Working Paper 24-005

The Crowdless Future? Generative AI and Creative Problem Solving

Léonard Boussioux
Jacqueline N. Lane
Miaomiao Zhang
Vladimir Jacimovic
Karim R. Lakhani

The Crowdless Future? Generative AI and Creative Problem Solving

Léonard Boussioux
University of Washington

Jacqueline N. Lane
Harvard Business School

Miaomiao Zhang
Harvard Business School

Vladimir Jacimovic
Harvard Business School
Continuum Labs

Karim R. Lakhani
Harvard Business School

Working Paper 24-005

We are grateful to Harvard Business School Research Associate Justin Ho, Program for Research in Markets & Organizations (PRIMO) fellow Stella Jia, who supported the data analysis for this project as well as Laboratory for Innovation Science at Harvard (LISH) lab manager Kate Powell who offered oversight and coordination for the research protocols.

Funding for this research was provided in part by Harvard Business School.

The Crowdless Future? Generative AI and Creative Problem Solving

Léonard Boussioux[1*], Jacqueline N. Lane[2*], Miaomiao Zhang[2],
Vladimir Jacimovic[2,3], & Karim R. Lakhani[2]

[1]University of Washington, Michael G. Foster School of Business; leobix@uw.edu
[2]Harvard Business School; jnlane@hbs.edu
[2]Harvard Business School; mzhang@hbs.edu
[3]ContinuumLab.AI; vladimir@continuumlab.ai
[2]Harvard Business School; klakhani@hbs.edu

*Léonard Boussioux and Jacqueline N. Lane share co-first authorship

This version has been accepted for publication in Organization Science. Please note it has not gone through final type-setting or copyediting by the journal.

Abstract

The rapid advances in generative artificial intelligence (AI) open up attractive opportunities for creative problem-solving through human-guided AI partnerships. To explore this potential, we initiated a crowdsourcing challenge focused on sustainable, circular economy business ideas generated by the human crowd and collaborative human-AI efforts using two alternative forms of solution search. The challenge attracted 125 global solvers from various industries, and we used strategic prompt engineering to generate the human-AI solutions. We recruited 300 external human evaluators to judge a randomized selection of 13 out of 234 solutions, totaling 3,900 evaluator–solution pairs. Our results indicate that while human crowd solutions exhibited higher novelty—both on average and for highly novel outcomes—human-AI solutions demonstrated superior strategic viability, financial and environmental value, and overall quality. Notably, human-AI solutions co-created through *differentiated* search, where human-guided prompts instructed the large language model (LLM) to sequentially generate outputs distinct from previous iterations, outperformed solutions generated through *independent* search. By incorporating "AI-in-the-loop" into human-centered creative problem-solving, our study demonstrates a scalable, cost-effective approach to augment the early innovation phases and lays the groundwork for investigating how integrating human-AI solution search processes can drive more impactful innovations.

Keywords: Generative AI, Large Language Models, creative problem-solving, organizational search, AI-in-the-loop, crowdsourcing, prompt engineering

Acknowledgments: We extend our heartfelt gratitude to Justin Ho, Stella Jia, and Kate Powell for their invaluable research assistance. Our work has been significantly enriched by the insightful comments and feedback from Charles Ayoubi, Nathan Rietzler, and seminar participants at various institutions, including the Laboratory for Innovation Science at Harvard (LISH), University of Washington Foster School of Business, GenAI Lab at the Professorship of Digital Marketing at the TUM School of Management, University of Toronto Rotman School of Management, MIT Sloan, Questrom School of Business Online Research on Digital Businesses, and Harvard Human-Computer Interaction groups. We are profoundly grateful to Editor-in-Chief Lamar Pierce for his exceptional guidance and support throughout the review process. His leadership and insights have been instrumental in elevating the quality of our work. We also extend our sincere appreciation to the three anonymous reviewers whose constructive feedback and thoughtful suggestions have substantially improved our manuscript. Their dedication and expertise have been invaluable in shaping this research. This work was made possible through the generous financial support from the HBS Division of Research and Faculty Development (DFRD) and LISH, for which we are deeply thankful. We acknowledge the use of GPT-4 and Claude-3 for writing assistance. Any remaining errors are our own responsibility.

Introduction

Organizations increasingly integrate artificial intelligence (AI) technologies into their work processes (Iansiti and Lakhani 2020), leveraging computational capabilities in identifying patterns (Choudhury et al. 2021), making predictions (Agrawal et al. 2018, Kim et al. 2024), and decision-making (Allen and Choudhury 2022, Kleinberg et al. 2018). These technological advancements have enabled AI to surpass human capabilities in a range of settings, such as chess (Silver et al. 2016), medical advice (Ayers et al. 2023), and talent management (Li et al. 2020, Tong et al. 2021). Although AI can perform exceptionally well in tasks with clear rules, patterns, and objectives (Lou and Wu 2021, Miric et al. 2023), it is less clear whether AI can aid in creative problem-solving tasks, which often require abstract, nuanced, and iterative thinking (Amabile 1983), social interactions (Fleming et al. 2007, Perry-Smith 2006, Wuchty et al. 2007), and broad search for distant knowledge and alternative perspectives (Jeppesen and Lakhani 2010, Katila and Ahuja 2002). This paper explores how generative AI—a type of AI technology capable of producing new content, such as text, images, audio, or video, based on patterns learned from existing data—can enhance creative problem-solving through human-guided AI partnerships. We propose two alternative modes of human-AI search against human crowdsourcing to investigate their respective abilities to generate novel, valuable, and high-quality solutions.

Creative problem-solving involves the generation of novel and valuable ideas (Amabile 1983, Leiponen and Helfat 2010). Novel solutions are original ideas that depart from existing knowledge, and valuable solutions are useful ideas that can be implemented to yield economic and social returns (Baumol 1993, Kaplan and Vakili 2015, Teodoridis et al. 2019). Yet, innovative activity is highly risky, and there can often be uncertainty regarding the best approach or path to solve a problem (Katila and Ahuja 2002, Laursen and Salter 2006, Leiponen and Helfat 2010). This uncertainty may be heightened when the problem draws upon multiple domains (Boudreau et al. 2011), is complex, or ill-structured (Nickerson and Zenger 2004, Simon 1973). Although the ability to generate and manage ideas is central to a firm's technological and competitive advantage (Hargadon and Bechky 2006, Van de Ven 1986), many organizations are constrained from innovating due to limited cognitive resources (Ocasio 1997, Rhee and Leonardi 2018), entrenched mental models (Barr et al. 1992), financial and social costs (Becker 1994, Glaeser et al. 2002), and organizational inertia (Tripsas 2009).

Creative problem-solving, often viewed as a search process for solutions, is a critical step in the innovation process (Benner and Tushman 2003, Katila and Ahuja 2002, March 1991). Firms aiming to enhance their chances of successful problem-solving can adopt a parallel path strategy, which utilizes various approaches to search for solutions. By exploring a wider range of potential solutions simultaneously, this parallel path strategy allows firms to expand the breadth of their solution search (Abernathy and Rosenbloom 1969, Leiponen and Helfat 2010, Nelson 1961). Crowd-based creative problem-solving increases the number of parallel paths by engaging multiple independent problem solvers possessing diverse knowledge and alternative methods (Boudreau et al. 2011, Jeppesen and Lakhani 2010). The recent advances in generative AI (Achiam et al. 2023, Bubeck et al. 2023, Wang et al. 2024) open up unprecedented opportunities to explore multiple parallel paths at relatively low costs, to increase the chances of achieving a high-quality outcome. These developments introduce a novel approach to creative problem-solving that fosters a collaborative partnership between humans and AI.

Generative AI systems, built by training complex algorithms on vast amounts of public and private data[1], are now accessible through user-friendly conversational interfaces. These systems offer cost-effective and efficient ways to generate a wide range of creative ideas. Human collaborators can prompt the models to produce and simulate diverse perspectives, broadening solution search at an unparalleled scale for just a few dollars (Girotra et al. 2023). Its capacity for delivering numerous cost-effective outcomes on demand and consistently throughout substantial workloads holds promise for augmenting organizational creative problem-solving (Dell'Acqua et al. 2023, Noy and Zhang 2023). In contrast, although crowdsourcing has previously been a viable solution to reduce costs and harness productivity gains

[1] For example, GPT-4, one of OpenAI's latest language models, is speculated to have 1.8 trillion parameters across 120 layers, approximately 10 times larger than GPT-3, potentially utilizing a mixture of experts (MoE) architecture with 16 expert models of around 111 billion parameters each. Its pre-training allegedly required an immense 2.15×10^{25} FLOPS, necessitating 25,000 A100 GPUs running for 90 to 100 days at an estimated hardware cost of $63 million, with the pre-training compute cost projected to be around $22 million; see https://www.semianalysis.com/p/gpt-4-architecture-infrastructure for details.

compared to internal methods (Paik et al. 2020), it has limitations (Piezunka and Dahlander 2019). In particular, crowdsourcing can require extensive planning and incur expenses of hundreds of thousands of dollars (Paik et al. 2020), and it can be difficult to manage the competing participant effects between incentives and efforts (Boudreau et al. 2011, 2016, Che and Gale 2003, Taylor 1995, Terwiesch and Xu 2008).

In this paper, we examine the effectiveness of collaborative problem-solving between humans and AI by comparing the novelty, value, and quality of solutions crowdsourced from humans to those generated by an individual strategically prompting a large language model (LLM). LLMs are a subset of AI designed to understand and produce text based on extensive training from published texts (Bubeck et al. 2023). Most generative AI studies using LLMs in organizational settings have focused on investigating the productivity effects of these technologies in the workplace (Brynjolfsson et al. 2023, Dell'Acqua et al. 2023, Noy and Zhang 2023, Otis et al. 2024). Moreover, recent research examining the impact of AI on creativity often focuses on well-understood domains (Girotra et al. 2023, Gómez-Rodríguez and Williams 2023, Guzik et al. 2023, Wang et al. 2024) and is typically conducted in controlled laboratory settings (Doshi and Hauser 2023, Hagendorff et al. 2023, Koivisto and Grassini 2023).

To understand how humans working alongside AI can shape the future of creative problem-solving, it is critical to further investigate their joint potential in real-world field settings and with complex, open-ended problems. We partnered with Continuum Lab, an AI firm, to develop a crowdsourcing challenge about new business ideas on the circular economy. Our study involved 234 human crowd (HC) and human-AI (HAI) solutions, evaluated by 300 external human judges, totaling 3,900 evaluator–solution pairs. We used different human prompt engineering techniques to enable alternative forms of human-AI partnerships for solution search and to demonstrate the range of capabilities of HAI approaches for creative problem-solving. Our findings indicate that whereas the HC solutions exhibit a higher level of novelty on average and at the upper end of the rating distribution, the HAI solutions are rated as higher in value regarding their strategic viability for successful implementation, as well as environmental and financial value. Overall, we find that the HAI solutions are rated higher on average in quality than the HC solutions.

Moreover, we investigate how different forms of HAI search processes impact the solutions' outputs. Specifically, we examine two modes of HAI collaboration, *independent* and *differentiated* search, that vary regarding the degree of human-prompted feedback to guide the LLM's search for solutions. Our analysis demonstrates that including human-crafted differentiation instructions that iteratively prompt the LLM to diversify each successive response effectively enhances the novelty of the outputs without compromising their value, resulting in higher overall quality and demonstrating the salience of human-led and AI-augmented creative problem-solving.

Our study contributes to the emerging literature on human-AI collaboration (Allen and Choudhury 2022, Anthony et al. 2023, Choudhary et al. 2023, Dell'Acqua et al. 2023, Lebovitz et al. 2022, Raisch and Fomina 2023) and offers insights into effectively integrating human and AI to solve innovative, open-ended problems at scale. Our work extends our understanding of how humans and AI agents can collaborate, moving from routine decision-making to solving new problems with untested solutions (Jeppesen and Lakhani 2010, Raisch and Fomina 2023). Our findings suggest that human prompt engineering to guide language models in generating creative outputs is a promising approach for adopting "AI-in-the-loop" workflows for creative problem-solving. We illustrate that human-AI approaches can efficiently produce novel and valuable outputs at minimal costs, facilitated by human guidance of the LLM's exploratory solution search space. Although the specific form or division of labor in human-AI collaboration will evolve with technological advances, this paper illustrates an adaptable framework for strategically integrating generative AI into creative problem-solving.

Creative Problem Solving and Human-AI Solution Search

Creative problem-solving can be conceived as a search for solutions on a landscape (Fleming and Sorenson 2001, Katila and Ahuja 2002, Levinthal 1997). This landscape contains peaks of exceptional opportunities and valleys with limited ones (Levinthal 1997). Most solvers tend to search locally, explore familiar neighborhoods, near previous successful solutions, and seek incremental improvements (Cyert and March 1963, Nelson and Winter 1982). However, some solvers may venture into uncharted areas, exploring solutions that deviate from existing ones to explore more distant areas of the solution landscape, potentially unlocking more innovative possibilities (Kaplan and Vakili 2015). This aligns with the observation that greater problem-solving success occurs when firms broaden their search for knowledge across various technological domains and geographic locations (Kneeland et al. 2020, Leiponen and Helfat 2010).

When the best method for solving a problem is uncertain, one strategy to enhance innovative search is to utilize a variety of different approaches, or "parallel paths," as this breadth can improve overall solution quality (Abernathy and Rosenbloom 1969, Leiponen and Helfat 2010, Nelson 1961). The parallel path effect suggests that developing multiple solutions to the same problem increases the likelihood of achieving a high-quality outcome (Boudreau et al. 2011, Dahan and Mendelson 2001). Arguably, utilizing various approaches is particularly critical when the objective is to maximize the quality of a few top ideas instead of many average ones (Girotra et al. 2010).

Crowdsourcing contests leverage a diverse pool of solvers with differing backgrounds and experiences to increase the number of parallel paths to solve a problem and improve solution quality (Jeppesen and Lakhani 2010, Lifshitz-Assaf 2018, Piezunka and Dahlander 2015, Riedl et al. 2024). However, crowdsourcing can be resource-intensive (Piezunka and Dahlander 2019) and statistically inefficient due to the volume of low-quality submissions (Bell et al. 2024). The quest for highly creative outcomes can be further complicated by diminishing contribution effort as the size of an innovation contest grows (Boudreau et al. 2011, 2016, Che and Gale 2003, Taylor 1995, Terwiesch and Xu 2008). Hence, crowdsourcing has some drawbacks, although it has been a highly effective approach for enhancing the parallel path effect. As AI systems like LLMs have capabilities that differ from those of humans, they promise to develop new forms of creative problem-solving (Raisch and Fomina 2023) and complement human intelligence (Choudhary et al. 2023).

Using LLMs to Advance Human-AI Creative Problem Solving

LLMs offer a new way of augmenting creative problem-solving for organizations. These models, trained on extensive data corpora, provide problem-solvers with unprecedented capabilities for interactive collaboration. One critical way to collaborate with LLMs is via strategic prompt engineering, where humans and AI explore the search space together, guided by carefully crafted human instructions. This study examines independent and differentiated search strategies, emphasizing the role of strategic prompt engineering in reducing the cost and improving the quality of outputs in human-AI collaborative search relative to traditional human crowdsourcing methods.

Technical Primer. AI is a broad field within computer science that seeks to create systems capable of performing tasks that typically require human intelligence. This includes activities such as learning, reasoning, problem-solving, perception, and understanding language. Machine Learning (ML), a subset of AI, focuses on algorithms that allow machines to analyze data, learn from it, and make predictions. Unlike traditional programming, ML models evolve their performance as they process more data, eliminating the need for explicit programming in every scenario.

Generative AI falls under the umbrella of ML and represents an approach where machines can generate new content or data that is similar but not necessarily identical to what they have been trained on. This can include anything from generating text, computer code, and music to creating images and videos. Generative AI leverages ML models, such as neural networks, trained on large datasets to produce outputs that mirror the input data distribution. LLMs are a type of generative AI specifically designed to process and generate human language. They are trained on vast corpora of textual data from the internet, books, and other text-based sources, allowing them to learn the intricacies of human language, including grammar, syntax, semantics, and context. This technical primer focuses on autoregressive LLMs, the foundation for models like ChatGPT, Gemini, and Claude. The training process of autoregressive LLMs involves several key components:

1. Tokenization: The input text is divided into smaller units called tokens, which can be words, subwords, or characters. This process allows the model to process the text more efficiently.
2. Embedding: Each token is mapped to a high-dimensional vector representation, capturing the semantic and syntactic relationships between tokens. This embedding layer allows the model to understand the meaning and context of words.
3. Transformer Architecture: LLMs commonly use transformer neural networks (Vaswani et al. 2017). Transformers utilize self-attention mechanisms, which allow the model to weigh the

importance of different tokens within a sequence, enabling it to capture long-range dependencies and context more effectively (Ash and Hansen 2023, Bahdanau et al. 2014).

4. Autoregressive Language Modeling: This approach trains the model to predict the next token in a sequence based on all preceding tokens. The model learns to generate text by iteratively predicting each subsequent token, conditioned on the previously generated ones. This process leverages self-supervised learning, where language comprehension is refined by predicting future elements of the text.
5. Optimization: The model's parameters are updated through an iterative process called gradient descent, which minimizes the difference between the model's predictions and the actual next tokens in the training data. This process allows the model to learn the patterns and relationships within the language. Due to a large number of parameters and complexity, optimizing LLMs requires substantial computational resources, particularly GPU-based computing power (Achiam et al. 2023).
6. Fine-tuning and Alignment: State-of-the-art LLMs are typically further refined through supervised learning, where the models are provided with human-annotated datasets that exemplify desired behaviors or task-specific outputs. Additionally, alignment and guardrail techniques are employed to promote helpful, safe, and human-aligned responses, ensuring that the models adhere to ethical guidelines and produce appropriate content (Bai et al. 2022, Ouyang et al. 2022).

During the text generation process, autoregressive LLMs use the learned patterns and relationships to calculate the probability distribution of the next token based on the input context. The model then samples a token from this distribution, with the sampling process influenced by the temperature parameter. A higher temperature leads to more diverse and creative outputs by allowing for the selection of lower probability tokens, while a lower temperature results in more deterministic and conservative generations by favoring high probability tokens (Bellemare-Pepin et al. 2024).

Strategic Prompt Engineering. Prompt engineering, the process of designing input prompts to guide the model's output (Brown et al. 2020), plays a crucial role in shaping the generated text (Battle and Gollapudi

2024, OpenAI 2024). As LLMs currently lack independent agency, the quality and relevance of their outputs heavily depend on humans' ability to skillfully craft prompts, emphasizing the necessary collaboration between humans and AI (Zamfirescu-Pereira et al. 2023). By carefully crafting prompts that provide context, instructions, or examples, humans can influence the output probability distribution, steering the model's output toward desired topics, styles, or formats. Effective prompt engineering is essential for aligning the model's outputs with specific tasks or domains, as it helps to constrain the vast space of possible generations to more relevant and coherent outputs. This "AI-in-the-loop" integration enables a synergistic HAI collaboration that can push the boundaries of traditional problem-solving approaches to produce creative solutions.

Anticipated Cost-Benefit Implications. Our study uses OpenAI's Generative Pretrained Transformer 4 (GPT-4), a representative example of advanced language models that operate based on similar foundational principles (see Appendix A for a detailed overview of the inference processes of LLMs). The advanced capabilities of LLMs, such as GPT-4, indicate a strong potential for application in creative problem-solving. Notably, using LLMS may streamline idea generation, making it a more cost-effective and efficient approach (Girotra et al. 2023). Unlike human participants, who typically require monetary or non-pecuniary incentives to engage in crowdsourcing contests (Jeppesen and Lakhani 2010, Terwiesch and Ulrich 2009), LLMs can continuously generate outputs for creative tasks without additional incentives. Moreover, LLMs can rapidly generate consistent solutions at a larger scale, substantially enriching the idea pool in much less time than conventional human crowdsourcing methods.

Human-AI Collaboration and the Production of Novel and Valuable Outputs. The interactive nature of LLMs, which enables humans to engage in conversations through personalized textual prompts, allows for novel forms of collaboration through the division of labor between humans and AI systems (Choudhary et al. 2023). Scholars have begun conceptualizing alternative forms of human-AI creative problem-solving, in which humans and AI can search together for creative outputs, for instance, along the dimensions of specialization of agents and the sequencing of tasks (Choudhary et al. 2023, He et al. 2023, Jia et al. 2023, Raisch and Fomina 2023). In this study, we draw upon these perspectives and explore two forms of HAI collaboration with LLMs for creative problem-solving: independent and differentiated search. These two forms of HAI solution search are depicted in Figure 1.

Although many formats of HAI collaboration can be envisioned, a key differentiating factor is the degree to which humans are involved in steering how the LLM searches for solutions. Due to the fundamental mechanism of LLMs, which involves calculating the probability distribution of the next word or token based on the input context and sampling from this distribution (Bubeck et al. 2023), LLMs tend to reflect more mainstream ideas unless otherwise directed (Anderson et al. 2024). Independent and differentiated search are two alternative approaches to elicit more creative responses.

With *independent search,* as shown in Figure 1, humans define the problem and provide an initial prompt, allowing the LLM to independently generate a potential solution through its own broad search capabilities by leveraging the LLM's training on the immense scope of data across diverse domains (Raisch and Fomina 2023). As an illustrative example, a human prompt like "Generate a creative solution for a new type of sustainable urban transportation" might solicit responses from the LLM, such as "A network of small, solar-powered pods running on elevated tracks above city streets" and "Sidewalks and bike paths that generate electricity through the kinetic energy of pedestrians and cyclists" which resemble independent contributions, such as what we might expect from the human crowd.

For independent search, the initial framing and scope of the exploration critically depend on the human's strategic guidance. Consequently, the human's initial guidance is crucial in broadly defining the scope and framing of the LLM's search process. For example, a role-playing prompt like "As a time traveler from 2100, you've seen the incredible advancements in sustainable urban transportation. Share with a city planner in 2025 the most groundbreaking and eco-friendly transportation solution from your era that has transformed city life. Generate a creative solution for a new type of sustainable urban transportation based on this futuristic insight" may provide solutions such as "elevated hyperloop tunnels consisting of elevated, vacuum-sealed tunnels that use magnetic levitation to transport passenger pods at high speeds" or "an electric water shuttle system that uses electric-powered shuttles that glide smoothly across waterways." The strategic guidance may allow the LLM to draw on diverse domains in its training data to make broader, more distant searches. The LLM's search may then be instructed to generate solutions that make substantial conceptual departures, where the human's prompting skill and initial guidance determine the overall direction and constraints of this expansive exploration.

Another approach is *differentiated search*, where, as illustrated in Figure 1, humans may provide an initial prompt and then insert differentiation instructions after each output to encourage the model to diversify its successive responses and explore a broader solution space. For example, the human may provide an initial prompt like "Generate creative solutions for a new type of sustainable urban transportation" and then after each output from the LLM, add the instruction "Make sure to tackle a different problem than the previous ones and propose a different solution." This iterative human guidance aims to promote solution diversity, reduce redundancy, encourage originality, and facilitate distant solution space exploration. While this process is iterative in nature, it focuses primarily on promoting solution diversity rather than in-depth refinement and collaboration between humans and AI.

--- Insert Figure 1 here ---

Although both independent and differentiated search approaches to HAI collaboration show promise for enhancing creative problem-solving, several open questions remain about the applicability of HAI partnerships to exploratory tasks involving solution search and the generation of novel solutions. These exploratory tasks differ from routine decision-making scenarios focusing on previously explored situations with known procedures and solution alternatives (Raisch and Fomina 2023), such as judge bail-or-release decisions (Kleinberg et al. 2018) or medical diagnoses (Lebovitz et al. 2022). The capabilities required for open-ended creative problem-solving may not be well-suited for HAI collaboration with current LLMs, especially when compared to the diverse perspectives offered by a broad human crowd.

First, when collaboratively searching for novel solutions, LLMs may inadvertently constrain their exploration due to their reliance on formal rationality—a decision-making mode grounded strictly in abstract rules, formal procedures, and established precepts, without nuanced consideration of contextual factors or personal perspectives (Lindebaum et al. 2020, Weber 1978). As creative problem-solving often draws inspiration from individual perspectives and situational factors (Amabile 1983, Perry-Smith 2006), there are concerns that HAI systems may be bound by their training data and constrained to searching for "myopic" or local solutions. This could lead the systems to overlook novel opportunities that necessitate conceptual leaps transcending formal rules (Kneeland et al. 2020) but may be more valuable based on their associations with past successes (Rindova and Petkova 2007).

Second, the outputs of HAI collaboration systems are limited to recombining patterns from the data used to train the LLM component. As a result, the outputs may be retrospective and ultimately confined by the specific data the LLM was exposed to during training. This contrasts with human cognition, which is inherently forward-looking and theory-based, enabling humans to transcend data and prediction to generate new data and observations and conduct experimentation (Felin and Holweg 2024, Gavetti and Levinthal 2000). These forward-looking theories guide human perception, search, and action, potentially serving as the source of highly novel recombinations, jumps, and applications (Katila and Ahuja 2002, Kneeland et al. 2020). While collaborating with AI may accelerate the generation of more incremental yet valuable solutions (Benner and Tushman 2003), pushing the boundaries towards radically new solutions may still require human expertise—particularly the collective input of the human crowd (Jeppesen and Lakhani 2010), which is unconstrained by the data-prediction modes of current AI systems (Felin and Holweg 2024).

Third, HAI outputs can exhibit failure modes, such as confabulation (generating plausible but unfounded content) and hallucination (producing outputs detached from training data or reality) (Ji et al. 2023), which may have mixed implications for creative problem-solving. On the one hand, some degree of confabulation and hallucination in the outputs could boost novelty by facilitating exploratory search that combines concepts from disparate domains in novel ways—aligning with the goals of distant search (Benner and Tushman 2003, March 1991). However, this factually incorrect knowledge could also compromise solution value by producing distorted or bizarre reflections of flawed data with limited practical value for adoption. Given the unclear potential of HAI collaboration for creative problem-solving, we investigate this question by comparing the novelty, value, and quality of HC and HAI outputs.

Research Design and Methods

Setting

Crowdsourcing Context. We partnered with Continuum Lab, an AI company, and Freelancer.com, an online marketplace, to launch a crowdsourcing challenge seeking new business ideas focused on sustainable, circular economy business opportunities. The circular economy is an economic framework that emphasizes the reuse and regeneration of materials or products to continue production in a sustainable or environmentally friendly way. Our choice of the circular economy as a backdrop for this study stems from its comprehensive scope, bridging disciplines such as environmental science, economics, design, and engineering. This interdisciplinary nature, coupled with its critical role in advancing sustainable development and addressing a range of economic, environmental, and social challenges, makes it a practical context to assess the creative problem-solving capabilities of human crowdsourcing and HAI collaboration facilitated by prompt engineering (Ivcevic and Grandinetti 2024). The challenge ran from January 30, 2023, to May 15, 2023. Participants were encouraged to submit real-life use cases of how companies can implement the circular economy concepts in their businesses. Participants were told that their ideas would be evaluated using five criteria: *Novelty*, *Strategic Viability* in terms of their feasibility and scalability of implementation, *Environmental Value*, *Financial Value*, and overall *Quality*.

All participants submitted their solutions using a Google Form. We also collected their demographic information, including their job title, geography, industry of application for their solution (a dropdown of 23 industries), and solution maturity (ideation, R&D, proof of concept, market testing, or full commercial). The contest received a total of 310 submissions. 148 participants received $10 for providing non-blank entries, and the best overall solution received a $1,000 prize. The crowdsourcing challenge had a total cost of $2,555, including a $75 platform fee. Of the 148 submissions, the research team deemed 125 eligible after filtering out off-topic or insufficiently detailed solutions.

Human-AI Collaboration and Prompt Engineering. We prompt-engineered with the GPT-4 Python API to generate various solutions to the same crowdsourcing challenge of developing sustainable, circular economy business ideas. One of the study's authors completed all prompt engineering techniques to generate the solutions using the default temperature parameter of 1.0. In this section, we describe our approach to solution generation. We first outline our use of two GPT-4 configurations to operationalize independent and differentiated search and then explain how we layer three prompt engineering techniques onto each configuration to diversify the range of the generated solutions.

We generate solutions using two alternative configurations of GPT-4: *multiple* and *single* instances. The multiple instance corresponds to independent search, where humans provide an initial prompt and allow the LLM to generate potential solutions independently. The single instance corresponds to differentiated search, where humans periodically provide instructions to differentiate the LLM's solution search. To the best of our knowledge, we are the first to report on the use and impact of these alternative configurations.

The multiple instance configuration starts from the default initialization of GPT-4 and uses a distinct instance to generate solutions. While the model and prompt remain identical across instances, the sampling methods employed in LLMs suggest that each instance can produce varied responses because the model samples from the probability distribution of possible next words or considers several high-probability next words rather than simply selecting the next word with the highest probability (See Appendix A.2 for more technical details).

The single instance configuration is an iterative prompting scheme in which the human engages in back-and-forth interactions with the LLM, inserting prompts one round after each to arrive at a desired output. More specifically, we use a single instance of GPT-4 to generate multiple solutions successively, one at a time, adding the following sentence in the context:

> *We will ask to answer these questions several times, and make sure each new answer tackles a different problem than the previous ones and proposes a different solution.*

We also add the following paragraph as a differentiation instruction each time after GPT-4 generates an output while also including all the previously GPT-4 generated solutions:

> *Make sure to tackle a different problem than the previous ones and propose a different solution. Also make sure your answers satisfy the evaluation criteria (novelty, environmental impact, financial impact, feasibility and scalability).*

By introducing a human-guided differentiation instruction between successive responses, a single instance of GPT-4 is more likely to diversify its successive responses from previous ones, enabling a potentially broader exploration of the search space than multiple instances, as LLMs tend to produce similar outputs given identical prompts (Meincke et al. 2024). The objective is to promote solution diversity, reduce redundancy, encourage originality, and facilitate distant exploration. This technique is based on "prompt-chaining" (DAIR.AI 2024), where the output of one prompt becomes the input or part of the input for the next prompt in the sequence (Saravia 2022).

Intuitively, the multiple instance configuration more closely aligns with the concept of independent crowd solvers, as each GPT-4 instance generates outputs independently, starting from a different default initialization and potentially exploring different areas of the problem and solution space. In contrast, the single instance configuration resembles an iterative process where distinct ideas are proposed successively.

We use additional prompt engineering techniques to generate solutions within the multiple and single instance configurations of GPT-4. As of mid-2023, when this research took place, techniques like one-shot or few-shot prompting (Brown et al. 2020), Chain-of-Thought processes (Wei et al. 2023), and role-playing prompts (Kong et al. 2023, Shanahan et al. 2023) were gaining traction.

Considering the evolving nature and understanding of prompt engineering, we layer on three alternative prompt engineering approaches to produce HAI solutions within the multiple and single instance configurations of GPT-4. Our baseline prompt includes the core problem description given to human solvers and a template to guide GPT-4 in answering in the same "Problem-Solution" format as the HC. This establishes a reference point for comparing HAI responses to their human counterparts, ensuring both received identical initial information. The prompt begins with the context, a concise description of the circular economy challenge, and the goal of idea generation. It then includes an example of circular economy as one-shot prompting and the different evaluation criteria accompanied by encouraging, positive, methodical wording emulating a chain-of-thought mechanism. Building upon the baseline, our second prompting approach introduces individual solver characteristics (job title, location, industry, solution maturity) in a role-playing technique for GPT-4. This approach aims to increase the contextual richness similar to human creativity, potentially enhancing the model's outputs to align better with human-produced solutions and generate more diverse creative responses. We change the individual solver characteristics described in the prompt for each new instance or single instance iteration to generate diverse solutions.

Finally, we role-play with expert, famous personas from 23 circular economy-relevant industries. This technique expands the model's input with diverse expert knowledge bases, facilitating more comprehensive information processing and the potential for generating innovative, industry-specific solutions grounded in practical applications. As previously, we change the expert personas described in the prompt for each new instance or iteration to explore different industry perspectives. Appendix B details the specific prompts used to generate the HAI solutions. The code used to generate the solutions is publicly available at https://github.com/leobix/creative.

API Costs and Time Spent. We generated 730 AI solutions, 315 each with multiple and single instances of GPT-4. Each solution was generated in 27.2 seconds on average (min = 5.9s, max = 80.8s, s.d. = 8.4s) from a Google Colab notebook and cost $0.037 on average. Hence the total direct cost of using this LLM was $27.01.

Evaluator Recruitment and Procedures

Our study (approved under Harvard University IRB23-0770) uses external human evaluators to judge the novelty, value, and quality of human and GPT-4 solutions. First, as shown in Figure 2, we recruited potential evaluators on Prolific.com in July 2023 and September 2023. For both recruitment sessions, we used a screening survey to screen potential evaluators for geographic location (US only) and age (18 years old or older), as well as for their level of interest, work experience, and knowledge of the circular economy through a multiple-choice skills test. Individuals who passed the screening filter showed at least a moderate level of interest, and either had two or more years of work experience or scored at least 60% (3 or more out of 5) on the skills test, were selected to participate in the evaluation survey (see Appendix C for survey instruments). Overall, we recruited 1,000 evaluators, of which 300 (or 30%) passed the screening survey. 145 of the 300 evaluators were from the first call and 155 from the second. We also collected demographic data on the evaluators' gender, highest level of education, field of study, and employment status.

Due to feasibility issues, such as scalability, cost, and time constraints with recruiting and managing many evaluators to review the entire set of 125 HC and 730 HAI prompt-engineered solutions, we randomly selected 234 solutions for human evaluation. Of these, 180 were HAI prompt engineered (90 single instance and 90 multiple instance), and 54 were HC submitted. We randomly selected a mix of HAI-generated responses for the HAI solutions, instructed with three alternative prompts, and evenly allocated between multiple and single instance configurations.

We used a blocked experimental design to randomize the HC and HAI solutions into distinct blocks. Each block contained ten HAI and three HC solutions, totaling 13 solutions per block. Within each block, there were five multiple instance and five single instance HAI solutions generated from the same prompt engineering approach. As shown in Figure 2, each evaluator was randomly assigned one of the 18 blocks of solutions to evaluate, i.e., evaluators were nested within solution blocks. Because prompt engineering approaches allow the LLM to explore different parts of its training, the HAI solutions generated using the same prompt engineering approach are likely to exhibit less variability than solutions generated across different prompting approaches. Therefore, our blocked randomization design choice allows for more precise comparisons between HAI and HC responses and between multiple and single instance model configurations. In other words, by strategically minimizing within-block variance, our approach ensures that each evaluator assesses HAI solutions that are more comparable than under a complete randomization design. This design choice aligns with the principle that optimal efficiency gain is achieved when the within-block variance is reduced while the between-block variance is maximized (Imbens and Rubin 2015).

Each evaluator, blind to the sources of the 13 solutions in their randomly assigned block, rated on each solution's novelty (*How different is it from existing solutions?*), strategic viability *(How likely is it to succeed and how scalable is it?*), environmental value (*How much does it benefit the planet?*), and financial value (*What financial value can it create for businesses?*), as well as overall quality (*Based on the four criteria above, what is the overall quality of the solution?*). Overall, each block was evaluated 16.67 times on average (min = 15, max = 18, s.d. = 0.88).

--- Insert Figure 2 here ---

To motivate effort and ensure thoroughness, we offered each evaluator $12 for completing the survey, with a bonus of $1 for each solution where they matched the consensus or the mode of quality rating for each solution. The mean bonus awarded was \$6.43 (s.d. = \$2.30, min = \$1, max = \$12). The total compensation per evaluator ranged from \$13 to \$24.

Variables

Dependent Variables. We use five main dependent variables, corresponding to the evaluator's *Novelty*, *Strategic Viability*, *Financial,* and *Environmental Value*, and *Quality* of each solution. To examine extreme outcomes, we create binary variables for *Top Novelty, Top Strategic Viability*, *Top Environmental Value,* and *Top Financial Value* to capture alternative dimensions of top value and *Top Quality*. Each of these binary variables is set to one if the evaluator gave the solution the top rating (out of 5 on the five-point Likert scale) and zero otherwise.

Independent Variables. Our main independent variable, *HAI*, is a dummy variable corresponding to whether the solution is HC (baseline) or HAI generated. We also report an alternative independent variable, *HAI instance*, a categorical variable that further differentiates the HAI solutions as *HAI Multiple Instance* or *HAI Single Instance*. This alternative independent variable enables us to develop deeper insights into how alternative configurations of GPT-4 influence the generated responses.

Other Variables. Our statistical analyses rely on the random assignment of evaluators to solutions. We add several covariates corresponding to the screening criteria (i.e., work experience, interest, and skills test score), the evaluators' demographic characteristics (i.e., gender, bachelor's degree or higher, STEM major, employment status), the solution word count, and the recruitment cohort (i.e., July or September 2023).

Table 1 shows the summary statistics (mean, median, standard deviation, minimum, and maximum) as well as the correlation matrix between the main variables. Table 2 cross-tabulates descriptive statistics across HC, HAI multiple instance as well as HAI single instance. Notably, HAI multiple instance and HC are comparable, but HAI single instance exceeds HC (Multiple instance: p = 0.324; Single instance: p = 0.011). At the top decile, we do not observe differences in quality (Multiple instance: p = 0.514; Single instance: p = 0.578).

--- Insert Tables 1 and 2 here ---

Estimation Approach

We analyze our data at the evaluator–solution block level. We use nested mixed-effects models or hierarchical linear models (Gelman and Hill 2006, Kenny et al. 2006), performed using the `lmerTest` package in R (Kuznetsova et al. 2017), to account for the interdependence of data around the evaluators and solution blocks resulting from our randomized block design, which exogenously assigned evaluators one of 18 blocks of solutions to evaluate. These models appropriately account for the nesting of evaluators within solution blocks by estimating random effects (i.e., random intercepts and slopes) for both the solution blocks and the evaluators. By modeling variability at both the evaluator and solution block levels, mixed-effects models can provide more accurate estimates and standard errors than ordinary least squares in the presence of nested data (Gelman and Hill 2006).

Results

Mixed Effects Models

Tables 3–6 report the mixed model results of each rating dimension, corresponding to HAI (Table 3) and HAI Instance (Table 4), and the results for the top rating on each dimension, corresponding to HAI (Table 5) and HAI Instance (Table 6). All models include the main effect of the solution source and control for the evaluator screening criteria as well as additional evaluator attributes, cohort, and solution word count. We report all results without controls in Appendix D.

Estimated Relationships Between Solution Source and Solution Ratings. In Table 3, Models 1 and 2 indicate that compared to HC solutions, the HAI solutions receive a lower novelty (Model 1: -0.140, $p < 0.001$). In contrast, Models 2, 3, and 4 indicate that the HAI solutions attained a higher value rating in terms of strategic viability (Model 2: 0.088, $p = 0.016$), environmental value (Model 3: 0.160, $p < 0.001$) and financial value (Model 4: 0.143, $p < 0.001$). Moreover, we observe that the HAI solutions achieved higher overall quality (Model 5: 0.101, $p = 0.002$).

Table 4 splits the solution source into HAI multiple and single instance configurations. The multiple instance configuration uses different AI instances to independently explore the solution space in parallel. In the single instance configuration, humans iteratively prompt and provide feedback to guide a single AI instance's exploratory search process.

In Model 1, we observe that the HAI solutions generated with multiple instances of GPT-4 are rated as significantly less novel than the HC solutions (Model 1: -0.217, $p < 0.001$), but there is no difference between the HC and single instance HAI solutions (Model 1: -0.056, $p = 0.156$). Using the `emmeans` package in R (Lenth et al. 2018), we perform pairwise comparisons to show that the coefficients for the *HAI Multiple Instance* and *Single Instance* solutions in Model 1 are significantly different from each other ($p < 0.001$). In Model 2, we find that while there is no difference in strategic viability between the HAI multiple instance and HC solutions (Model 2: 0.047, $p = 0.239$), the HAI single instance solutions are rated as having higher strategic viability (Model 2: 0.133, $p = 0.001$). Moreover, the post-hoc test indicates that the coefficients are statistically different from each other ($p = 0.038$).

Moving to environmental and financial value in Models 3 and 4, we observe that both the HAI multiple and single instance of GPT-4 produced solutions of higher environmental (multiple: 0.136, $p < 0.001$; single: 0.186, $p < 0.001$) and financial value (multiple: 0.160, $p < 0.001$; single: 0.126, $p < 0.001$) than the HC solutions. Lastly, there is no difference in the quality of the solutions between the HAI multiple instance and HC solutions (Model 5: 0.049, $p = 0.165$), and the HAI single instance solutions are rated as higher in quality (Model 5: 0.159, $p < 0.001$). We note that the *Multiple* and *Single Instance* coefficients are statistically different from each other ($p < 0.001$).

--- Insert Tables 3 and 4 here ---

Estimated Relationships Between Solution Source and Top Solution Ratings. Turning to Table 5, which examines the top solution ratings, Model 1 indicates that HAI solutions are 7.9 percentage points less likely to achieve the top novelty rating than HC solutions (Model 1: -0.079, $p < 0.001$). In contrast, we do not find any noticeable differences between the HAI and HC solutions across any of the other dimensions in Models 2–5: *Top Strategic Viability* (Model 2: 0.004, $p = 0.710$); *Top Environmental Value* (Model 3: 0.017, $p = 0.209$); *Top Financial Value* (Model 4: 0.005, $p = 0.658$), and *Top Quality* (Model 5: -0.008, $p = 0.462$).

Next, Table 6 investigates the relationships between the HAI single and multiple instance configurations and the HC solutions for the top ratings along each dimension. In Model 1, we observe that both HAI multiple and single instance configurations are rated as less likely to achieve the top novelty rating (multiple instance: -0.091, $p < 0.001$; single instance: -0.065, $p < 0.001$). This said, there are no other observable differences between either the HAI multiple or single instance configurations and the HC solutions in terms of top ratings in Models 2–5: *Top Strategic Viability* (multiple: -0.009, $p = 0.469$; single: 0.019, $p = 0.138$); *Top Environmental Value* (multiple: 0.014, $p = 0.350$; single: 0.021, $p = 0.176$); *Top Financial Value* (multiple: 0.013, $p = 0.264$; single: -0.004, $p = 0.705$); *Top quality* (multiple: -0.017; $p = 0.173$; single: 0.001, $p = 0.924$).

In summary, our results in Tables 3–6 indicate that whereas the HC solutions are rated as more innovative, both on average and top novelty, the HAI solutions are rated as more valuable on average, in terms of strategic viability, financial and environmental value, as well as higher in overall quality. It is noteworthy that the single instance HAI solutions tend to receive higher ratings for novelty, strategic viability, and overall quality compared to the HAI multiple instance solutions. A possible explanation is that the human-guided differentiation instruction in the single instance configuration will likely force different or unique answers that may push the model towards greater novelty. An important insight of the single instance configuration is that we can achieve more novel responses (see Table 4) without compromising the perceived value of the responses. Appendix E provides further text analysis of the HC and HAI solutions.

--- Insert Tables 5 and 6 here ---

Discussion

We began this paper with the question: How will generative AI reshape creative problem-solving? Specifically, it is unclear whether an individual's use of generative AI can yield creative outputs that address open-ended and challenging organizational and societal problems (Ivcevic and Grandinetti 2024). To investigate this question, we partner with Continuum Lab, an AI firm, to launch a crowdsourcing challenge to identify sustainable, circular economy business opportunities, and compare the novelty and value of their outputs to those generated via HAI collaboration, facilitated with different prompt engineering approaches.

We subsequently invite human evaluators to assess the novelty and value of the submitted solutions without revealing their sources as HC- or HAI-generated.

Our analysis reveals that while the HC solutions demonstrate higher novelty, the HAI solutions are rated higher in value, in terms of their strategic viability to be implemented as innovations to deliver environmental and financial benefits. When considering all factors collectively, the HAI solutions are deemed superior in quality compared to the HC solutions. Our findings resonate with recent laboratory studies involving well-defined tasks, which indicate that outputs generated through HAI collaboration are approaching human-level creativity (Doshi and Hauser 2023, Franceschelli and Musolesi 2023, Girotra et al. 2023, Gómez-Rodríguez and Williams 2023, Guzik et al. 2023). Notably, our study provides one of the earliest empirical analyses of collaborative human-AI systems applied to creative problem-solving in a real-world setting, conducted after the widespread adoption of foundation generative AI models like GPT-4.

The diverging patterns of novelty and value observed between the HC and the HAI solutions may be explained by the differing search behaviors involved in the respective approaches. Specifically, while HAI leverages foundation generative AI models trained on vast datasets, the solutions generated by these models still tend to occupy more incremental search spaces that are proximal to existing solutions rather than exploring highly novel spaces, possibly due to the models' training methods on past data (Felin and Holweg 2024), or formal rationality (Lindebaum et al. 2020, Weber 1978) and the underlying statistical convergence to the "mean" answer for the prompt. As a result, HAI solutions trend towards being less innovative than the HC's but delivering higher value and overall quality. In contrast, we observe that the HC exhibits more variable search behaviors, both at the bottom and top end of the novelty distribution. Some of this heterogeneity was not observed in our analyses since the research team filtered out irrelevant and off-topic HC solutions before human evaluation. Nonetheless, consistent with the multiple parallel paths approach to innovation (Abernathy and Rosenbloom 1969, Leiponen and Helfat 2010), this variability may enable the HC to surface a greater proportion of highly novel "extreme" solution outliers (Dahan and Mendelson 2001) that current default AI systems are unlikely to produce. This said, we surmise that ultimately the novelty and value of HAI outputs will depend on how humans engage with AI to guide its solution search behaviors.

As evidence, our study employed both single instance and multiple instance configurations of GPT-4 to investigate how different forms of HAI collaboration impact solution outcomes. In the single instance configuration, humans iteratively prompted the same GPT-4 instance, instructing its search process and generating differentiated responses. In contrast, the multiple instance configuration involved independent GPT-4 instances generating solutions without iterative human input during the search process. Our results demonstrate that for current LLM capabilities, the single instance configuration with iterative human prompts can effectively increase the novelty of outputs while preserving their perceived value. This finding underscores the importance of human involvement in directing how the LLM explores the solution space. Consistent with Anthony et al. (2023), the observed differences between the independent (multiple instance) and differentiated (single instance) HAI search processes highlight that human and LLM agents offer distinct and complementary roles, both essential for effective collaborative problem-solving. More advanced search forms could involve humans providing more substantive feedback and engaging in a more conversational back-and-forth with the AI to explore and refine solutions.

Lastly, we emphasize the substantial time and cost savings from leveraging HAI approaches to creative problem-solving. In our specific study, whereas the HC solutions cost $2,555 and 2,520 hours to develop, the final HAI solutions were generated in only 5.5 hours with $27.01. This contrast indicates that generative AI presents a promising, time- and cost-effective alternative to augmenting conventional approaches to creative problem-solving.

The Future of Human-AI Creative Problem Solving

Our work builds and contributes to recent theoretical work that conceptualizes human and AI agents as interacting counterparts within a system with distinct yet complementary roles (Anthony et al. 2023). In contrast to prior studies that compare human and AI capabilities independently, akin to a horse race, our research perceives knowledge creation as a process that is shaped and emerges from the interactions between hybrid human and AI systems. In this regard, we extend the literature on HAI collaboration from routine decision-making (Allen and Choudhury 2022, Lebovitz et al. 2022) to the novel context of creative problem-solving, where the task at hand involves exploring new problems and searching for unknown solutions (Raisch and Fomina 2023). These unique characteristics of creative problem-solving render it particularly conducive to generative AI, which can create new outputs from learned patterns (Bubeck et al. 2023), as opposed to predictive AI, which anticipates future outcomes based on past learned patterns.

Our findings offer several implications for the future of creative problem-solving and innovation within organizations. First, investing in workforce training to build "AI-literate" expertise might emerge. While our study employed foundational prompting approaches, more advanced methods will likely be developed as organizations gain experience leveraging LLMs. As prompt engineering and similar interaction techniques mature, their impact on enhancing human-AI creativity is expected to grow. Such HAI approaches could allow organizations to strategically reallocate resources toward later innovation stages like solution refinement and implementation (Perry-Smith and Mannucci 2017).

Second, our findings indicate that experimenting with different forms of HAI collaboration in organizational settings—involving the inputs of the HC—could further enhance quality in problem-solving tasks. Building on evidence of distinct strengths—HC in generating novel ideas and HAI in enhancing value—a promising avenue may be to experiment with a sequential approach, in which human crowds brainstorm novel solutions, followed by HAI systems refining these ideas to improve their value, or vice versa. Alternatively, a joint search strategy could also be beneficial where HC and HAI outputs are combined iteratively. We posit that by experimenting (Levine et al. 2023) with different forms of creative problem-solving and leveraging the complementary strengths of humans and AI, we will advance organizational theory and unlock novel forms of HAI collaboration to identify optimal divisions of labor that go beyond those explored in our current study.

Third, our findings suggest that organizations integrating AI into their workflows may need to adopt robust processes to ensure the responsible and unbiased use of LLMs in creative endeavors. Despite their training on wide-ranging datasets, HAI responses were consistently rated less likely to achieve top novelty than the HC. This tendency may result from various factors, including the fine-tuning and alignment techniques used in LLM training or the inherent limitation that LLMs, despite their generative capabilities, are trained on historical data (Felin and Holweg 2024). The substantial time and cost savings from HAI collaboration for creative problem-solving only make addressing these issues more pressing. Along with other studies, we caution that excessive dependence on LLMs may undermine human creativity and output diversity (Dell'Acqua et al. 2023, Doshi and Hauser 2023, Stevenson et al. 2022). Particularly concerning is the potential for evaluators to prioritize ideas conforming to established success patterns, leading to a bias favoring incremental innovations over more radical breakthroughs (Dewar and Dutton 1986).

As AI capabilities advance, human roles in the creative process will inevitably evolve, prompting deeper reflections on the nature of creativity itself and the unique value humans contribute. These potential paradigm shifts underscore the importance of thoughtfully delineating responsibilities between human and AI collaborators while recognizing them as complementary components within an interactive system.

Methodological Considerations, Limitations, and Future Directions

Despite employing a highly practical approach to study HAI creative problem-solving in a rapidly advancing field, our study has limitations that point to avenues for future research. First, our study's reliance on evaluators based solely in the U.S. limits the generalizability of our findings. This geographical constraint may skew the evaluation of solutions, as cultural and contextual understandings of novelty and value vary globally (Jang 2017). Consequently, the insights derived from this study may not fully encapsulate the diverse perspectives that evaluators from different cultural backgrounds could offer, potentially affecting the applicability of our conclusions across international contexts.

Second, we recruited crowd rather than domain-specific experts to evaluate solutions. This approach may impact the perceived novelty and value of the generated solutions. Experts, with their deep domain knowledge, might assess the solutions differently, focusing on aspects laypersons might overlook (Boudreau et al. 2016, Mollick and Nanda 2016). Although research suggests that it can sometimes be costly to recruit experts to evaluate a multitude of ideas (Bell et al. 2024) and that the crowd can be a good proxy of expert opinions in creative contexts (Mollick and Nanda 2016), the crowd's evaluation may nonetheless not fully capture the nuanced domain understanding that experts may bring, potentially leading to an underestimation or overestimation of the solutions' novelty and applicability. Future studies should expand the evaluators' pool to a more globally diverse representation of crowds and experts.

Third, the quality of the HAI outputs in our study may have been influenced by the training data, the model setting, and the specific prompt engineering strategies employed by one of the study team's authors. The configuration of LLMs, particularly the temperature parameter, may play a critical role in determining the creativity and relevance of the outputs. Although rigorous research is needed (Renze and Guven 2024), higher temperature settings may lead to more creative, statistically rare responses, while lower settings tend to produce more conservative and relevant outputs (Chen et al. 2021).

Additionally, since LLM outputs are guided by human interaction with the AI system, we must interpret our results carefully. The coauthor who performed the HAI collaboration in this study possesses deep technical expertise in prompt engineering and invested significant time and effort in refining the prompts to generate the LLM solutions. Furthermore, we limited the degree of heterogeneity in the prompts for generating solutions to maintain a quality threshold in the responses and enable effective solution generation at scale. It is also important to note that our study aims to achieve multiple objectives in both novelty and value of the LLM outputs. Prioritizing one aspect over the other may lead to compromises, which we do not investigate specifically.

These considerations in our approach to establishing realistic yet scalable forms of HAI collaboration for creative problem-solving highlight the promise of future research in investigating different forms of HAI solution search. Expanding the view of HAI collaboration to explore its possibilities for problem formulation and evaluation could also yield valuable insights.

Lastly, in our research, we focused on the capabilities of a single LLM. One possibility is incorporating more sophisticated applications, including domain-specific knowledge (Yager 2023) and adjusting for emotional tone (Yin et al. 2024) to improve LLMs' capabilities to offer more nuanced and contextually appropriate solutions. In addition, an intriguing avenue for further elevating LLM creativity is to build on the collective insight of multi-modal (Yin et al. 2023) and multi-agent systems that collaborate and compete with one another (Wang et al. 2023, Xi et al. 2023). Moreover, Retrieval-Augmented

Generation systems could enable LLMs to access and process both external and proprietary (i.e., firm-specific) knowledge bases, enhancing factual accuracy and enriching their responses (Lewis et al. 2020). Additionally, the emergence of LLMs with web-browsing capabilities may further expand the range of information available to these models, potentially improving their performance in creative problem-solving tasks. Last, beyond the family of GPT-4-level models, an array of open-source LLMs are swiftly advancing and beginning to rival the capabilities of the closed-source ones. Importantly, because these alternative LLMs might be trained on different datasets and use alternative fine-tuning approaches or guardrails impacting their outputs, their collaborative output could offer more creative recombinations than responses from a single GPT-4 model.

Despite these limitations, our findings have important implications for creative problem-solving, as they demonstrate the feasibility of HAI collaboration for solution generation. By providing a proof of concept, our study lays the groundwork for leveraging HAI collaboration to search through the solution space effectively and efficiently. This approach holds promise for enhancing the creative problem-solving process and unlocking new avenues for organizational innovative activities. Looking forward, the rapid advancement in generative AI capabilities holds tremendous promise for transforming human-centered innovation processes through synergistic human-AI integration, known as AI-in-the-loop.

References

Abernathy WJ, Rosenbloom RS (1969) Parallel strategies in development projects. *Manag. Sci.* 15(10):B-486-B-505.

Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S (2023) Gpt-4 technical report. *ArXiv Prepr. ArXiv230308774*.

Agrawal A, Gans J, Goldfarb A (2018) *Prediction Machines: The Simple Economics of Artificial Intelligence* (Harvard Business Review Press).

Allen R, Choudhury P (2022) Algorithm-augmented work and domain experience: The countervailing forces of ability and aversion. *Organ. Sci.* 33(1):149–169.

Amabile TM (1983) The social psychology of creativity: A componential conceptualization. *J. Pers. Soc. Psychol.* 45(2):357.

Anderson BR, Shah JH, Kreminski M (2024) Homogenization Effects of Large Language Models on Human Creative Ideation. *ArXiv Prepr. ArXiv2402.01536.*

Anthony C, Bechky BA, Fayard AL (2023) “Collaborating” with AI: Taking a system view to explore the future of work. *Organ. Sci.* 34(5):1672–1694.

Ash E, Hansen S (2023) Text algorithms in economics. *Annu. Rev. Econ.* 15:659–688.

Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, Faix DJ, Goodman AM, Longhurst CA, Hogarth M (2023) Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* 183(6):589–596.

Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. *ArXiv Prepr. ArXiv14090473*.

Bai Y, Kadavath S, Kundu S, Askell A, Kernion J, Jones A, Chen A, Goldie A, Mirhoseini A, McKinnon C (2022) Constitutional ai: Harmlessness from ai feedback. *ArXiv Prepr. ArXiv221208073*.

Barr PS, Stimpert JL, Huff AS (1992) Cognitive change, strategic action, and organizational renewal. *Strateg. Manag. J.* 13(S1):15–36.

Baumol WJ (1993) Formal entrepreneurship theory in economics: Existence and bounds. *J. Bus. Ventur.* 8(3):197–210.

Battle R, Gollapudi T (2024) The Unreasonable Effectiveness of Eccentric Automatic Prompts. *ArXiv Prepr. ArXiv240210949*

Becker GS (1994) *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*, 3rd ed. (The University of Chicago Press).

Bell JJ, Pescher C, Tellis GJ, Füller J (2024) Can AI help in ideation? A theory-based model for idea screening in crowdsourcing contests. *Mark. Sci.* 43(1):54–72.

Bellemare-Pepin A, Lespinasse F, Thölke P, Harel Y, Mathewson K, Olson J., Bengio Y, Jerbi K (2024) Divergent Creativity in Humans and Large Language Models. *ArXiv Prepr. ArXiv2405.13012*.

Benner MJ, Tushman ML (2003) Exploitation, exploration, and process management: The productivity dilemma revisited. *Acad. Manage. Rev.* 28(2):238–256.

Boudreau KJ, Guinan EC, Lakhani KR, Riedl C (2016) Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Manag. Sci.* 62(10):2765–2783.

Boudreau KJ, Lacetera N, Lakhani KR (2011) Incentives and problem uncertainty in innovation contests: An empirical analysis. *Manag. Sci.* 57(5):843–863.

Brand J, Israeli A, Ngwe D (2023) Using gpt for market research. *Available SSRN 4395751*.

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, et al. (2020) Language Models are Few-Shot Learners. *CoRR* abs/2005.14165.

Brynjolfsson E, Li D, Raymond LR (2023) *Generative AI at work* (National Bureau of Economic Research).

Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, Lee P, Lee YT, Li Y, Lundberg S (2023) Sparks of artificial general intelligence: Early experiments with gpt-4. *ArXiv Prepr. ArXiv230312712*.

Che YK, Gale I (2003) Optimal design of research contests. *Am. Econ. Rev.* 93(3):646–671.

Chen M, Tworek J, Jun H, Yuan Q, Pinto HP de O, Kaplan J, Edwards H, et al. (2021) Evaluating Large Language Models Trained on Code.

Choudhury P, Allen RT, Endres MG (2021) Machine learning for pattern discovery in management research. *Strateg. Manag. J.* 42(1):30–57.

Choudhary V, Marchetti A, Shrestha YR, Puranam P (2023) Human-AI Ensembles: When Can They Work? *J. Manag.*:01492063231194968.

Cyert RM, March JG (1963) A behavioral theory of the firm. *Englewood Cliffs NJ* 2(4):169–187.

Dahan E, Mendelson H (2001) An extreme-value model of concept testing. *Manag. Sci.* 47(1):102–116.

DAIR.AI (2024) Prompt Chaining. Prompting Guide. Retrieved March 21, 2024, from https://www.promptingguide.ai/techniques/prompt_chaining.

Dell'Acqua F, McFowland E, Mollick ER, Lifshitz-Assaf H, Kellogg K, Rajendran S, Krayer L, Candelon F, Lakhani KR (2023) Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harv. Bus. Sch. Technol. Oper. Mgt Unit Work. Pap.* (24–013).

Dewar RD, Dutton JE (1986) The adoption of radical and incremental innovations: An empirical analysis. *Manag. Sci.* 32(11):1422–1433.

Doshi AR, Hauser O (2023) Generative artificial intelligence enhances creativity. *Available SSRN*.

Felin T, Holweg M (2024) Theory Is All You Need: AI, Human Cognition, and Decision Making. *Hum. Cogn. Decis. Mak. Febr. 23 2024*.

Fleming L, Mingo S, Chen D (2007) Collaborative brokerage, generative creativity, and creative success. *Adm. Sci. Q.* 52(3):443–475.

Fleming L, Sorenson O (2001) Technology as a complex adaptive system: evidence from patent data. *Res. Policy* 30(7):1019–1039.

Franceschelli G, Musolesi M (2023) On the creativity of large language models. *ArXiv Prepr. ArXiv230400008*.

Gavetti G, Levinthal D (2000) Looking forward and looking backward: Cognitive and experiential search. *Adm. Sci. Q.* 45(1):113–137.

Gelman A, Hill J (2006) *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Cambridge University Press).

Girotra K, Meincke L, Terwiesch C, Ulrich KT (2023) Ideas are dimes a dozen: Large language models for idea generation in innovation. *Available SSRN 4526071*.

Girotra K, Terwiesch C, Ulrich KT (2010) Idea generation and the quality of the best idea. *Manag. Sci.* 56(4):591–605.

Glaeser EL, Laibson D, Sacerdote B (2002) An Economic Approach to Social Capital*. *Econ. J.* 112(483):F437–F458.

Gómez-Rodríguez C, Williams P (2023) A confederacy of models: A comprehensive evaluation of LLMs on creative writing. *ArXiv Prepr. ArXiv231008433*.

Guzik EE, Byrge C, Gilde C (2023) The originality of machines: AI takes the Torrance Test. *J. Creat.* 33(3):100065.

Hagendorff T, Fabi S, Kosinski M (2023) Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nat. Comput. Sci.* 3(10):833–838.

Hargadon AB, Bechky BA (2006) When collections of creatives become creative collectives: A field study of problem solving at work. *Organ. Sci.* 17(4):484–500.

He VF, Shrestha YR, Puranam P, Miron-Spektor E (2023) Searching Together: A Theory of Human-AI Co-Creativity.

Henderson R, Cockburn I (1994) Measuring competence? Exploring firm effects in pharmaceutical research. *Strateg. Manag. J.* 15(S1):63–84.

Iansiti M, Lakhani KR (2020) *Competing in the age of AI: Strategy and leadership when algorithms and networks run the world* (Harvard Business Press).

Imbens GW, Rubin DB (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences* (Cambridge University Press).

Ivcevic Z, Grandinetti M (2024) Artificial intelligence as a tool for creativity. *J. Creat.*:100079.

Jang S (2017) Cultural brokerage and quality in multicultural teams. *Organ. Sci.* 28(6):993–1009.

Jeppesen LB, Lakhani KR (2010) Marginality and problem-solving effectiveness in broadcast search. *Organ. Sci.* 21(5):1016–1033.

Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang YJ, Madotto A, Fung P (2023) Survey of hallucination in natural language generation. *ACM Comput. Surv.* 55(12):1–38.

Jia N, Luo X, Fang Z, Liao C (2023) When and how artificial intelligence augments employee creativity. *Acad. Manage. J.* (ja).

Kaplan S, Vakili K (2015) The double-edged sword of recombination in breakthrough innovation. *Strateg. Manag. J.* 36(10):1435–1457.

Katila R, Ahuja G (2002) Something old, something new: A longitudinal study of search behavior and new product introduction. *Acad. Manage. J.* 45(6):1183–1194.

Kenny D, Kashy D, Cook W, Simpson J (2006) *Dyadic Data Analysis* (The Guildford Press, New York).

Kim H, Glaeser EL, Hillis A, Kominers SD, Luca M (2024) Decision authority and the returns to algorithms. *Strateg. Manag. J.* 45(4): 619–648.

Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S (2018) Human decisions and machine predictions. *Q. J. Econ.* 133(1):237–293.

Kneeland MK, Schilling MA, Aharonson BS (2020) Exploring uncharted territory: Knowledge search processes in the origination of outlier innovation. *Organ. Sci.* 31(3):535–557.

Koivisto M, Grassini S (2023) Best humans still outperform artificial intelligence in a creative divergent thinking task. *Sci. Rep.* 13(1):13601.

Kong A, Zhao S, Chen H, Li Q, Qin Y, Sun R, Zhou X (2023) Better Zero-Shot Reasoning with Role-Play Prompting. *ArXiv Prepr. ArXiv230807702*

Kumar A, Davenport T (2023) How to make generative AI greener. *Harv. Bus. Rev.* 20.

Kuznetsova A, Brockhoff PB, Christensen RHB (2017) lmerTest Package: Tests in Linear Mixed Effects Models. *J. Stat. Softw.* 82(13):1–26.

Laursen K, Salter A (2006) Open for innovation: the role of openness in explaining innovation performance among UK manufacturing firms. *Strateg. Manag. J.* 27(2):131–150.

Lebovitz S, Lifshitz-Assaf H, Levina N (2022) To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organ. Sci.* 33(1):126–148.

Leiponen A, Helfat CE (2010) Innovation objectives, knowledge sources, and the benefits of breadth. *Strateg. Manag. J.* 31(2):224–236.

Lenth R, Love J, Herve M (2018) *emmeans: Estimated Marginal Means, aka Least-Squares Means* (CRAN, https://cran.r-project.org/package=emmeans).

Levine SS, Schilke O, Kacperczyk O, Zucker LG (2023) Primer for Experimental Methods in Organization Theory. *Organ. Sci.* 34(6):1997–2025.

Levinthal DA (1997) Adaptation on rugged landscapes. *Manag. Sci.* 43(7):934–950.

Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W tau, Rocktäschel T (2020) Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv. Neural Inf. Process. Syst.* 33:9459–9474.

Li D, Raymond LR, Bergman P (2020) *Hiring as exploration* (National Bureau of Economic Research).

Lifshitz-Assaf H (2018) Dismantling knowledge boundaries at NASA: The critical role of professional identity in open innovation. *Adm. Sci. Q.* 63(4):746–782.

Lindebaum D, Vesa M, Den Hond F (2020) Insights from "the machine stops" to better understand rational assumptions in algorithmic decision making and its implications for organizations. *Acad. Manage. Rev.* 45(1):247–263.

Lingo EL, O'Mahony S (2010) Nexus work: Brokerage on creative projects. *Adm. Sci. Q.* 55(1):47–81.

Lou B, Wu L (2021) AI on Drugs: Can Artificial Intelligence Accelerate Drug Development? Evidence from a Large-Scale Examination of Bio-Pharma Firms. *Manag. Inf. Syst. Q.* 45(3):1451–1482.

March JG (1991) Exploration and exploitation in organizational learning. *Organ. Sci.* 2(1):71–87.

Meincke L, Mollick ER, Terwiesch C (2024) Prompting Diverse Ideas: Increasing AI Idea Variance. *ArXiv Prepr. ArXiv240201727*.

Miric M, Jia N, Huang KG (2023) Using supervised machine learning for large-scale classification in management research: The case for identifying artificial intelligence patents. *Strateg. Manag. J.* 44(2):491–519.

Mollick E, Nanda R (2016) Wisdom or madness? Comparing crowds with expert evaluation in funding the arts. *Manag. Sci.* 62(6):1533–1553.

Nelson RR (1961) Uncertainty, learning, and the economics of parallel research and development efforts. *Rev. Econ. Stat.*:351–364.

Nelson RR, Winter SG (1982) The Schumpeterian tradeoff revisited. *Am. Econ. Rev.* 72(1):114–132.

Nickerson JA, Zenger TR (2004) A knowledge-based theory of the firm—The problem-solving perspective. *Organ. Sci.* 15(6):617–632.

Noy S, Zhang W (2023) Experimental evidence on the productivity effects of generative artificial intelligence. *Available SSRN 4375283*.

Ocasio W (1997) Towards an attention-based view of the firm. *Strateg. Manag. J.* 18(S1):187–206.

OpenAI (2024) Strategy: Write Clear Instructions.

Otis N, Clarke RP, Delecourt S, Holtz D, Koning R (2023) The Uneven Impact of Generative AI on Entrepreneurial Performance. *Available SSRN 4671369*.

Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A (2022) Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* 35:27730–27744.

Paik JH, Scholl M, Sergeev R, Randazzo S, Lakhani KR (2020) Innovation contests for high-tech procurement. *Res.-Technol. Manag.* 63(2):36–45.

Perry-Smith JE (2006) Social yet creative: The role of social relationships in facilitating individual creativity. *Acad. Manage. J.* 49(1):85–101.

Perry-Smith JE, Mannucci PV (2017) From creativity to innovation: The social network drivers of the four phases of the idea journey. *Acad. Manage. Rev.* 42(1):53–79.

Piezunka H, Dahlander L (2015) Distant search, narrow attention: How crowding alters organizations' filtering of suggestions in crowdsourcing. *Acad. Manage. J.* 58(3):856–880.

Piezunka H, Dahlander L (2019) Idea rejected, tie formed: Organizations' feedback on crowdsourced ideas. *Acad. Manage. J.* 62(2):503–530.

Poetz MK, Schreier M (2012) The value of crowdsourcing: can users really compete with professionals in generating new product ideas? *J. Prod. Innov. Manag.* 29(2):245–256.

Raisch S, Fomina K (2023) Combining human and artificial intelligence: Hybrid problem-solving in organizations. *Acad. Manage. Rev.*

Riedl C, Grad T, Lettl C (2024) Competition and Collaboration in Crowdsourcing Communities: What Happens When Peers Evaluate Each Other? *Organ. Sci.*

Renze M, Guven E (2024) The Effect of Sampling Temperature on Problem Solving in Large Language Models. *ArXiv Prepr. ArXiv240205201*.

Rhee L, Leonardi PM (2018) Which pathway to good ideas? A n attention-based view of innovation in social networks. *Strateg. Manag. J.* 39(4):1188–1215.

Rindova VP, Petkova AP (2007) When is a new thing a good thing? Technological change, product form design, and perceptions of value for product innovations. *Organ. Sci.* 18(2):217–232.

Saravia E (2022) *Prompt Engineering Guide*. Accessed June 22, 2024. https://www.promptingguide.ai/.

Shanahan M, McDonell K, Reynolds L (2023) Role play with large language models. *Nature* 623(7987):493–498.

Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M (2016) Mastering the game of Go with deep neural networks and tree search. *Nature* 529(7587):484–489.

Simon HA (1973) The structure of ill structured problems. *Artif. Intell.* 4(3–4):181–201.

Stevenson C, Smal I, Baas M, Grasman R, van der Maas H (2022) Putting GPT-3's Creativity to the (Alternative Uses) Test. *ArXiv Prepr. ArXiv220608932*.

Taylor CR (1995) Digging for golden carrots: An analysis of research tournaments. *Am. Econ. Rev.*:872–890.

Teodoridis F, Bikard M, Vakili K (2019) Creativity at the knowledge frontier: The impact of specialization in fast-and slow-paced domains. *Adm. Sci. Q.* 64(4):894–927.

Terwiesch C, Ulrich KT (2009) *Innovation tournaments: Creating and selecting exceptional opportunities* (Harvard Business Press).

Terwiesch C, Xu Y (2008) Innovation contests, open innovation, and multiagent problem solving. *Manag. Sci.* 54(9):1529–1543.

Tong S, Jia N, Luo X, Fang Z (2021) The Janus face of artificial intelligence feedback: Deployment versus disclosure effects on employee performance. *Strateg. Manag. J.* 42(9):1600–1631.

Tripsas M (2009) Technology, identity, and inertia through the lens of "The Digital Photography Company." *Organ. Sci.* 20(2):441–460.

Tushman ML, Anderson P (1986) Technological discontinuities and organizational environments. *Adm. Sci. Q.*:439–465.

Van de Ven AH (1986) Central problems in the management of innovation. *Manag. Sci.* 32(5):590–607.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. *Adv. Neural Inf. Process. Syst.* 30.

Wang L, Ma C, Feng X, Zhang Z, Yang H, Zhang J, Chen Z, Tang J, Chen X, Lin Y (2023) A survey on large language model based autonomous agents. *ArXiv Prepr. ArXiv230811432*.

Weber M (1978) *Economy and society: An outline of interpretive sociology* (University of California press).

Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, Chi E, Le Q, Zhou D (2023) Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.

Wuchty S, Jones BF, Uzzi B (2007) The increasing dominance of teams in production of knowledge. *Science* 316(5827):1036–1039.

Xi Z, Chen W, Guo X, He W, Ding Y, Hong B, Zhang M, Wang J, Jin S, Zhou E (2023) The rise and potential of large language model based agents: A survey. *ArXiv Prepr. ArXiv230907864*.

Yager KG (2023) Domain-specific chatbots for science using embeddings. *Digit. Discov.* 2(6):1850–1861.

Yin S, Fu C, Zhao S, Li K, Sun X, Xu T, Chen E (2023) A Survey on Multimodal Large Language Models. *ArXiv Prepr. ArXiv230613549*.

Yin Z, Wang H, Horio K, Kawahara D, Sekine S (2024) Should We Respect LLMs? A Cross-Lingual Study on the Influence of Prompt Politeness on LLM Performance. *ArXiv Prepr. ArXiv240214531*.

Zamfirescu-Pereira JD, Wong RY, Hartmann B, Yang Q (2023) Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. *Proc. 2023 CHI Conf. Hum. Factors Comput. Syst.* CHI '23. (Association for Computing Machinery, New York, NY, USA).

Zhou Y, Muresanu AI, Han Z, Paster K, Pitis S, Chan H, Ba J (2022) Large language models are human-level prompt engineers. *ArXiv Prepr. ArXiv221101910*.

Table 1. Summary Statistics and Correlation Between Main Variables (N = 3,900)

		Mean	Med	SD	Min	Max	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	Novelty	3.412	3	1.047	1	5	1.000													
2	Strategic Viability	3.257	3	1.071	1	5	0.261	1.000												
3	Environmental Value	3.754	4	0.954	1	5	0.419	0.454	1.000											
4	Financial Value	3.211	3	1.013	1	5	0.405	0.524	0.496	1.000										
5	Quality	3.378	3	0.973	1	5	0.554	0.707	0.646	0.680	1.00									
6	HAI M/S*	1.154	1	0.769	0	2	0.000	0.045	0.082	0.051	0.069	1.000								
7	Experience	4.680	3.5	5.172	0	45.5	0.001	-0.000	-0.002	0.038	0.023	0.000	1.000							
8	Interest	3.913	4	0.16	3	5	0.061	0.091	0.098	0.116	0.116	0.000	0.134	1.000						
9	Score	2.447	2.5	1.158	0	5	-0.069	-0.075	-0.080	-0.124	-0.092	0.000	-0.156	-0.058	1.000					
10	Female	0.370	0	0.483	0	1	0.039	0.042	0.071	0.048	0.070	0.000	-0.161	0.039	0.235	1.000				
11	Bachelor’s	0.617	1	0.486	0	1	-0.018	0.001	-0.006	-0.019	-0.021	0.000	0.034	0.076	0.038	0.008	1.000			
12	STEM	0.467	0	0.499	0	1	0.001	0.002	0.001	0.007	-0.005	0.000	0.142	0.009	-0.084	-0.080	0.064	1.000		
13	Employed	0.860	1	0.347	0	1	0.026	0.020	-0.027	0.042	0.015	0.000	0.127	0.063	-0.102	-0.208	0.235	0.012	1.000	
14	Cohort	0.517	1	0.499	0	1	0.101	0.050	0.111	0.109	0.074	0.000	0.160	0.085	-0.151	-0.074	-0.008	0.022	0.129	1.000
15	Word Count	237.769	238	114.243	35	1049	0.030	-0.012	0.053	0.055	0.031	0.209	-0.008	-0.023	0.004	0.011	0.003	0.008	-0.033	-0.009

Notes: All values of $|\rho| > 0.03$ are significant at $p < 0.05$. *HAI M/S instances were equally split and each corresponded to 38.5% of the observations.

Table 2. Cross-Tabulation of Summary Statistics Across Solution Sources

	Human Crowd (HC)	**Human-AI (HAI) Multiple Instance**	**Human-AI (HAI) Single Instance**
N Ideas	54	90	90
Average Word Count	204	231	265
		Mean Rating Across All Solutions	
Novelty	3.508	3.230	3.469
	(1.127)	(1.040)	(0.993)
p-value (vs HC)	—	0.013	0.657
p-value (vs multiple instance)	—	—	0.002
Strategic Viability	3.194	3.236	3.315
	(1.125)	(1.037)	(1.070)
p-value (vs HC)	—	0.445	0.069
p-value (vs multiple instance)	—	—	0.165
Environmental Value	3.616	3.763	3.827
	(1.037)	(0.934)	(0.913)
p-value (vs HC)	—	0.009	0.000
p-value (vs multiple instance)	—	—	0.120
Financial Value	3.086	3.237	3.239
	(1.051)	(1.002)	(0.996)
p-value (vs HC)	—	0.004	0.009
p-value (vs multiple instance)	—	—	0.636
Quality	3.292	3.345	3.461
	(1.058)	(0.952)	(0.936)
p-value (vs HC)	—	0.324	0.007
p-value (vs multiple instance)	—	—	0.011
		Mean Rating for Solutions in Top Decile	
Novelty	4.364	3.901	4.003
p-value (vs HC)	—	0.000	0.000
p-value (vs multiple instance)	—	—	0.184
Strategic Viability	3.789	3.797	3.886
p-value (vs HC)	—	0.881	0.132
p-value (vs multiple instance)	—	—	0.091
Environmental Value	4.197	4.308	4.267
p-value (vs HC)	—	0.043	0.181
p-value (vs multiple instance)	—	—	0.466
Financial Value	3.701	3.801	3.702
p-value (vs HC)	—	0.304	0.985
p-value (vs multiple instance)	—	—	0.087
Quality	3.950	3.902	3.975
p-value (vs HC)	—	0.514	0.578
p-value (vs multiple instance)	—	—	0.340

Table 3. Human Crowd vs Human-AI Nested Mixed Effects Models of Evaluator Ratings

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**
	Novelty	**Strategic Viability**	**Env't Value**	**Financial Value**	**Quality**
HAI Solution	-0.140***	0.088*	0.160***	0.143***	0.101**
	(0.035)	(0.037)	(0.029)	(0.033)	(0.032)
Intercept	3.222***	2.979***	3.375***	2.658***	2.956***
	(0.214)	(0.213)	(0.224)	(0.217)	(0.204)
N	3900	3900	3900	3900	3900
# blocks	18	18	18	18	18
# evaluators	300	300	300	300	300
Screening criteria	Y	Y	Y	Y	Y
Other controls	Y	Y	Y	Y	Y
Log-Likelihood	-5439.9 df = 15	-5568.3 df = 15	-4798.1 df = 15	-5169.7 df = 15	-5085.0 df = 15

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Notes. This table presents mixed-model (hierarchical linear modeling) results from evaluator ratings of solution novelty, strategic viability, environmental value, financial value, and quality with 300 evaluators nested in eighteen solution blocks. All models include the screening criteria of work Experience, level of interest, and knowledge test score, along with covariates such as gender, highest level of education, major, employment Status, cohort session, and solution word count. Standard errors are in parentheses.

Table 4. Human Crowd vs Human-AI Multiple and Single Instance Nested Mixed Effects Models of Evaluator Ratings

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**
	Novelty	**Strategic Viability**	**Env't Value**	**Financial Value**	**Quality**
HAI Multiple Instance	-0.217***	0.047	0.136***	0.160***	0.049
	(0.039)	(0.040)	(0.032)	(0.036)	(0.035)
HAI Single Instance	-0.056	0.133**	0.186***	0.126***	0.159***
	(0.039)	(0.041)	(0.033)	(0.037)	(0.036)
Intercept	3.242***	2.990***	3.382***	2.654***	2.970***
	(0.214)	(0.213)	(0.224)	(0.217)	(0.204)
N	3900	3900	3900	3900	3900
# blocks	18	18	18	18	18
# evaluators	300	300	300	300	300
Screening criteria	Y	Y	Y	Y	Y
Other controls	Y	Y	Y	Y	Y
Log-Likelihood	-5430.9 df = 16	-5567.7 df = 16	-4799.1 df = 16	-5171.7 df = 16	-5081.1 df = 16

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Notes. This table presents mixed-model (hierarchical linear modeling) results from evaluator ratings of solution novelty, strategic viability, environmental value, financial value, and quality with 300 evaluators nested in eighteen solution blocks. All models include the screening criteria of work Experience, level of interest, and knowledge test score, along with covariates such as gender, highest level of education, major, employment Status, cohort session, and solution word count. Standard errors are in parentheses.

Table 5. Human Crowd vs Human-AI Nested Mixed Effects Models of Top Evaluator Ratings

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**
	Top Novelty	**Top Strategic Viability**	**Top Env't Value**	**Top Financial Value**	**Top Quality**
HAI Solution	-0.079*** (0.012)	0.004 (0.012)	0.017 (0.014)	0.005 (0.010)	-0.008 (0.011)
Intercept	0.176** (0.068)	0.019 (0.061)	0.085 (0.098)	0.005 (0.056)	0.042 (0.067)
N	3900	3900	3900	3900	3900
# blocks	18	18	18	18	18
# evaluators	300	300	300	300	300
Screening criteria	Y	Y	Y	Y	Y
Other controls	Y	Y	Y	Y	Y
Log-Likelihood	-1357.0 df = 15	-1059.1 df = 15	-1814.5 df = 15	-614.9 df = 15	-966.0 df = 15

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Notes. This table presents mixed-model (hierarchical linear modeling) results from evaluator ratings of top solution novelty, strategic viability, environmental value, financial value, and quality with 300 evaluators nested in eighteen solution blocks. All models include the screening criteria of work Experience, level of interest, and knowledge test score, along with covariates such as gender, highest level of education, major, employment Status, cohort session, and solution word count. Standard errors are in parentheses.

Table 6. Human Crowd vs Human-AI Multiple and Single Instance Nested Mixed Effects Models of Top Evaluator Ratings

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**
	Top Novelty	**Top Strategic Viability**	**Top Env't Value**	**Top Financial Value**	**Top Quality**
HAI Multiple Instance	-0.091*** (0.014)	-0.009 (0.031)	0.014 (0.015)	0.013 (0.011)	-0.017 (0.012)
HAI Single Instance	-0.065*** (0.014)	0.019 (0.013)	0.021 (0.015)	-0.004 (0.011)	0.001 (0.012)
Intercept	0.179** (0.068)	0.022 (0.061)	0.086 (0.098)	0.003 (0.056)	0.044 (0.067)
N	3900	3900	3900	3900	3900
# blocks	18	18	18	18	18
# evaluators	300	300	300	300	300
Screening criteria	Y	Y	Y	Y	Y
Other controls	Y	Y	Y	Y	Y
Log-Likelihood	-1358.1 df = 16	-1059.4 df = 16	-1817.8 df = 16	-617.1 df = 16	-968.2 df = 16

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Notes. This table presents mixed-model (hierarchical linear modeling) results from evaluator ratings of top solution novelty, strategic viability, environmental value, financial value, and quality with 300 evaluators nested in eighteen solution blocks. All models include the screening criteria of work Experience, level of interest, and knowledge test score, along with covariates such as gender, highest level of education, major, employment Status, cohort session, and solution word count. Standard errors are in parentheses.

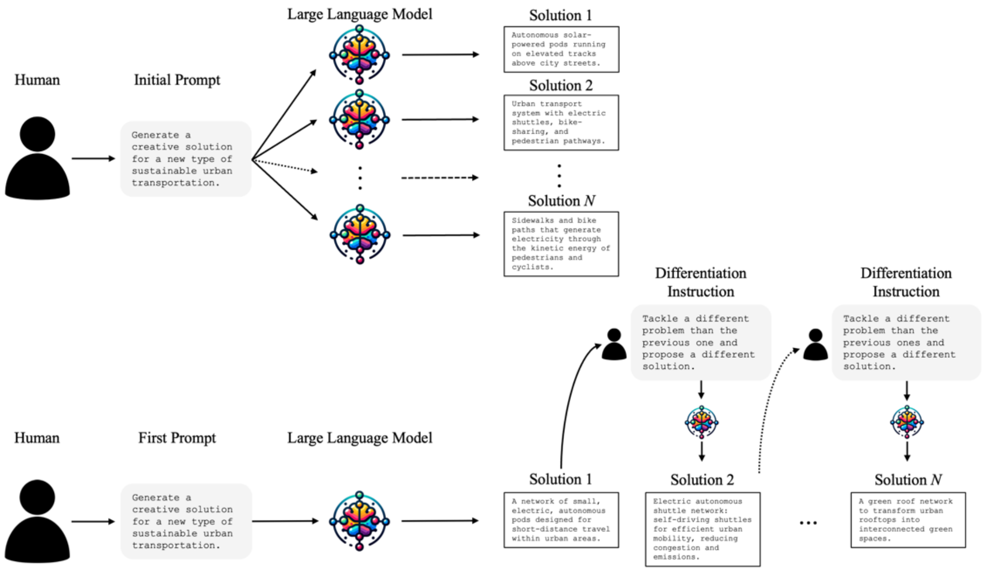
Figure 1. Conceptualization of Two Alternative Forms of Human-AI Solution Search

Independent Search

The human user provides the initial prompt, and the LLM generates multiple independent solutions.

Differentiated Search

The human user provides the first prompt and inserts a differentiation prompt to diversify the subsequent solution after each individual solution the LLM generates.



Initial Prompt: Generate a creative solution for a new type of sustainable urban transportation.

Solution 1: Autonomous solar-powered pods running on elevated tracks above city streets.

Solution 2: Urban transport system with electric shuttles, bike-sharing, and pedestrian pathways.

Solution *N*: Sidewalks and bike paths that generate electricity through the kinetic energy of pedestrians and cyclists.

First Prompt: Generate a creative solution for a new type of sustainable urban transportation.

Differentiation Instruction: Tackle a different problem than the previous one and propose a different solution.

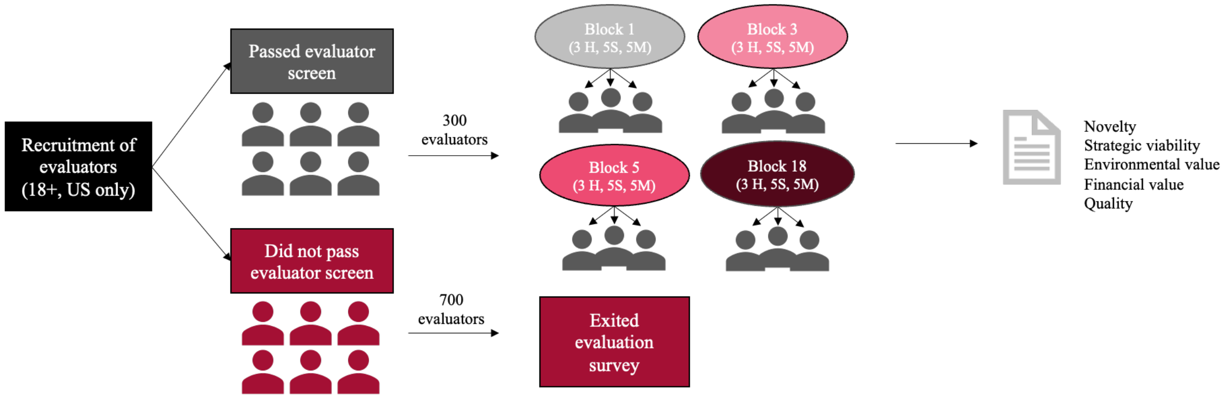
Differentiation Instruction: Tackle a different problem than the previous ones and propose a different solution.

Solution 1: A network of small, electric, autonomous pods designed for short-distance travel within urban areas.

Solution 2: Electric autonomous shuttle network: self-driving shuttles for efficient urban mobility, reducing congestion and emissions.

Solution *N*: A green roof network to transform urban rooftops into interconnected green spaces.

Figure 2. Flow of Evaluator Recruitment and Procedures



Recruited evaluators (18+, U.S. only) on Prolific

Each potential evaluator completed an initial screen for their interest, work experience, and knowledge in the circular economy. Evaluators with moderate or high interest and either 2+ years of work experience or 60% or more on the knowledge skills test passed the screen. Out of 1000 individuals screened, 300 passed.

Each of the 300 evaluators passing the circular economy screen was then randomly assigned one out of 18 blocks of circular economy solutions. Each block contained 13 solutions: 3 human (H) and 10 AI solutions. Of the AI solutions, 5 were generated using a single (S) instance of GPT-4, and 5 were generated using multiple (M) instances of GPT-4.

Each evaluator rated each assigned solution on five criteria: novelty, strategic viability, environmental value, financial value, and overall quality.