the evaluations as questionnaire items and analyze them with Fleiss' Kappa coefficients to evaluate rater agreement. Our analysis shows a moderate level of consistency in GPT-4's performance, with all Fleiss' Kappa values surpassing the 0.4 threshold. The specific Fleiss' Kappa values for the different criteria were the following. Relevance: 0.42, Innovation: 0.40, and Insightfulness: 0.49. Thus, GPT-4 evaluations can be seen as consistent across the three criteria.

5.3.2 Comparative Analysis of GPT-4's Evaluations Against Novice and Expert Human Evaluators. We compare the ratings given GPT-4 to those given by novices and experts to the 148 ideas generated by either humans, GPT-3, or in collaboration. The ratings were given to each idea for each of the three criteria: Relevance, Innovation, and Insightfulness. To compare the GPT-4 evaluations to human raters, we conducted the following steps: (a) compared the given rating distributions, (b) compared evaluations for the top and bottom ideas as ranked by the experts' ratings; (c) computed the Pearson correlation between GPT-4 ranking of ideas and the experts' ranking; (d) compared the ratings given by GPT-4, novices, and experts, across the three criteria, to the ideas that were chosen by the teams as their final ideas.

Unlike GPT-4, Expert and Novice evaluators had diverging opinions and medium to low internal consistency across the three criteria. A Shapiro-Wilk Test on the raw rating distribution of Experts evaluations found that the null hypothesis of a normal distribution is rejected with a p-value « 0.001 for the ratings of all three criteria: Relevance, Innovation, and Insightfulness. Similarly, the Shapiro-Wilk Test on the raw rating distribution of Novice evaluations found that the null hypothesis of a normal distribution is rejected with a p-value of « 0.01 for all three criteria.

(a) First, we compare the ratings distributions across the evaluator groups. Figure 5 depicts the distribution of ratings on a Likert scale of 1 to 5 given by Experts (lower panel), Novices (middle panel), and GPT-4 (upper panel) for the 148 ideas across the three criteria. For each idea and criterion, the rating was calculated as the average of the ratings given by the corresponding rater group, either Experts, Novices, or GPT-4, to that idea. The ratings distributions demonstrate that the Experts were more critical than the Novices and that GPT-4 gives relatively high ratings to ideas. GPT-4

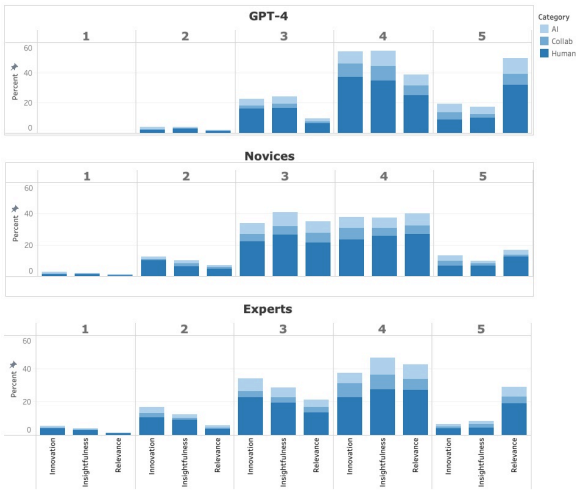


Figure 5: The Distribution of ratings on a 1 to 5 Likert scale given to ideas generated in the Brainwriting process. Ideas were generated by either humans, GPT-3, or as a collaboration. Every idea was assessed based on three criteria: its relevance, depth of insight, and level of innovation. All 148 ideas were rated by Experts, Novices, and the GPT-4 rating engine. The lower panel depicts the distribution of ratings given by Experts to ideas in each of the criteria. The middle panel depicts ratings given by Novices, and the upper panel the rates given by GPT-4.

gave much more ratings of 5 than novices and experts and much less ratings of 2 and 1. Specifically, it gave a lower rating of 1 to only one idea, for its Insightfulness. GPT-4 gave an average rating of 4.19 for relevance, 3.72 for innovation, and 3.68 for insightfulness.

Clearly, there is no agreement between either of the groups, and hence also not with that of GPT-4. We then continue to examine the similarity in ranking of ideas, and the ratings given to the final ideas as chosen by the teams.

(b) We created a ranking of the ideas for each rater group, Experts, Novices, and GPT-4. The ranking of the ideas was computed as follows. For each rater group, the rating of an idea by that rater group was computed by averaging the ratings given by the group members for each of the criteria and then by summing these values. For example, in the case of the Expert rater group, the average rating given by the three experts to each of the criteria Relevance, Innovation, and Insightfulness was computed, and the idea's final rating was computed as the sum of these three average values. per each criteria and summing it. Thus, an idea with an Experts average rating of 4 for relevance, 2.75 for innovation, and 2.375 for insightfulness received an aggregated rating of 9.125, and was ranked 24 out of 148 ideas.

From the Expert ranked idea list, we chose the four highest and lowest-ranked ideas and compared their ranking to their ranking on the GPT-4 ranked idea list. On the Expert ranking list, the top four received ratings of 13, 12.5, 12.5, 12.5. The lowest-ranking ideas received ratings of 5.75, 5.33, 5.25, 5.25. Of the four ideas ranked highest by the Experts, one was also in the second place on GPT-4

list, and the rest were in the top half of the list. Out of the four ideas ranked lowest by the experts, three were in the bottom 6 places on GPT-4 ranked list. The fourth ranked lowest idea by the Experts, was ranked in the middle of the list by GPT-4.

Comparing the top and bottom four between experts and novices, we found that out of the experts' four top rated ideas two were also rated at the top by novices. The two other ideas were not at the top of the novices' list. There was no agreement at the bottom part of the ranked list, as all ideas that were rated lowest by the experts appeared in the lower quarter, however not in the bottom of the novices' list. When comparing the novices' and GPT-4's top and bottom elements, we find that there is a high agreement.

(c) To quantify the relationship between the rankings provided by different groups, we computed Pearson correlation coefficients. The comparison yielded a coefficient of 0.556 between expert and GPT-4 ratings, 0.547 between novice and GPT-4 ratings, and 0.602 between expert and novice ratings. These results indicate a moderate positive linear relationship among the three ranked lists.

Thus, we can conclude that overall, GPT-4's ranking of ideas is generally in agreement with the Experts' and novices' rankings.

(d) Lastly, we examine the evaluations given by the GPT-4 evaluation engine to the ideas that were ultimately chosen by student teams, and compare these to the expert and novices corresponding ratings. Table 4 summarizes the the ratings of the final ideas chosen by teams. For the majority of instances, all rater groups, namely experts, novices, and the GPT-4 evaluation engine, assigned higher ratings to the final selected ideas compared to the average rating they assigned to all ideas.

We have shown that both the expert and the novice raters had diverging opinions on many of the ideas. While the majority of the final project ideas received a higher rating than the average idea from the experts (but team 5's idea), their evaluations of the ideas along the three criteria differ substantially, as reflected in the relatively high standard deviation values. Similar disagreement, although to a lesser degree, exists also among the novice raters' evaluations.

Among the evaluations of the teams chosen ideas, the largest disagreement between the rater groups exists between the experts and GPT-4 for team 5 chosen idea. The largest difference exists for the evaluation of the Innovation of the idea, receiving a low average score of 2.00 by the experts, compared with an average rating of 4.93 from GPT-4. Interestingly, this idea received the highest rating for Innovation from the novice raters among the chosen ideas.

Overall, our analysis shows that GPT-4 evaluation engine did not rate below the average the ideas that were chosen as final by the ideas.

5.3.3 Summary of findings for RQ2. The GPT-4 evaluation engine gave high ratings to all of the ideas that were ultimately chosen by student teams as can be seen in Table 4. We further observed a robust level of internal consistency among the ratings generated by the GPT-4 engine, as evidenced by elevated values of Fleiss' Kappa exceeding 0.4 across all three criteria: Relevance, Innovation, Insightfulness. Unlike GPT-4, Expert and Novice evaluators had diverging opinions, and medium to low internal consistency across the three criteria. The distributions of evaluations reveal that Experts were more critical than Novices, and that GPT-4 gives relatively high ratings to ideas.

We evaluated the alignment of idea rankings between experts, novices, and GPT-4. A notable correlation was observed, especially between the highest and lowest-rated ideas. Top ideas as rated by experts were generally also favored by GPT-4, with a similar pattern evident in the novice evaluations. The Pearson correlation coefficients – 0.556 between experts and GPT-4, 0.547 between novices and GPT-4, and 0.602 between experts and novices – suggested a moderate positive linear relationship among the three groups' rankings. This consistency across human and AI evaluations highlights GPT-4's potential as a viable tool for preliminary idea filtering, aligning closely with human judgment in identifying high-quality ideas.

The fact that none of the chosen ideas received low ratings by GPT-4 is encouraging - it means that, if GPT-4 had been used to provide feedback for teams during the ideation process, it would not have filtered out ideas that were considered to be good by the teams. At the same time, it also appears that, had GPT-4 been used to provide feedback during the ideation process, teams could have safely discarded ideas that were rated low by GPT-4. After all, none of the ideas that were rated low by GPT-4 were ultimately chosen. (Note that we used GPT-4 to evaluate ideas only after the ideation sessions were completed, so these evaluations were not available to teams.)

6 DISCUSSION

In this paper we propose a framework for collaborative group-AI Brainwriting and study two dimensions of such integration. First, we study the use of an LLM for enhancing the idea generation process. Second, we explore the use of an LLM for evaluating ideas during the convergence phase, in which three criteria of the ideas are evaluated: their relevance to the problem statement, the originality and creativity of the idea, i.e., how innovative it is, and the extent to which the idea reflects a profound and nuanced understanding of the problem statement, which we refer to as the insightfulness of the idea. We conduct a user study that uses the framework for an idea generation process as part of a college-level interaction design course, and conduct a set of evaluations of the process, its outcomes, and the potential use of an LLM for the evaluation process.

Here we discuss our findings, focusing on addressing the two research questions we introduced in the introduction. We then discuss implications for HCI education and practice.

6.1 Discussion of results for RQ1: Does the use of an LLM during the divergence stage of collaborative group Brainwriting enhance the idea generation process and its outcome?

In their reflections, 50% of the students found the use of GPT-3 helpful in providing unique or expanded perspective on the problem statement and its possible solutions. Findings from our semantic and LPA analyses of the idea space, indicate that indeed GPT-3 contributed both ideas that were somewhat different from those generated by humans, as well as included more technical and usage details. These findings indicate that integrating an LLM into the Brainwriting ideation process could provide support for both

Table 4: Comparison of the evaluations of experts, novices, and GPT-4 for the chosen final project ideas of each team, as described in Table 3

Rater	Criterion		Team 1	Team 2	Team 3	Team 4	Team 5	Average over all ideas
Expert	Relevance	Avg	3.75	4.25	4.00	3.67	3.00	3.57
		Stdev	0.96	0.50	1.00	0.58	1.83	1.10
	Innovation	Avg	3.00	3.25	3.33	3.00	2.00	2.79
		Stdev	1.41	0.96	2.08	1.00	0.82	1.10
	Insightfulness	Avg	3.00	3.25	3.67	3.67	2.25	3.01
		Stdev	1.15	1.26	1.53	0.58	1.26	1.11
Novice	Relevance	Avg	3.67	3.50	4.17	3.17	3.33	3.38
		Stdev	0.52	0.55	0.75	0.75	0.82	0.95
	Innovation	Avg	2.83	3.67	3.50	3.83	3.83	3.11
		Stdev	0.98	0.52	1.05	0.98	0.75	1.07
	Insightfulness	Avg	3.50	3.67	3.50	3.33	3.17	3.13
		Stdev	0.55	0.52	0.84	1.03	0.98	0.96
GPT-4	Relevance	Avg	4.80	4.73	4.52	4.03	4.57	4.19
		Stdev	0.41	0.45	0.51	0.32	0.50	0.82
	Innovation	Avg	3.77	3.90	3.52	4.57	4.93	3.72
		Stdev	0.43	0.31	0.51	0.50	0.25	0.80
	Insightfulness	Avg	3.87	4.27	3.93	3.87	4.33	3.68
		Stdev	0.43	0.69	0.53	0.43	0.48	0.80

divergent thinking - producing a wide range of different ideas, and convergent thinking - incremental, step-by-step development of the details of a solution [74]. Indeed, the set of chosen ideas (see Table 3) illustrates that GPT-3 provided enhancements to the ideation process - all 5 teams chose project ideas that either combine GPT-3-Generated ideas with Human-Generated ideas, or are based on a GPT-3-Generated idea.

However, in our study, about 30% of the students pointed out that GPT-3 tends to be redundant and lacks creativity. How can we increase the novelty and creativity of the ideas contributed by an LLM to a collaborative group-AI ideation process? One possibility is through prompt engineering. In our study, students prompt the GPT-3 model directly, but integrating an LLM model into a custom interface, which implements back-end prompt engineering could potentially cause the LLM to provide better assistance for users during ideation. Several tools demonstrate the use of back-end prompt engineering within the context of education (e.g.[24, 40]) and decision making (e.g. [55]).

Applying this approach, we can help users to utilize prompts that challenge conventional molds. One direction is through connecting seemingly unrelated concepts in a way that invokes conceptual blending - a cognitive process in which distinct ideas are combined to create a new, unique concept [17]. Wang and colleagues have demonstrated the feasibility of this approach with a system that automatically suggests conceptual blends [82]. Another possibility is to adopt an approach similar to "Six Thinking Hats" [9], where different prompts are constructed, each defining a different persona for the LLM and hence leading to ideas that are provided in different style and represent different perspectives. Yet another approach might be an adaptation of the process proposed by Kahneman and colleagues to reduce noise in decision making – the authors propose that decision maker teams approach a problem by separating it into well-defined and separate focus areas [36]. For us, this could mean crafting different prompts that aim to elicit ideas about different aspect of the question at hand, e.g. a technological implementation, or an issue of aesthetics.

6.2 Discussion of results for RQ2: How can LLMs assist to evaluate ideas during the convergence stage of a collaborative group Brainwriting process?

The GPT-4 evaluation engine gave relatively high ratings to all of the ideas that were ultimately chosen by student teams, see Table 3. The fact that none of the chosen ideas received low ratings by GPT-4 is encouraging - it indicates that, if GPT-4 had been used to provide feedback for teams during the ideation process, it would not have filtered out ideas that were considered to be good by the teams.

At the same time, based on the moderate positive linear relationship between Expert and GPT-4 engine review scores, it also appears that, had GPT-4 been used to provide feedback during the ideation process, teams could have safely discarded ideas that were rated low by GPT-4. After all, none of the ideas that were rated low by GPT-4 were ultimately chosen, and none of the ideas that were rated low by Experts were rated high by GPT-4.

A final note on how LLMs can be used in supporting idea evaluation relates to the statistical terms of noise and bias [36]. Statistically, we saw that GPT-4 made consistent decisions as we asked it to evaluate each idea 29 times; thus, the noise in GPT-4 decisions was low. However, on average, GPT-4 and Expert evaluations differed from each other, representing a statistical bias. It is clear that this statistical observation in our data can translate into future versions of an LLM system that attempts to support ideation but provides feedback with harmful biases.

6.3 Implications for HCI education and practice

While generative AI have created new opportunities for supporting designers [27, 51], the structured integration of AI into design courses remains challenging [18]. In this paper we contribute a practical framework for collaborative group-AI Brainwriting that could be applied in HCI education and practice. We evaluated this framework with college students as part of their project work in a tangible interaction design course. The integration of co-creation processes with AI was aligned with the learning goals for the course, which aims to address some of the challenges that designers face when working with AI as design material [14, 32, 69, 81, 84]. Here we discuss the implications of our findings for HCI education and practice.

6.3.1 Expanding Ideas. Our findings demonstrate that integrating co-creation processes with AI into the ideation process of novice designers, could enhance the divergence stage where a wider range of different ideas is explored.

From our experience teaching tangible and embodied interaction design over the years [hidden for anonymity], students or novice designers who are new to TEI often limit their early ideation to traditional forms of interaction such as mobile phone apps and screen-based wearables. Results from our brainwriting activity indicate that using an LLM during ideation helped students to expand their ideas, and to consider different approaches (see Figure 4). While the creativity exhibited by GPT-3 itself was sometimes limited when prompted for producing new ideas, when it was prompted to expand on specific students' ideas, it often provided new modalities and suggested novel features that diverged from traditional graphical user interfaces (see section 5.3.2).

6.3.2 Prompt Engineering. The comments made by the students in our study make it clear that sometimes they struggled with creating effective prompts for GPT-3. This is an important issue, since our goal is to support ideation for teams with diverse levels of experience working with LLMs, not only professionals with training in the usage of the latest LLM technologies. While back-end prompting is one approach to address this challenge, it is clear that novice designers also require instruction on constructing effective prompts. It is thereby important to develop training materials for interaction designers in best practices of prompt engineering and to encourage them to consider how best to provide domain, task, and interaction style specific keywords with their prompts.

6.3.3 Increased Creativity through Shifting Attention. Tversky and Chou suggest that shifting attention between different problems fosters divergent thought and enhance creativity [74]. Future variation of our proposed framework for collaborative group-AI Brainwriting, could shift the group attention so that an LLM is prompted multiple times, where each prompt is focused on a different problem or aspect of the problem. Future research should explore such strategies for increasing the creativity of LLM-generated ideas.

6.3.4 Limitations of Non-Human Agents. The proposed group-AI Brainwriting process could be considered within the realm of more-than-human, post-humanist interaction design methods [19, 30, 79, 80] where agency is distributed among humans and non-human agents such as LLMs. When applying such methods it is important to remember that AI-based non-human agents are trained upon and import "traditional, humanist forms of logic and language" [77]. Thus, co-creation ideation processes might yield ideas that embody and amplify human social biases. While we did not identify specific social biases in the ideas produced by the proposed group-AI collaboration in response to the given problem statement, future work should probe for ideas that contain bias regarding specific groups or concepts. Future work could also develop methods of filtering out ideas that contain such bias.

6.3.5 Evaluating Ideas. In our study, we used the GPT-4 evaluation engine only after the ideation sessions were completed, so these evaluations were not available to teams. As we continue to work towards providing such LLM-generated evaluations to users, there are several issues to consider. First, such use of LLMs falls into the trend identified by Janssen et al. [33] that automation is increasingly being used by users with varied levels of expertise in using automated and AI-powered tools. LLM-generated feedback needs to be explained for designers with varying levels of training, such that they can appropriately calibrate their trust in the system [38, 42], understand it, and apply it effectively [25]. Second, our findings demonstrate that LLM-based idea evaluation could potentially filter out low-rated ideas in early stages of the process. This is promising, since teams of future or novice designers could receive early feedback, which provides direction and allows them to focus their time on developing the more promising ideas. Finally, as van Dijk warns us non-human agent still embody human biases [77], before making an LLM-based idea evaluation engine available to users, it is important to probe for and identify potential biases in its output.

6.4 Limitations

A clear limitation of our work is that we only examined the use of LLMs in ideation with novice designers, using a specific ideation process (Brainwriting), using a single problem statement, and within the context of HCI education. The students were also all novice users of GPT-3. Therefore, the study may not generalize to cases where the groups consist of expert LLM users, expert designers, or users that are assisted by highly trained prompt engineers. Furthermore, the study may not generalize to cases where the participants themselves are experts in the innovative domain, to different innovation domains or to other educational disciplines. Another limitation is that our work lacks an exploration of the long-term impact of integrating AI into HCI education, focusing primarily on immediate outcomes. Nevertheless, our study demonstrates the feasibility of enhancing enhancing Brainwriting with LLMs, and open avenues for future work in the intersection of AI, HCI and education, including developing custom interfaces and conducting longitudinal studies.

7 CONCLUSION

We expect that collaboration between humans and LLMs is one of several radical changes in the way in which humans will utilize machines in the coming years (cf. [33]). In this work we explore one potential scenario of such a collaboration, when an LLM supports a collaborative ideation process of a team. Our focus is on Brainwriting, and we explore how an LLM can enhance the ideas generated by the team using Brainwriting within an educational

context, as well as how it can help broaden the number of topics that are explored by the team. Our results indicate that LLMs can be useful in both aspects. Furthermore, we found that LLM-based idea evaluations hold promise in identifying both good ideas and poor ideas, which in the future could be useful feedback to teams as they work through the Brainwriting process, with the caveat that the system must be carefully designed such that its feedback is explainable and avoids propagating biases derived from human-generated data.

ACKNOWLEDGMENTS

This work was in part supported by NSF grant CMMI-1840085. The authors are grateful to Marios Constantinides and Duncan Brumby for generously contributing their time in our early conversations exploring the use of LLMs in group ideation. We also thank Marysabel Morales and Josephine Ramirez for assisting with early explorations of the data.

REFERENCES

[1] I Elaine Allen and Christopher A Seaman. 2007. Likert scales and data analyses. *Quality progress* 40, 7 (2007), 64–65.

[2] Kristina Andersen, Ron Wakkary, Laura Devendorf, and Alex McLean. 2019. Digital Crafts-Machine-Ship: Creative Collaborations with Machines. *Interactions* 27, 1 (dec 2019), 30–35. https://doi.org/10.1145/3373644

[3] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association, Washington, D.C.

[4] CompVis Group and Runway and Stability AI. 2022. Stable Diffusion Online. https://stablediffusionweb.com/. Accessed: 02-08-2023.

[5] Conceptboard. 2023. Brainwriting Technique Free Template. https://conceptboard.com/blog/brainwriting-technique-free-template/. Accessed: 12-09-2023.

[6] Conceptboard. 2023. Secure Collaboration Tool for Hybrid Teams - Conceptboard. https://conceptboard.com/. Accessed: 14-09-2023.

[7] Lauren E Coursey, Ryan T Gertner, Belinda C Williams, Jared B Kenworthy, Paul B Paulus, and Simona Doboli. 2019. Linking the divergent and convergent processes of collaborative creativity: The impact of expertise levels and elaboration processes. *Frontiers in Psychology* 10 (2019), 699.

[8] David H. Cropley, Caroline Theurer, Sven Mathijssen, and Rebecca L. Marrone. 2023. Fit-for-Purpose Creativity Assessment: Using Machine Learning to Score a Figural Creativity Test. *PsyArXiv Preprints* N/A, N/A (2023), N/A. Available online at PsyArXiv.

[9] Edward De Bono. 1999. *Six Thinking Hats.* Back Bay Books, New York.

[10] Douglas L. Dean, Jillian M. Hender, Thomas Lee Rodgers, and Eric L. Santanen. 2006. Identifying Quality, Novel, and Creative Ideas: Constructs and Scales for Idea Evaluation. *J. Assoc. Inf. Syst.* 7 (2006), 30. https://api.semanticscholar.org/CorpusID:15910404

[11] Dennis J. Devine, Laura D. Clayton, Jennifer L. Philips, Benjamin B. Dunford, and Sarah B. Melner. 1999. Teams in Organizations. *Small Group Research* 30, 6 (dec 1999), 678–711. https://doi.org/10.1177/104649649903000602

[12] Michael Diehl and Wolfgang Stroebe. 1987. Productivity loss in brainstorming groups: Toward the solution of a riddle. *Journal of personality and social psychology* 53, 3 (1987), 497.

[13] Anil R Doshi and Oliver Hauser. 2023. Generative artificial intelligence enhances creativity. *Available at SSRN* N/A, N/A (2023), N/A.

[14] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 278–288. https://doi.org/10.1145/3025453.3025739

[15] Steven Dow, Julie Fortuna, Dan Schwartz, Beth Altringer, Daniel Schwartz, and Scott Klemmer. 2011. Prototyping Dynamics: Sharing Multiple Designs Improves Exploration, Group Rapport, and Results. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 2807–2816. https://doi.org/10.1145/1978942.1979359

[16] Jeffrey H. Dyer, Hal B Gregersen, and Clayton Christensen. 2008. Entrepreneur behaviors, opportunity recognition, and the origins of innovative ventures. *Strategic Entrepreneurship Journal* 2, 4 (2008), 317–338. https://doi.org/10.1002/sej.59 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/sej.59

[17] Gilles Fauconnier and Mark Turner. 1998. Conceptual integration networks. *Cognitive Science* 22, 2 (1998), 133–187. https://doi.org/10.1016/S0364-0213(99)80038-X

[18] Rahel Flechtner and Aeneas Stankowski. 2023. AI Is Not a Wildcard: Challenges for Integrating AI into the Design Curriculum. In *Proceedings of the 5th Annual Symposium on HCI Education* (Hamburg, Germany) *(EduCHI '23)*. Association for Computing Machinery, New York, NY, USA, 72–77. https://doi.org/10.1145/3587399.3587410

[19] Elisa Giaccardi and Johan Redström. 2020. Technology and More-Than-Human Design. *Design Issues* 36, 4 (09 2020), 33–44. https://doi.org/10.1162/desi_a_00612 arXiv:https://direct.mit.edu/desi/article- pdf/36/4/33/1857682/desi_a_00612.pdf

[20] Rony Ginosar, Hila Kloper, and Amit Zoran. 2018. PARAMETRIC HABITAT: Virtual Catalog of Design Prototypes. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. Association for Computing Machinery, New York, NY, USA, 1121–1133. https://doi.org/10.1145/3196709.3196813

[21] K Girotra, L Meincke, C Terwiesch, and KT Ulrich. 2023. Ideas are dimes a dozen: large language models for idea generation in innovation (SSRN Scholarly Paper 4526071).

[22] Toshali Goel, Orit Shaer, Catherine Delcourt, Quan Gu, and Angel Cooper. 2023. Preparing Future Designers for Human-AI Collaboration in Persona Creation. In *Proceedings of the 2nd Annual Meeting of the Symposium on Human-Computer Interaction for Work*. ACM Press, New York, NY, USA, 1–14.

[23] Google. 2023. Bard: Chat-Based AI Tool from Google, Powered by PaLM 2. https://bard.google.com/. Accessed: 14-09-2023.

[24] Jieun Han, Haneul Yoo, Yoo Lae Kim, Jun-Hee Myung, Minsun Kim, Hyunseung Lim, Juho Kim, Tak Yeon Lee, Hwajung Hong, So-Yeon Ahn, and Alice H. Oh. 2023. RECIPE: How to Integrate ChatGPT into EFL Writing Education. In *Proceedings of the Tenth ACM Conference on Learning @ Scale*. ACM, New York, NY, USA, 1–8. https://api.semanticscholar.org/CorpusID:258823196

[25] AKM Bahalul Haque, AKM Najmul Islam, and Patrick Mikalef. 2023. Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. *Technological Forecasting and Social Change* 186 (2023), 122120.

[26] Andrew Hargadon. 2003. *How breakthroughs happen: The surprising truth about how companies innovate.* Harvard Business Press, Boston, MA.

[27] Harvard Business Review. 2022. How Generative AI Is Changing Creative Work. https://hbr.org/2022/11/how-generative-ai-is-changing-creative-work. Accessed: 01-08-2023.

[28] Scarlett R. Herring, Chia-Chen Chang, Jesse Krantzler, and Brian P. Bailey. 2009. Getting Inspired! Understanding How and Why Examples Are Used in Creative Design Practice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 87–96. https://doi.org/10.1145/1518701.1518717

[29] Peter A. Heslin. 2009. Better than brainstorming? Potential contextual boundary conditions to brainwriting for idea generation in organizations. *Journal of Occupational and Organizational Psychology* 82, 1 (2009), 129–145. https://doi.org/10.1348/096317908X285642 arXiv:https://bpspsychub.onlinelibrary.wiley.com/doi/pdf/10.1348/096317908X285642

[30] Sarah Homewood, Marika Hedemyr, Maja Fagerberg Ranten, and Susan Kozel. 2021. Tracing Conceptions of the Body in HCI: From User to More-Than-Human. 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 258, 12 pages. https://doi.org/10.1145/3411764.3445656

[31] Charles McLaughlin Hymes and Gary M Olson. 1992. Unblocking brainstorming through the use of a simple group editor. In *Proceedings of the 1992 ACM conference on Computer-supported cooperative work*. ACM Press, New York, NY, USA, 99–106.

[32] Nanna Inie, Jeanette Falk, and Steve Tanimoto. 2023. Designing Participatory AI: Creative Professionals' Worries and Expectations about Generative AI. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 82, 8 pages. https://doi.org/10.1145/3544549.3585657

[33] Christian P Janssen, Stella F Donker, Duncan P Brumby, and Andrew L Kun. 2019. History and future of human-automation interaction. *International journal of human-computer studies* 131 (2019), 99–107.

[34] Frans Johansson. 2004. *The medici effect.* Penerbit Serambi, Jakarta, Indonesia.

[35] Martin Jonsson and Jakob Tholander. 2022. Cracking the Code: Co-Coding with AI in Creative Programming Education. In *Proceedings of the 14th Conference on Creativity and Cognition* (Venice, Italy) *(C&C '22)*. Association for Computing Machinery, New York, NY, USA, 5–14. https://doi.org/10.1145/3527927.3532801

[36] Daniel Kahneman, Olivier Sibony, and Cass R Sunstein. 2021. *Noise: a flaw in human judgment.* Hachette UK, London, UK.

[37] Jingoog Kim and Mary Lou Maher. 2023. The effect of AI-based inspiration on human design ideation. *International Journal of Design Creativity and Innovation* 11, 2 (2023), 81–98. https://doi.org/10.1080/21650349.2023.2167124 arXiv:https://doi.org/10.1080/21650349.2023.2167124

[38] Lars Krupp, Steffen Steinert, Maximilian Kiefer-Emmanouilidis, Karina E Avila, Paul Lukowicz, Jochen Kuhn, Stefan Küchemann, and Jakob Karolus. 2023. Unreflected Acceptance–Investigating the Negative Consequences of ChatGPT-Assisted Problem Solving in Physics Education. *arXiv preprint arXiv:2309.03087* N/A, N/A (2023), N/A.

[39] Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics* 22, 1 (1951), 79–86.

[40] Harsh Kumar, Ilya Musabirov, Mohi Reza, Jiakai Shi, Anastasia Kuzminykh, Joseph Jay Williams, and Michael Liut. 2023. Impact of guidance and interaction strategies for LLM use on Learner Performance and perception. https://arxiv.org/abs/2310.13712

[41] Brian Lee, Savil Srivastava, Ranjitha Kumar, Ronen Brafman, and Scott R. Klemmer. 2010. Designing with Interactive Example Galleries. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 2257–2266. https://doi.org/10.1145/1753326.1753667

[42] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[43] J McCormack and A Dorin. 2014. Generative Design: A Paradigm for Design Research. *Futureground - DRS International Conference* N/A, N/A (2014), 17–21.

[44] Meta Research. 2023. LLaMA: Open and Efficient Foundation Language Models. https://research.facebook.com/publications/llama-open-and-efficient-foundation-language-models/. Accessed: 14-09-2023.

[45] Midjourney. 2022. Midjourney. https://www.midjourney.com/. [Accessed 01-08-2023].

[46] Miro. 2023. First Idea to Final Innovation: It All Lives Here. https://miro.com/product-overview/. Accessed: 14-09-2023.

[47] Miro. 2023. Miro AI. https://miro.com/ai/. Accessed: 09-09-2023.

[48] Osnat Mokryn and Hagit Ben-Shoshan. 2021. Domain-based Latent Personal Analysis and its use for impersonation detection in social media. *User Modeling and User-Adapted Interactions* 31, 4 (2021), 785–828.

[49] René Morkos. 2023. Council Post: Generative AI: It's Not All ChatGPT — forbes.com. https://www.forbes.com/sites/forbestechcouncil/2023/04/24/generative-ai-its-not-all-chatgpt/?sh=151ea40a32ef. [Accessed 01-08-2023].

[50] MURAL. 2023. Work Better Together with Mural's Visual Work Platform. https://www.mural.co/. Accessed: 14-09-2023.

[51] Thomas Olsson and Kaisa Väänänen. 2021. How Does AI Challenge Design Practice? *Interactions* 28, 4 (jun 2021), 62–64. https://doi.org/10.1145/3467479

[52] OpenAI. 2022. DALL·E 2. https://openai.com/dall-e-2. Accessed: 2-08-2023.

[53] OpenAI. 2023. GPT-4 — openai.com. https://openai.com/gpt-4. Accessed: 14-09-2023.

[54] Alex F Osborn. 1953. *Applied imagination.* Charles Scribner's Son's, New York, USA.

[55] Jeongeon Park, Bryan Min, Xiaojuan Ma, and Juho Kim. 2023. Choicemates: Supporting unfamiliar online decision-making with multi-agent conversational interactions. https://arxiv.org/abs/2310.01331

[56] Paul B Paulus and Mary T Dzindolet. 1993. Social influence processes in group brainstorming. *Journal of personality and social psychology* 64, 4 (1993), 575.

[57] Paul B Paulus and Huei-Chuan Yang. 2000. Idea generation in groups: A basis for creativity in organizations. *Organizational behavior and human decision processes* 82, 1 (2000), 76–87.

[58] Billy Perrigo. 2023. Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. https://time.com/6247678/openai-chatgpt-kenya-workers/

[59] Anuradha Reddy. 2022. Artificial everyday creativity: creative leaps with AI through critical making. *Digital Creativity* 33, 4 (2022), 295–313. https://doi.org/10.1080/14626268.2022.2138452

[60] Kevin Roose. 2022. The Brilliance and Weirdness of ChatGPT. https://www.nytimes.com/2022/12/05/technology/chatgpt-ai-twitter.html

[61] root. 2022. noda - mind mapping in virtual reality, solo or group — noda.io. https://noda.io/. [Accessed 09-09-2023].

[62] Vildan Salikutluk, Dorothea Koert, and Frank Jäkel. 2023. Interacting with Large Language Models: A Case Study on AI-Aided Brainstorming for Guesstimation Problems. In *HHAI 2023: Augmenting Human Intellect*. IOS Press, Amsterdam, Netherlands, 153–167.

[63] Albrecht Schmidt, Passant Elagroudy, Fiona Draxler, Frauke Kreuter, and Robin Welsch. 2024. Simulating the Human in HCD with ChatGPT: Redesigning Interaction Design with AI. *Interactions* 31, 1 (jan 2024), 24–31. https://doi.org/10.1145/3637436

[64] Orit Shaer and Angelora Cooper. 2023. Integrating Generative Artificial Intelligence to a Project Based Tangible Interaction Course. *IEEE Pervasive Computing* 23, 1 (2023), 5. https://doi.org/10.1109/MPRV.2023.3346548

[65] Joon Gi Shin, Janin Koch, Andrés Lucero, Peter Dalsgaard, and Wendy E. Mackay. 2023. Integrating AI in Human-Human Collaborative Ideation. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 355, 5 pages. https://doi.org/10.1145/3544549.3573802

[66] Pao Siangliulue, Kenneth C. Arnold, Krzysztof Z. Gajos, and Steven P. Dow. 2015. Toward Collaborative Ideation at Scale: Leveraging Ideas from Others to Generate More Creative and Diverse Ideas. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 937–945. https://doi.org/10.1145/2675133.2675239

[67] Dominik Siemon. 2023. Let the computer evaluate your idea: evaluation apprehension in human-computer collaboration. *Behaviour & Information Technology* 42, 5 (2023), 459–477.

[68] Wolfgang Stroebe, Bernard A. Nijstad, and Eric F. Rietzschel. 2010. Chapter Four - Beyond Productivity Loss in Brainstorming Groups: The Evolution of a Question. In *Advances in Experimental Social Psychology*, Mark P. Zanna and James M. Olson (Eds.). Vol. 43. Academic Press, Amsterdam, Netherlands, 157–203. https://doi.org/10.1016/S0065-2601(10)43004-X

[69] Hariharan Subramonyam, Colleen Seifert, and Eytan Adar. 2021. Towards A Process Model for Co-Creating AI Experiences. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference* (Virtual Event, USA) *(DIS '21)*. Association for Computing Machinery, New York, NY, USA, 1529–1543. https://doi.org/10.1145/3461778.3462012

[70] Ivan E. Sutherland. 1963. Sketchpad: A Man-Machine Graphical Communication System. In *Proceedings of the May 21-23, 1963, Spring Joint Computer Conference* (Detroit, Michigan) *(AFIPS '63 (Spring))*. Association for Computing Machinery, New York, NY, USA, 329–346. https://doi.org/10.1145/1461551.1461591

[71] Ivan Edward Sutherland. 2003. *Sketchpad: A man-machine graphical communication system.* Technical Report UCAM-CL-TR-574. University of Cambridge, Computer Laboratory. https://doi.org/10.48456/tr-574

[72] The New York Times. 2023. What's the Future for A.I.? — nytimes.com. https://www.nytimes.com/2023/03/31/technology/ai-chatbots-benefits-dangers.html. Accessed: 01-08-2023.

[73] Jakob Tholander and Martin Jonsson. 2023. Design Ideation with AI - Sketching, Thinking and Talking with Generative Machine Learning Models. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 1930–1940. https://doi.org/10.1145/3563657.3596014

[74] Barbara Tversky and Juliet Y. Chou. 2011. Creativity: Depth and Breadth. In *Design Creativity 2010*, Toshiharu Taura and Yukari Nagai (Eds.). Springer London, London, 209–214.

[75] Brygg Ullmer, Orit Shaer, Ali Mazalek, and Caroline Hummels. 2022. *Weaving Fire into Form: Aspirations for Tangible and Embodied Interaction* (1 ed.). Vol. 44. Association for Computing Machinery, New York, NY, USA.

[76] Priyan Vaithilingam, Tianyi Zhang, and Elena L Glassman. 2022. Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. In *Chi conference on human factors in computing systems extended abstracts*. ACM, NY, USA, 1–7.

[77] Jelle van Dijk. 2020. Post-Human Interaction Design, Yes, but Cautiously. In *Companion Publication of the 2020 ACM Designing Interactive Systems Conference* (Eindhoven, Netherlands) *(DIS' 20 Companion)*. Association for Computing Machinery, New York, NY, USA, 257–261. https://doi.org/10.1145/3393914.3395886

[78] Mathias Peter Verheijden and Mathias Funk. 2023. Collaborative Diffusion: Boosting Designerly Co-Creation with Generative AI. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 73, 8 pages. https://doi.org/10.1145/3544549.3585680

[79] Ron Wakkary. 2020. Nomadic Practices: A Posthuman Theory for Knowing Design. *International Journal of Design* 14, 3 (2020), 117.

[80] Ron Wakkary. 2021. *Things we could design: For more than human-centered worlds*. MIT Press, Boston, MA, USA.

[81] Qiaosi Wang, Michael Madaio, Shaun Kane, Shivani Kapania, Michael Terry, and Lauren Wilcox. 2023. Designing Responsible AI: Adaptations of UX Practice to Meet Responsible AI Challenges. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 249, 16 pages. https://doi.org/10.1145/3544548.3581278

[82] Sitong Wang, Savvas Petridis, Taeahn Kwon, Xiaojuan Ma, and Lydia B Chilton. 2023. PopBlends: Strategies for Conceptual Blending with Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 435, 19 pages. https://doi.org/10.1145/3544548.3580948

[83] Chauncey Wilson. 2013. Using Brainwriting For Rapid Idea Generation. https://www.smashingmagazine.com/2013/12/using-brainwriting-for-rapid-idea-generation/

[84] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-Examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376301