Towards an Evaluation of LLM-Generated Inspiration by Developing and Validating Inspiration Scale

Hyungyu Shin
KAIST
hyungyu.sh@kaist.ac.kr

Seulgi Choi
KAIST
igules8925@kaist.ac.kr

Ji Yong Cho
LG AI Research
Cornell University
jiyong.cho@lgresearch.ai

Sahar Admoni
Technion
saharad@campus.technion.ac.il

Hyunseung Lim
KAIST
charlie9807@kaist.ac.kr

Taewan Kim
KAIST
taewan@kaist.ac.kr

Hwajung Hong
KAIST
hwajung@kaist.ac.kr

Moontae Lee
LG AI Research
University of Illinois Chicago
moontae.lee@lgresearch.ai

Juho Kim
KAIST
juhokim@kaist.ac.kr

ABSTRACT

Researchers seek inspiration during the research process. Large Language Models (LLMs) have the potential to inspire researchers to make progress in their research, especially in the ideation process, but it is challenging to assess this capability. We envision (1) developing a scale—*Inspiration scale*—that captures key elements of inspiration, (2) evaluating the capability of existing LLMs for inspiring researchers in the research ideation process, and (3) further transforming the developed scale into an auto-assessment rubric for LLMs to align human-perceived and machine-assessed inspiration. In this paper, we develop a list of items for human evaluators by (1) compiling metrics for inspiration through a systematic literature review and (2) contextualizing them in the context of research ideation. We discuss the next steps to validate our scale, evaluate LLMs using the scale, and develop an auto-assessment rubric aligned with our original scale.

CCS CONCEPTS

• **Computing methodologies → Natural language generation**; • **Human-centered computing → Human computer interaction (HCI)**; **HCI design and evaluation methods**; • **General and reference → Evaluation**.

KEYWORDS

Large Language Models, Evaluation, Creativity, Inspiration, Scale development

1 INTRODUCTION

In a scientific research process, researchers seek inspirations in performing various tasks such as identifying research opportunities by analyzing prior literature, designing and conducting experiments, and ideating future research directions [41]. Inspirations[1] involve *evocation* (i.e., there need to be external stimuli), *transcendence* (i.e., an individual realizes novelty on the stimuli), and *motivation* (i.e., the stimuli motivates an individual to act) [111]. In this sense, Large Language Models (LLMs) have a potential to meet the characteristics by providing novel ideas (*evocation* and *transcendence*) on topics that researchers are interested in, which could help researchers create actionable to-do items for research progress (*motivation*).

It is questionable whether and how well LLMs can inspire researchers in a scientific research process. One way to answer the question is to measure how inspiring LLM responses are in scientific QA [69] settings, but it is unclear what metrics should be employed to specifically inform the strengths and weaknesses of the models providing inspiration. Prior research used a set of related metrics to evaluate how inspiring LLM responses are (e.g., whether generated texts are helpful [47], novel [35], and creative [26]), but the metrics in the prior research are highly diverse and context-specific depending on the research focus. Such diversity makes it challenging to understand an overall landscape of aspects that should be considered when evaluating the capability of LLMs to provide inspiration. Developing a general-purpose, validated scale for assessing LLM-generated inspiration can offer a straightforward yet thorough method for evaluating their inspiration capability, using fine-grained metrics to examine and compare the performance of LLMs.

Our ultimate goal is to (1) develop and validate an inspiration scale for evaluating inspirations by following a standard methodology for developing scales [32, 61, 85] and (2) evaluate how inspirational LLMs are in a scientific QA task. We acknowledge that a single standard inspiration scale may not perfectly capture the unique nature of diverse research domains and processes. Nevertheless, the inspiration scale can serve as a good default for assessing the inspirational capability of LLMs, offering researchers the opportunity to adapt the scale according to their specific interests. In this paper, we report (1) results of a systematic literature review that compiles metrics used for measuring inspiration and (2) a list of items (i.e., questions for human evaluation) after contextualizing the metrics to research process (Figure 1). Then we discuss future work to achieve the ultimate goal.

[1] We denote "inspirations" as artifacts that aim to facilitate inspiration. We use the term "inspiration" to indicate a conceptual entity.

1st HEAL Workshop at CHI Conference on Human Factors in Computing Systems, May 12, Honolulu, HI, USA
2024.

A. Idea Quality

Metrics

- Domain-specific quality metrics
- Novelty
- Originality
- Creativity
- Feasibility
- Elaboration
- Conventional
- Completeness

Novelty of ideas
- The response breaks away from common solutions and offers truly unique perspectives.
- The response offers insightful solutions that haven't been explored before.
- The answer is infused with innovative and thought-provoking approaches.

Feasibility of ideas
- The response offers practical and attainable solutions.
- The response provides ideas feasible to execute.
- Ideas in the response can be realistically implemented.
- The response provides ideas that are realistic within the given context.

Applicability of ideas
- The response offers adaptable information readily applicable to various situations.
- The response provides a foundation easily tailored to specific needs and contexts.
- The response provides a versatile foundation for further exploration and application.

B. Idea Space

Metrics

- Quantity
- Diversity
- Evenness
- Depth

Breadth of knowledge - diversity of ideas
- The response presents a wealth of different approaches, ensuring a well-rounded perspective.
- The response showcases a broad spectrum of solutions, encouraging further consideration.
- The response offers a wide range of ideas, showcasing different perspectives and approaches.

Breadth of knowledge - sufficiency in number of ideas
- The response offers enough ideas to spark creativity and stimulate exploration.
- The response provides a sufficient starting point for productive brainstorming and problem-solving.

Breadth of knowledge - no redundancy between ideas
- The provided options are clearly distinct, avoiding redundancy and overlap.
- The suggestions cover a broad spectrum, encompassing various possibilities without being repetitive.

C. Impact of ideas on users

Metrics

- Inspiring
- Usefulness
- Surprise
- Task influence
- Helpfulness
- Satisfying
- Motivational
- Distractive

Perceived utility
- The suggestions are helpful for exploring different possibilities.
- The response is useful in the context of brainstorming ideas.
- This information helps to expand the scope of potential solutions.
- The response offers valuable input for exploring different approaches.

Promoting creative thinking
- The response prompts me to develop an approach I had not previously considered.
- The response stimulates me to explore new directions of solving the problem.
- The response sparks my curiosity to try out alternative solutions.
- The response encourages me to think of a new way to tackle the issue.

Impact on my ideation process
- The response questions my initial assumptions or beliefs, encouraging me to explore different viewpoints.
- The response influences my thought process, providing me with fresh perspectives and methods.

D. Social acceptance

Metrics

- Appropriateness
- Flexibility
- Value
- Realistic
- Acceptance

Research community
- The response directly tackles the current challenges identified by researchers in my field of study.
- The response meets the specific needs and goals of the research community of my interest.
- The response aligns with the current state of knowledge and ongoing research endeavors.
- The response closely corresponds with the latest research trends and priorities within the relevant research community.

Ethical consideration
- The response suggests approaches that align with widely accepted ethical principles.
- The response presents solutions that are unlikely to raise ethical concerns for most people.
- The response avoids solutions that could be seen as harmful or morally questionable.
- The response champions ethically sound approaches to the problem.

E. Human Alignment

Metrics

- Relevance
- Elaboration
- Fluency
- Understandability
- Repetitiveness

Specificity
- The response is rich in information and specific examples.
- The response offers concrete details and data to support its claims and arguments.
- The response lacks depth and specifics

Comprehensibility
- The response presents a structured and logical flow of information, aiding comprehension.
- The information is well-organized and easy to follow, making it clear and understandable.
- The provided information is clearly defined and avoids ambiguity, ensuring every point is precise.
- The details are clearly presented and easy to follow, even for complex topics

Relevancy
- The response aligns with the subject matter of my inquiry.
- The response answers the essence of my question.
- The response directly addresses my question.

Consistency
- The response presents coherent and consistent ideas, avoiding any conflicting perspectives.
- The response demonstrates a unified and well-integrated flow of thought, free of contradictions.
- The response clearly communicates ideas free of ambiguity, preventing any misinterpretations or contradictions.

Figure 1: Results of contextualization. Metrics found from the systematic review were grouped into five themes and translated into items in the context of research ideation.

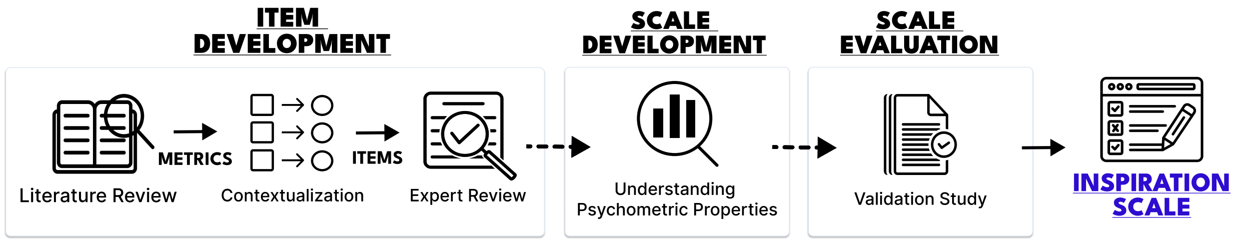

Figure 2: Overall approach for developing an inspiration scale. We develop items by compiling metrics for evaluating inspiration from systematic literature review and contextualizing the metrics into a research process. After getting feedback on the developed items from experts, we are planning to polish the items by understanding psychometric properties, conducting exploratory factor analysis. Then we validate the items as a scale by conducting confirmatory factor analysis and an additional validation study.

2 RELATED WORK

We review research about (1) the concept of inspiration and their effects and (2) systems that offer inspirations and their evaluation.

2.1 Effects of inspiration

Inspiration is the process of being stimulated by external artifacts [80, 103]. Some other perspectives define inspiration as the external stimuli [45, 113]. The inspiration helps individuals to be more creative in various ways. Specifically, the inspiration enriches a creative process and influences human behavior, leading to positive experiences. In terms of creative process, inspiration assists in solve problems [45], influences cognitive mechanisms [113], and alters problem framing [65]. Once getting inspired, individuals feel a desire to express [34], generate new and diverse ideas [103], and consider a wide range of perspectives [103]. It results in feeling excitement, satisfaction, a sense of coalescence, and an arousal of long-term memory [37, 45]. However, inspiration can also have negative consequences such as leading design fixations and disrupting designers' thinking [45].

Evaluating the effects of inspirations is mostly done by human evaluations with various focuses. One of the popular methods is to design Likert-scale questions that ask users (i.e., those who receive inspirations) about how they perceive the given inspirations [16, 54, 121]. External judges are recruited for evaluating the quality of inspirations in the domain and understanding how the inspirations affected the creative process of users [15, 16]. We aim to develop and validate an inspiration scale for evaluating inspirations so that researchers can not only measure inspirations with a validated scale but also be informed about a landscape of evaluating inspirations as a guideline.

2.2 Systems for offering inspirations

Research has introduced systematic approaches for providing inspiration where some of the approaches are (1) offering predefined sets of examples [20, 84], (2) sampling ideas through computational exploration of the idea space [23], and (3) adaptively offering ideas based on the user's ongoing creations [21, 46]. As research is a highly creative process, systems for facilitating effective inspirations for researchers have been introduced. For instance, research introduced techniques for suggesting novel ideas [116], searching related work [66, 87], adapting ideas [43, 52], and generating review of a paper [25].

To show the effects of inspirations offered by systems, research measured various dimensions of provided inspirations, including creativity [4, 103], usefulness [20, 121], and how it influences the creative process [54, 104]. However, research has employed diverse sets of metrics with various descriptions of what the metrics mean. It remains unclear what the comprehensive set of metrics is for measuring inspiration. We aim to develop and validate an inspiration scale via a systematic methodology for scale development.

Creativity Supporting Tools (CSTs) [102] are related systems where the goal is to assist users' creative process. The evaluation of CSTs mainly focuses on usability of the tools and quality of users' creative outcomes. For assessing the usability, Creativity Support Index (CSI) [18] has been widely used in addition to general usability assessment metrics such as NASA-TLX [40] and SUS [10]. Assessing the creative outcome is mostly done by human evaluations of domain experts. In this work, we focus on measuring LLM responses instead of a holistic evaluation of LLM as a tool for offering inspirations. We believe that our scale can be informative for evaluating tools for offering inspirations by incorporating usability perspectives of the process.

3 OVERALL APPROACH FOR DEVELOPING INSPIRATION SCALE

Figure 2 shows an overall approach for developing and validating an inspiration scale, following a standard methodology [6]. In the first phase, we develop a set of items (i.e., questions for the evaluation) by taking an inductive approach: (1) conducting a systematic literature review to create a list of relevant metrics with inspiration, (2) creating items by contextualizing the metrics to research process, and (3) reflecting expert feedback on the created items. The second phase is scale development, where we run a human evaluation study (N = 150) with the created items. Using the evaluation results, we perform exploratory factor analysis [30] to identify core factors

HCI	[5, 7, 13, 15, 16, 20–23, 26, 27, 33, 38, 42, 48, 52–54, 63, 67, 70, 73, 74, 77, 78, 83, 84, 90, 98–101, 103–105, 107, 108, 110, 114, 119, 121, 124, 127]
AI/LLM	[2, 3, 8, 12, 14, 17, 19, 26, 28, 29, 39, 44, 51, 55, 57–60, 62, 64, 68, 72, 75, 76, 82, 86, 88, 89, 91, 92, 94, 95, 115, 118–120, 126, 128]
Cognitive Science	[31, 36, 50, 71, 79, 81, 93, 96, 97, 106, 109, 112, 117, 123]

Table 1: The list of papers containing human evaluation with metrics, identified from the systematic literature review.

that describe the items and remove potentially redundant items. The final phase is scale evaluation, which evaluates our items as a scale. In other words, we evaluate whether the items capture key properties of inspirations via a validation study.

In this paper, we report the results of a systematic literature review and a list of items through contextualization (Phase 1). Then we describe future work (Phase 2 and 3).

4 SYSTEMATIC LITERATURE REVIEW

To create items for developing an inspiration scale, we conducted a systematic literature review as an inductive approach of item development.

4.1 Procedure

Figure 3 shows a diagram illustrating the overall procedure. We sampled papers that include evaluation of inspirations given by a system (e.g., ideas [23, 75], feedback [9, 56], and images [74]) until the year of 2023 through web search. Since the concept of inspiration has been discussed in various research fields, we targeted three fields of research: (1) AI/LLM, (2) HCI, and (3) Cognitive Science. Specifically, our goal was to review papers that (1) asked users to perform creative tasks, (2) offered inspirations to support the task, and (3) evaluated the provided inspirations. As such, our search keyword includes “inspiration”, “creativity”, “ideation”. Also, we included keywords “measure” and “Likert” for sampling papers that include human evaluation with specific metrics as evaluating ideas have been mostly done by human (i.e., users and external judges). Finally, we included field names (e.g., “LLM” for AI/LLM, and “Human-Computer Interaction” for HCI).

From the search results (523 papers in total), we filtered papers that contain evaluations of inspirations from the system by reading their evaluation methodology. Then we listed the name (e.g., “Unexpectedness”) and description of the metrics (e.g., “how unexpected a prompt was”) from each paper, which resulted in 97 papers (Table 1) with 202 metrics in total (allowing duplicates between papers). To come up with a list of unique metrics, we assigned a label for each of the raw metrics where a label represents a single metric. In other words, we assigned the same label for raw metrics if the description of the metrics took the same perspective. Three authors assigned labels for 20% of the metrics together and made a consensus around how to assign labels. Then, each of the three authors individually assigned labels for 80% of the metrics and resolved conflicts together. Finally, 30 unique labels have been identified, and each label corresponds to a metric.

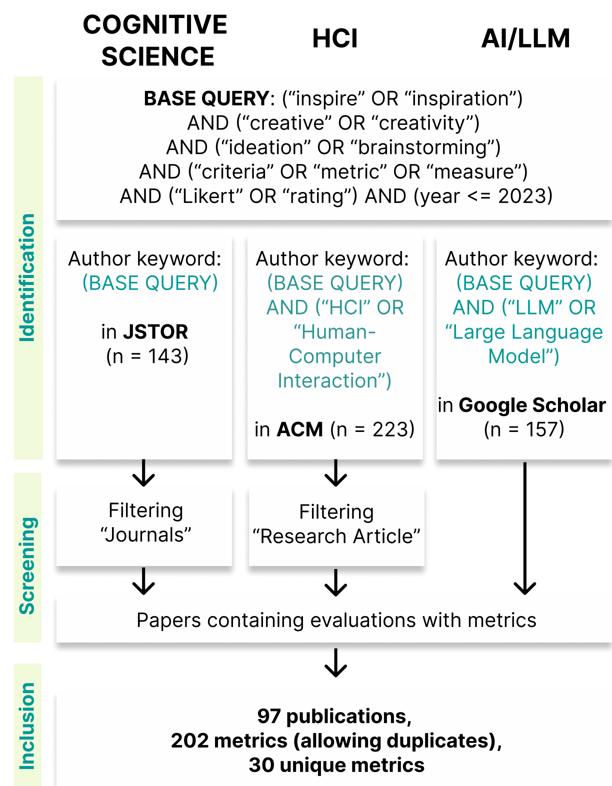

Figure 3: The systematic literature review process.

4.2 Result

Figure 1 shows the list of metrics identified from the systematic literature review. We recognized 5 themes of the metrics based on the target of measures: evaluating (1) idea quality, (2) idea space, (3) impact of ideas on users, (4) social acceptance, and (5) human alignment.

4.2.1 Idea quality. One of the major evaluation targets is the quality of ideas, where commonly used metrics include *Novelty*, *Originality*, *Creativity*, and *Feasibility*. Research also introduces domain-specific quality evaluation metrics depending on the context (e.g., Coherency for story writing task [21] and Aesthetics for image generation task [7]). Research often invites expert judges to evaluate the quality of ideas.

4.2.2 Idea space. The space of ideas is another important theme of evaluation in ideation tasks. The subject of idea creation can be both systems and users. Metrics include *Quantity*, *Diversity*, and *Evenness*. Such metrics can be employed not only for human evaluation but also for quantitative evaluation through operationalization (e.g., computing Diversity as the mean pairwise distance between ideas [16]).

4.2.3 Impact of ideas on users. It is important to understand how the ideas affect users' creative process, cognitive process, and overall experience. Research used metrics such as *Inspiring*, *Usefulness*, and *Surprise*, and *Task Influence*. Researchers measured the impact of ideas by asking such questions to the users [54, 105] or examining how users' creative process have been altered after the users browsed system-offered ideas [16].

4.2.4 Social acceptance. Research also examined whether the ideas can be socially accepted by taking a broader perspective. Metrics include *Appropriateness*, *Flexibility*, and *Acceptance*. We found, however, that few research discussed ethical considerations of the ideas, which is an important consideration in AI [49]. We add a few items regarding the ethical perspectives, emphasizing the importance of the community.

4.2.5 Human alignment. The idea description needs to be aligned with human values. Metrics include *Relevance*, *Elaboration*, and *Understandability*. Metrics in this theme can be generally employed in contexts other than ideation tasks. Researchers may employ other related metrics about human alignments as well (e.g., *Factuality* [125]).

5 CONTEXTUALIZING THE METRICS TO RESEARCH PROCESS

Since the list of metrics is organized from various papers with different contexts, we developed items by contextualizing the metrics into a research process. Our approach is to develop multiple items for each theme so that the items cover important dimensions discussed in the prior literature.

We took an iterative approach to create clear items that avoid ambiguity and multiple interpretations. First, we wrote the description of the metrics, starting with "The response" as our evaluation target is LLM response. With the initial list of items, we conducted a pilot human evaluation study where the three authors rated LLM responses using the initial items. The LLM responses were for questions that ask potential future research directions, given a discussion section of a research paper. For items that are unclear or have multiple possible interpretations, we rewrote the items or decomposed the items into multiple items to make them clearer. We iterated the process until we have clear items.

As a result, we designed 48 initial items (Figure 1). Note that the items will be reduced to a smaller number of items where we expect to have approximately 10 items in the final version, considering the practicality of the evaluation. Prior research recommended that the initial pool of items could be five times as large as the final version [122].

We are planning to conduct an expert review session on the items to understand whether the items represent the key characteristics of inspiration in the research process. Our plan is to invite experienced researchers (e.g., faculty or postdoctoral-level) in different research fields (e.g., AI/LLM, HCI, and Cognitive Science) to get feedback from diverse perspectives.

6 FUTURE WORK

To evaluate inspirations of LLMs via both human evaluation and automatic evaluation, our future work addresses (1) understanding psychometric properties and validating the scale and (2) evaluating inspirations of multiple LLMs and analyzing their strengths and weaknesses in offering inspirations via both human evaluation and automatic evaluation.

6.1 Understanding psychometric properties and validating the scale

To develop a validated scale, we perform two studies: (1) a study to understand psychometric properties and further polish the items and (2) another study to validate the list of items as a scale.

6.1.1 Study 1. Understanding psychometric properties. The scale development process involves identifying core factors that describe the overall items and reducing redundant items [6, 111]. Therefore, we aim to conduct a study with 150 participants via crowdsourcing. In the study, we ask participants to rate a GPT-4 response, which is potentially inspiring for the participants, using the developed items in a scenario of brainstorming future research ideas. As our scenario requires brainstorming research ideas, our target population of the participants is those who have prior research experiences.

It is not straightforward to generate GPT-4 responses that could be inspiring for the participants (i.e., researchers). To provide inspiring experiences, we (1) ask participants to upload a pdf of a paper that they are interested in, (2) generate an ideation question using GPT-4 by leveraging the paper contents (See Figure 4 for the prompt), (3) ask participants to regenerate or revise the question in a way that the response is expected to provide inspirations to them, and (4) ask participants to rate the response generated by GPT-4 for the question. In this way, we can expect that the participants would rate responses that potentially offer inspirations.

After collecting the ratings, we conduct exploratory factor analysis [30] to identify the core factors. Based on the factors, we can further reduce the items by removing highly correlating items. We will also drop items that are not closely related to any of major groups. An item reliability test [24] could further drop items that are not closely correlates with items in the same factor group.

6.1.2 Study 2. Validating the scale. In this study, we aim to validate our scale. The procedure is similar to Study 1, but we perform confirmatory factor analysis [11] and reliability test [24] on the evaluation results.

In the study, we also include other validated measures in psychology (e.g., cognitive load, enjoyment, and Inspiration Scale [111]) to show discriminant validity and convergent validity of our scale. We further validate our scale by comparing the ratings of the responses to different types of questions. Here, we assume that certain questions are more prone to be inspirational and ratings for those will be higher. For example, asking about solutions to a problem would be more likely to produce inspiring responses compared to asking about the definition of a concept.

6.2 Evaluating inspirations of LLMs

Using the scale, we evaluate how well existing LLMs offer inspirations to researchers and analyze strengths and weaknesses of LLMs in providing inspirations to researchers. We follow a similar methodology for evaluating LLM capabilities [1, 125].

[[Title, Abstract, Introduction, and Discussion section of paper]]

{{ *Title, Abstract, Introduction, and Discussion section of the user's paper* }}

[[Definition of ideation question]]

Ideation questions in the context of research process explore ways to extend current research by either improving upon it or by applying the findings to different contexts.

[[Examples]]

What potential future research directions could be explored to further enhance the effectiveness and efficiency of multi-task offline reinforcement learning, particularly in terms of integrating adaptive learning algorithms or exploring different domains and applications beyond robotic manipulation and drone navigation?

How could the OnIS framework be further developed to enhance its robustness and adaptability in dynamically changing environments, and what are the potential applications in real-world scenarios where environmental unpredictability is a significant challenge?

[[Instruction]]

Now, generate three self-contained questions for ideation. Avoid the use of the phrase 'in the text.' Exclude any second-person pronouns like 'you,' so no questions should start by 'can you.' Spell out all the acronyms.

Figure 4: The prompt used for generating ideation questions that are expected to produce inspiring ideas to researchers. In addition to the definition of the ideation question, we put the title, abstract, introduction, and discussion contents of a paper that the researcher is interested. In this way, we can generate questions that address specific contexts of the paper, which is likely to produce inspiring ideas about the paper.

In the evaluation process, we conduct both human evaluation and automatic evaluation to see the feasibility of using LLMs as an evaluator. We report the correlation between the two evaluation results and discuss what automatic evaluation can measure well and not. Also, we can further transform the developed scale into a rubric for automatic evaluation that better aligns with human-perceived inspiration.

ACKNOWLEDGMENTS

This work was supported by LG AI Research.

REFERENCES

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

[2] Victor Nikhil Antony and Chien-Ming Huang. 2023. ID.8: Co-Creating Visual Stories with Generative AI. arXiv:2309.14228 [cs.HC]

[3] Trevor Ashby, Braden K Webb, Gregory Knapp, Jackson Searle, and Nancy Fulda. 2023. Personalized Quest and Dialogue Generation in Role-Playing Games: A Knowledge Graph- and Language Model-based Approach. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 290, 20 pages. https://doi.org/10.1145/3544548.3581441

[4] Trevor Ashby, Braden K Webb, Gregory Knapp, Jackson Searle, and Nancy Fulda. 2023. Personalized Quest and Dialogue Generation in Role-Playing Games: A Knowledge Graph-and Language Model-based Approach. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

[5] Suyun Sandra Bae, Oh-Hyun Kwon, Senthil Chandrasegaran, and Kwan-Liu Ma. 2020. Spinneret: Aiding creative ideation through non-obvious concept associations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[6] Godfred O Boateng, Torsten B Neilands, Edward A Frongillo, Hugo R Melgar-Quiñonez, and Sera L Young. 2018. Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Frontiers in public health* 6 (2018), 149.

[7] Stephen Brade, Bryan Wang, Mauricio Sousa, Sageev Oore, and Tovi Grossman. 2023. Promptify: Text-to-image generation through interactive prompt exploration with large language models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–14.

[8] Herbie Bradley, Andrew Dai, Hannah Teufel, Jenny Zhang, Koen Oostermeijer, Marco Bellagente, Jeff Clune, Kenneth Stanley, Grégory Schott, and Joel Lehman. 2023. Quality-Diversity through AI Feedback. arXiv:2310.13032 [cs.CL]

[9] Herbie Bradley, Andrew Dai, Hannah Teufel, Jenny Zhang, Koen Oostermeijer, Marco Bellagente, Jeff Clune, Kenneth Stanley, Grégory Schott, and Joel Lehman. 2023. Quality-Diversity through AI Feedback. *arXiv preprint arXiv:2310.13032* (2023).

[10] John Brooke. 1996. Sus: a "quick and dirty'usability. *Usability evaluation in industry* 189, 3 (1996), 189–194.

[11] Timothy A Brown. 2015. *Confirmatory factor analysis for applied research*. Guilford publications.

[12] Celina Burlin. 2023. Explainability to enhance creativity: A human-centered approach to prompt engineering and task allocation in text-to-image models for design purposes.

[13] Ana Caraban, Loukas Konstantinou, and Evangelos Karapanos. 2020. The nudge deck: A design support tool for technology-mediated nudging. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 395–406.

[14] Tuhin Chakrabarty, Vishakh Padmakumar, Faeze Brahman, and Smaranda Muresan. 2023. Creativity Support in the Age of Large Language Models: An Empirical Study Involving Emerging Writers. *ArXiv* abs/2309.12570 (2023). https://api.semanticscholar.org/CorpusID:262217523

[15] Joel Chan, Steven Dang, and Steven P Dow. 2016. Comparing different sensemaking approaches for large-scale ideation. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2717–2728.

[16] Joel Chan, Steven Dang, and Steven P Dow. 2016. Improving crowd innovation with expert facilitation. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1223–1235.

[17] Ellen Zhang Chang. 2022. *Surrounding Dialogue Generation using Deep Learning with Adapters*. Master's thesis. NTNU.

[18] Erin Cherry and Celine Latulipe. 2014. Quantifying the creativity support of digital tools through the creativity support index. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21, 4 (2014), 1–25.

[19] DaEun Choi, Sumin Hong, Jeongeon Park, John Joon Young Chung, and Juho Kim. 2023. CreativeConnect: Supporting Reference Recombination for Graphic Design Ideation with Generative AI. arXiv:2312.11949 [cs.HC]

[20] David Chung and Rung-Huei Liang. 2015. Understanding the usefulness of ideation tools with the grounding lenses. In *Proceedings of the Third International Symposium of Chinese CHI*. 13–22.

[21] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A Smith. 2018. Creative writing with a machine in the loop: Case studies on slogans and stories. In *23rd International Conference on Intelligent User Interfaces*. 329–340.

[22] Andrew Clayphan, Anthony Collins, Christopher Ackad, Bob Kummerfeld, and Judy Kay. 2011. Firestorm: a brainstorming application for collaborative group work at tabletops. In *Proceedings of the ACM international conference on interactive tabletops and surfaces*. 162–171.

[23] Samuel Rhys Cox, Yunlong Wang, Ashraf Abdul, Christian Von Der Weth, and Brian Y. Lim. 2021. Directed diversity: Leveraging language embedding distances for collective creativity in crowd ideation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–35.

[24] Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *psychometrika* 16, 3 (1951), 297–334.

[25] Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. 2024. MARG: Multi-Agent Review Generation for Scientific Papers. *arXiv preprint arXiv:2401.04259* (2024).

[26] Zijian Ding, Arvind Srinivasan, Stephen MacNeil, and Joel Chan. 2023. Fluid Transformers and Creative Analogies: Exploring Large Language Models' Capacity for Augmenting Cross-Domain Analogical Creativity. In *Proceedings of the 15th Conference on Creativity and Cognition*. 489–505.

[27] Graham Dove and Sara Jones. 2014. Using data to stimulate creative thinking in the design of new products and services. In *Proceedings of the 2014 conference on Designing interactive systems*. 443–452.

[28] Kevin Dunnell, Trudy Painter, Andrew Stoddard, and Andy Lippman. 2023. Latent Lab: Large Language Models for Knowledge Exploration. *arXiv preprint arXiv:2311.13051* (2023).

[29] Noyan Evirgen. 2023. *Exploring User-Centric Generative Models: Advancing User Control, Comprehension, and Creative Capacity*. Ph. D. Dissertation. University of California, Los Angeles.

[30] Leandre R Fabrigar and Duane T Wegener. 2011. *Exploratory factor analysis*. Oxford University Press.

[31] Yue Fan, H Chad Lane, and ̈Omer Delialioğlu. 2022. Open-ended tasks promote creativity in Minecraft. *Educational Technology & Society* 25, 2 (2022), 105–116.

[32] Corey Ford and Nick Bryan-Kinns. 2023. Towards a Reflection in Creative Experience Questionnaire. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[33] Veronika Gamper, Andreas Butz, and Klaus Diepold. 2017. Sooner or later? immediate feedback as a source of inspiration in electronic brainstorming. In *Proceedings of the 29th Australian Conference on Computer-Human Interaction*. 182–190.

[34] Jacquline Faye Gilson. 2015. *An exploration into inspiration in heritage interpretation through virtual World Café*. Royal Roads University (Canada).

[35] Karan Girotra, Lennart Meincke, Christian Terwiesch, and Karl T Ulrich. 2023. Ideas are dimes a dozen: Large language models for idea generation in innovation. *Available at SSRN 4526071* (2023).

[36] Karan Girotra, Christian Terwiesch, and Karl T Ulrich. 2010. Idea generation and the quality of the best idea. *Management science* 56, 4 (2010), 591–605.

[37] Nicole M Gnezda. 2011. Cognition and emotions in the creative process. *Art Education* 64, 1 (2011), 47–52.

[38] Shihui Guo, Yubin Shi, Pintong Xiao, Yinan Fu, Juncong Lin, Wei Zeng, and Tong-Yee Lee. 2023. Creative and progressive interior color design with eye-tracked user preference. *ACM Transactions on Computer-Human Interaction* 30, 1 (2023), 1–31.

[39] Jennifer Haase and Paul H.P. Hanel. 2023. Artificial muses: Generative artificial intelligence chatbots have risen to human-level creativity. *Journal of Creativity* 33, 3 (Dec. 2023), 100066. https://doi.org/10.1016/j.yjoc.2023.100066

[40] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[41] Tom Hope, Doug Downey, Daniel S Weld, Oren Etzioni, and Eric Horvitz. 2023. A computational inflection for scientific discovery. *Commun. ACM* 66, 8 (2023), 62–73.

[42] Tom Hope, Ronen Tamari, Daniel Hershcovich, Hyeonsu B Kang, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2022. Scaling Creative Inspiration with Fine-Grained Functional Aspects of Ideas. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.

[43] Tom Hope, Ronen Tamari, Hyeonsu Kang, Daniel Hershcovich, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2021. Scaling creative inspiration with fine-grained functional facets of product ideas. *arXiv preprint arXiv:2102.09761* (2021).

[44] Md Naimul Hoque, Tasfia Mashiat, Bhavya Ghai, Cecilia Shelton, Fanny Chevalier, Kari Kraus, and Niklas Elmqvist. 2024. The HaLLMark Effect: Supporting Provenance and Transparent Use of Large Language Models in Writing with Interactive Visualization. arXiv:2311.13057 [cs.HC]

[45] Guanhua Hou et al. 2023. The impact of design creativity: Inspirations and timing of stimulation. *Telematics and Informatics Reports* 12 (2023), 100105.

[46] Chieh-Yang Huang, Shih-Hong Huang, and Ting-Hao Kenneth Huang. 2020. Heteroglossia: In-situ story ideation with the crowd. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[47] Chieh-Yang Huang, Saniya Naphade, Kavya Laalasa Karanam, and Ting-Hao'Kenneth' Huang. 2023. Conveying the Predicted Future to Users: A Case Study of Story Plot Prediction. *arXiv preprint arXiv:2302.09122* (2023).

[48] Anniek Jansen and Sara Colombo. 2023. Mix & Match Machine Learning: An Ideation Toolkit to Design Machine Learning-Enabled Solutions. In *Proceedings of the Seventeenth International Conference on Tangible, Embedded, and Embodied Interaction*. 1–18.

[49] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature machine intelligence* 1, 9 (2019), 389–399.

[50] Gita Venkataramani Johar, Morris B Holbrook, and Barbara B Stern. 2001. The role of myth in creative advertising design: Theory, process and outcome. *Journal of Advertising* 30, 2 (2001), 1–25.

[51] Boxiang Dong Kai Wang and Junjie Ma. 2023. Testing Computational Assessment of Idea Novelty in Crowdsourcing. *Creativity Research Journal* 0, 0 (2023), 1–14.

[52] Hyeonsu B Kang, Xin Qian, Tom Hope, Dafna Shahaf, Joel Chan, and Aniket Kittur. 2022. Augmenting scientific creativity with an analogical search engine. *ACM Transactions on Computer-Human Interaction* 29, 6 (2022), 1–36.

[53] Youwen Kang, Zhida Sun, Sitong Wang, Zeyu Huang, Ziming Wu, and Xiaojuan Ma. 2021. MetaMap: Supporting visual metaphor ideation through multi-dimensional example-based exploration. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[54] Jeongyeon Kim, Sangho Suh, Lydia B Chilton, and Haijun Xia. 2023. Metaphorian: Leveraging Large Language Models to Support Extended Metaphor Creation for Science Writing. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 115–135.

[55] Jeongyeon Kim, Sangho Suh, Lydia B Chilton, and Haijun Xia. 2023. Metaphorian: Leveraging Large Language Models to Support Extended Metaphor Creation for Science Writing. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (<conf-loc>, <city>Pittsburgh</city>, <state>PA</state>, <country>USA</country>, </conf-loc>) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 115–135. https://doi.org/10.1145/3563657.3595996

[56] Nam Wook Kim, Grace Myers, and Benjamin Bach. 2023. How Good is ChatGPT in Giving Advice on Your Visualization Design? *arXiv preprint arXiv:2310.09617* (2023).

[57] Nam Wook Kim, Grace Myers, and Benjamin Bach. 2024. How Good is ChatGPT in Giving Advice on Your Visualization Design? arXiv:2310.09617 [cs.HC]

[58] Sang Soo Kim. 2019. Exploitation of shared knowledge and creative behavior: the role of social context. *Journal of Knowledge Management* 24 (12 2019), 279–300. https://doi.org/10.1108/JKM-10-2018-0611

[59] Tae Soo Kim, Yoonjoo Lee, Minsuk Chang, and Juho Kim. 2023. Cells, Generators, and Lenses: Design Framework for Object-Oriented Interaction with Large Language Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (<conf-loc>, <city>San Francisco</city>, <state>CA</state>, <country>USA</country>, </conf-loc>) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 4, 18 pages. https://doi.org/10.1145/3586183.3606833

[60] Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2023. EvalLM: Interactive Evaluation of Large Language Model Prompts on User-Defined Criteria. arXiv:2309.13633 [cs.HC]

[61] René F Kizilcec and Emily Schneider. 2015. Motivation as a lens to understand online learners: Toward data-driven design with the OLEI scale. *ACM Transactions on Computer-Human Interaction (TOCHI)* 22, 2 (2015), 1–24.

[62] Hyung-Kwon Ko, Gwanmo Park, Hyeon Jeon, Jaemin Jo, Juho Kim, and Jinwook Seo. 2023. Large-scale Text-to-Image Generation Models for Visual Artists' Creative Works. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)*. ACM. https://doi.org/10.1145/3581641.3584078

[63] Janin Koch, Andrés Lucero, Lena Hegemann, and Antti Oulasvirta. 2019. May AI? Design ideation with cooperative contextual bandits. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[64] Alexander Kwan. 2023. *Piniverse: A Versatile AI-Powered Framework for Streamlining Digital Content Interaction and Empowering User Creativity*. Ph. D. Dissertation. WORCESTER POLYTECHNIC INSTITUTE.

[65] Elisa Kwon, Vivek Rao, and Kosa Goucher-Lambert. 2023. Understanding inspiration: Insights into how designers discover inspirational stimuli using an AI-enabled platform. *Design Studies* 88 (2023), 101202.

[66] Dan Lahav, Jon Saad Falcon, Bailey Kuehl, Sophie Johnson, Sravanthi Parasa, Noam Shomron, Duen Horng Chau, Diyi Yang, Eric Horvitz, Daniel S Weld, et al. 2022. A search engine for discovery of scientific challenges and directions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 11982–11990.

[67] BRENDAN LE, LUIS A LEIVA, HAQ AMAN, ARI KINTISCH, GABRIELLE BUFREM, and JEFF HUANG. 2020. Sketchy: Drawing Inspiration from the Crowd. (2020).

[68] Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM. https://doi.org/10.1145/3491102.3502030

[69] Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Hong-in Lee, and Moontae Lee. 2023. QASA: advanced question answering on scientific articles. In *International Conference on Machine Learning*. PMLR, 19036–19052.

[70] Florian Lehmann, Niklas Markert, Hai Dang, and Daniel Buschek. 2022. Suggestion lists vs. continuous generation: Interaction design for writing with generative models on mobile devices affect text length, wording and perceived authorship. In *Proceedings of Mensch und Computer 2022*. 192–208.

[71] Jan Marco Leimeister, Michael Huber, Ulrich Bretschneider, and Helmut Krcmar. 2009. Leveraging crowdsourcing: activation-supporting components for IT-based ideas competition. *Journal of management information systems* 26, 1 (2009), 197–224.

[72] Xin-Ye Li, Jiang-Tian Xue, Zheng Xie, and Ming Li. 2023. Think Outside the Code: Brainstorming Boosts Large Language Models in Code Generation.

[73] Yuyu Lin, Jiahao Guo, Yang Chen, Cheng Yao, and Fangtian Ying. 2020. It is your turn: Collaborative ideation with a co-creative robot through sketch. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[74] Vivian Liu, Han Qiao, and Lydia Chilton. 2022. Opal: Multimodal image generation for news illustration. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–17.

[75] Yiren Liu, Si Chen, Haocong Chen, Mengxia Yu, Xiao Ran, Andrew Mo, Yiliu Tang, and Yun Huang. 2023. How AI Processing Delays Foster Creativity: Exploring Research Question Co-Creation with an LLM-based Agent. *arXiv preprint arXiv:2310.06155* (2023).

[76] Ryan Louie. 2023. *Human-AI Interface Layers: Enhancing Communication of Intent for AI-Assisted Creative Pursuits and Social Experiences*. Ph. D. Dissertation. Northwestern University.

[77] Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J Cai. 2020. Novice-AI music co-creation via AI-steering tools for deep generative models. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[78] Ioanna Lykourentzou, Faez Ahmed, Costas Papastathis, Irwyn Sadien, and Konstantinos Papangelis. 2018. When crowds give you lemons: Filtering innovative ideas using a diverse-bag-of-lemons strategy. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–23.

[79] Brenda Massetti. 1996. An empirical examination of the value of creativity support systems on idea generation. *MIS quarterly* (1996), 83–97.

[80] Serena Mastria, Sergio Agnoli, Giovanni Emanuele Corazza, Michele Grassi, and Laura Franchin. 2023. What inspires us? An experimental analysis of the semantic meaning of irrelevant information in creative ideation. *Thinking & Reasoning* 29, 4 (2023), 698–725.

[81] Ravi Mehta, Darren W Dahl, and Rui Zhu. 2017. Social-recognition versus financial incentives? Exploring the effects of creativity-contingent external rewards on creative performance. *Journal of Consumer Research* 44, 3 (2017), 536–553.

[82] Lucas Memmert, Izabel Cvetkovic, and Eva A. C. Bittner. 2023. Human-AI Collaboration in Conceptualizing Design Science Research studies: Perceived Helpfulness of Generative Language Model's Suggestions. In *31st European Conference on Information Systems - Co-creating Sustainable Digital Futures, ECIS 2023, Kristiansan, Norway, June 11-16, 2023*, Margunn Aanestad, Stefan Klein, Monideepa Tarafdar, Shengnan Han, Sven Laumer, and Isabel Ramos (Eds.). https://aisel.aisnet.org/ecis2023_rp/405

[83] Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. 2022. Co-Writing Screenplays and Theatre Scripts with Language Models: An Evaluation by Industry Professionals. CoRR abs/2209.14958 (2022), 102 pages.

[84] Simone Mora, Francesco Gianni, and Monica Divitini. 2017. Tiles: a card-based ideation toolkit for the internet of things. In *Proceedings of the 2017 conference on designing interactive systems*. 587–598.

[85] Elizabeth A Nick, David A Cole, Sun-Joo Cho, Darcy K Smith, T Grace Carter, and Rachel L Zelkowitz. 2018. The online social support scale: measure development and validation. *Psychological assessment* 30, 9 (2018), 1127.

[86] Ibukun Olatunji. 2023. Why try to build try to build a co-creative poetry system that makes people feel that they have "creative superpowers"? (2023).

[87] Srishti Palani, Aakanksha Naik, Doug Downey, Amy X Zhang, Jonathan Bragg, and Joseph Chee Chang. 2023. Relatedly: Scaffolding Literature Reviews with Existing Related Work Sections. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

[88] Marios Papachristou, Longqi Yang, and Chin-Chia Hsu. 2024. Leveraging Large Language Models for Collective Decision-Making. arXiv:2311.04928 [cs.CL]

[89] Jooyeon Park and Chang Soo Sung. 2023. The impact of generative AI tools on the development of entrepreneurial career intentions. (2023).

[90] Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. 2023. Anglekindling: Supporting journalistic angle ideation with large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[91] Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. 2023. AngleKindling: Supporting Journalistic Angle Ideation with Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 225, 16 pages. https://doi.org/10.1145/3544548.3580907

[92] Savvas Petridis, Ben Wedin, James Wexler, Aaron Donsbach, Mahima Pushkarna, Nitesh Goyal, Carrie J. Cai, and Michael Terry. 2023. ConstitutionMaker: Interactively Critiquing Large Language Models by Converting Feedback into Principles. arXiv:2310.15428 [cs.HC]

[93] Alain Pinsonneault, Henri Barki, R Brent Gallupe, and Norberto Hoppen. 1999. Electronic brainstorming: The illusion of productivity. *Information Systems Research* 10, 2 (1999), 110–133.

[94] Mohi Reza, Nathan Laundry, Ilya Musabirov, Peter Dushniku, Zhi Yuan ""Michael"" Yu, Kashish Mittal, Tovi Grossman, Michael Liut, Anastasia Kuzminykh, and Joseph Jay Williams. 2023. ABScribe: Rapid Exploration of Multiple Writing Variations in Human-AI Co-Writing Tasks using Large Language Models. arXiv:2310.00117 [cs.HC]

[95] Jeba Rezwana. 2023. *Towards designing engaging and ethical human-centered AI partners for human-AI co-creativity*. Ph. D. Dissertation. The University of North Carolina at Charlotte.

[96] Sara Rosengren, Micael Dahlén, and Erik Modig. 2013. Think outside the ad: Can advertising creativity benefit more than the advertiser? *Journal of advertising* 42, 4 (2013), 320–330.

[97] Lisa DaVia Rubenstein, Gregory L Callan, and Lisa M Ridgley. 2018. Anchoring the creative process within a self-regulated learning framework: Inspiring assessment methods and future research. *Educational Psychology Review* 30 (2018), 921–945.

[98] Ruben Schlagowski, Fabian Wildgrube, Silvan Mertes, Ceenu George, and Elisabeth André. 2022. Flow with the beat! Human-centered design of virtual environments for musical creativity support in VR. In *Proceedings of the 14th Conference on Creativity and Cognition*. 428–442.

[99] Johannes Schleith, Milda Norkute, Mary Mikhail, and Daniella Tsar. 2022. Cognitive Strategy Prompts: Creativity Triggers for Human Centered AI Opportunity Detection. In *Proceedings of the 14th Conference on Creativity and Cognition*. 29–37.

[100] Yang Shi, Nan Cao, Xiaojuan Ma, Siji Chen, and Pei Liu. 2020. Emog: Supporting the sketching of emotional expressions for storyboarding. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[101] Yang Shi, Yang Wang, Ye Qi, John Chen, Xiaoyao Xu, and Kwan-Liu Ma. 2017. IdeaWall: Improving creative collaboration through combinatorial visual stimuli. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 594–603.

[102] Ben Shneiderman. 2007. Creativity support tools: accelerating discovery and innovation. *Commun. ACM* 50, 12 (2007), 20–32.

[103] 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*. 937–945.

[104] Pao Siangliulue, Joel Chan, Krzysztof Z Gajos, and Steven P Dow. 2015. Providing timely examples improves the quantity and quality of generated ideas. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*. 83–92.

[105] Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L Glassman. 2023. Where to hide a stolen elephant: Leaps in creative writing with multimodal machine intelligence. *ACM Transactions on Computer-Human Interaction* 30, 5 (2023), 1–57.

[106] Andrew T Stephen, Peter Pal Zubcsek, and Jacob Goldenberg. 2016. Lower connectivity is better: The effects of network structure on redundancy of ideas and customer innovativeness in interdependent ideation tasks. *Journal of Marketing Research* 53, 2 (2016), 263–279.

[107] Sangho Suh, Sydney Lamorea, Edith Law, and Leah Zhang-Kennedy. 2022. PrivacyToon: Concept-driven Storytelling with Creativity Support for Privacy Concepts. In *Designing Interactive Systems Conference*. 41–57.

[108] Sangho Suh, Jian Zhao, and Edith Law. 2022. Codetoon: Story ideation, auto comic generation, and structure mapping for code-driven storytelling. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–16.

[109] Simon Taggar. 2002. Individual creativity and group ability to utilize individual creative resources: A multilevel model. *Academy of management Journal* 45, 2 (2002), 315–330.

[110] Jaime Teevan and Lisa Yu. 2017. Bringing the wisdom of the crowd to an individual by having the individual assume different roles. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition*. 131–135.

[111] Todd M Thrash and Andrew J Elliot. 2003. Inspiration as a psychological construct. *Journal of personality and social psychology* 84, 4 (2003), 871.

[112] Olivier Toubia and Oded Netzer. 2017. Idea generation, creativity, and prototypicality. *Marketing science* 36, 1 (2017), 1–20.

[113] Luis A Vasconcelos and Nathan Crilly. 2016. Inspiration and fixation: Questions, methods, findings, and challenges. *Design Studies* 42 (2016), 1–32.

[114] Qian Wan and Zhicong Lu. 2023. GANCollage: A GAN-Driven Digital Mood Board to Facilitate Ideation in Creativity Support. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 136–146.

[115] Boheng Wang, Haoyu Zuo, Zebin Cai, Yuan Yin, Peter Childs, Lingyun Sun, and Liuqing Chen. 2023. A Task-Decomposed AI-Aided Approach for Generative Conceptual Design. https://doi.org/10.1115/DETC2023-109087

[116] Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2023. SciMON: Scientific Inspiration Machines Optimized for Novelty. *arXiv preprint arXiv:2305.14259* (2023).

[117] Shijuan Wang and Masao Murota. 2016. Possibilities and limitations of integrating peer instruction into technical creativity education. *Instructional Science* 44 (2016), 501–525.

[118] Sitong Wang, Zheng Ning, Anh Truong, Mira Dontcheva, Dingzeyu Li, and Lydia B Chilton. 2023. PodReels: Human-AI Co-Creation of Video Podcast Teasers. *arXiv preprint arXiv:2311.05867* (2023).

[119] 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*. 1–19.

[120] Yunlong Wang, Shuyuan Shen, and Brian Y Lim. 2023. RePrompt: Automatic Prompt Editing to Refine AI-Generative Art Towards Precise Expressions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM. https://doi.org/10.1145/3544548.3581402

[121] Yunlong Wang, Priyadarshini Venkatesh, and Brian Y Lim. 2022. Interpretable Directed Diversity: Leveraging Model Explanations for Iterative Crowd Ideation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–28.

[122] Irving B Weiner. 2003. *Handbook of psychology, history of psychology*. Vol. 1. John Wiley & Sons.

[123] Haiyang Yang, Amitava Chattopadhyay, Kuangjie Zhang, and Darren W Dahl. 2012. Unconscious creativity: When can unconscious thought outperform conscious thought? *Journal of Consumer Psychology* 22, 4 (2012), 573–581.

[124] Kexin Bella Yang, Tomohiro Nagashima, Junhui Yao, Joseph Jay Williams, Kenneth Holstein, and Vincent Aleven. 2021. Can crowds customize instructional materials with minimal expert guidance? Exploring teacher-guided crowdsourcing for improving hints in an ai-based tutor. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–24.

[125] Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2023. Flask: Fine-grained language model evaluation based on alignment skill sets. *arXiv preprint arXiv:2307.10928* (2023).

[126] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story Writing With Large Language Models. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 841–852. https://doi.org/10.1145/3490099.3511105

[127] Chao Zhang, Cheng Yao, Jiayi Wu, Weijia Lin, Lijuan Liu, Ge Yan, and Fangtian Ying. 2022. StoryDrawer: A Child–AI Collaborative Drawing System to Support Children's Creative Visual Storytelling. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.

[128] Gaoxia Zhu, Xiuyi Fan, Chenyu Hou, Tianlong Zhong, Peter Seow, Annabel Chen Shen-Hsing, Preman Rajalingam, Low Kin Yew, and Tan Lay Poh. 2023. Embrace Opportunities and Face Challenges: Using ChatGPT in Undergraduate Students' Collaborative Interdisciplinary Learning. *arXiv preprint arXiv:2305.18616* (2023).