Working Paper 24-005

Generative AI and Creative Problem Solving

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

Harvard Business School

Generative AI and Creative Problem Solving

Léonard Boussioux
University of Washington
Massachusetts Institute of Technology

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.

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 & Massachusetts Institute of Technology, Operations Research Center
[2]Harvard Business School & Digital Data and Design (D^3) Institute at Harvard
[3]Continuum Labs

*Corresponding author: jnlane@hbs.edu

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

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 and assessed the novelty, value, and creativity of solutions created by both the human crowd and the collaborative efforts of one human and AI. The challenge attracted 125 global solvers from various industries and the human-prompted AI solutions were generated using strategic prompt engineering. We recruited 300 evaluators to judge a randomized selection of 13 out of 234 solutions, totaling 3,900 evaluator–solution pairs. Our findings demonstrate that the solutions generated through human-AI collaboration matched the creativity of those from the human solvers. Whereas the human-AI solutions provided more value, the human-only solutions were more innovative—both on average and for highly novel outcomes. Our study explores the potential for incorporating "AI-in-the-loop" into creative problem-solving, offering a scalable and cost-efficient method for enhancing the early phases of innovation. Our research paves the way for future exploration of how AI can be integrated into creative processes to foster more effective innovation.

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

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. We also appreciate participants at the HEC AI and Entrepreneurship conference, Utah Winter Strategy Conference, Berkeley Open Innovation Conference, USC AI in Management Conference, BU Online Seminar on Digital Businesses, WINDS Doctoral Consortium, the Data to Actionable Knowledge Lab and the Human-Computer Interaction groups at Harvard for their feedback. We used GPT-4 and Gemini Advanced for writing assistance. This work was supported by HBS DRFD and conducted within the Digital Data Design Institute at Harvard. All errors remain our own.

The best answer to the question, "Will computers ever be as smart as humans?" is probably "Yes, but only briefly."

—Vernor Vinge

Introduction

Organizations increasingly integrate artificial intelligence (AI) technologies into their work processes, leveraging their strengths in identifying patterns (Choudhury et al. 2021), making predictions (Agrawal et al. 2018, Kim et al. 2023), and conducting simulations (Horton 2023). These technological advancements have enabled AI to surpass human capabilities in a range of settings, such as healthcare (Lebovitz et al. 2022), criminal justice (Kleinberg et al. 2018), 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 is capable of creative problem-solving, which often requires 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). In this paper, we explore how generative AI—a type of artificial intelligence technology capable of producing new content based on human prompting—can improve problem-solving abilities in creative tasks.

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 yield economic and social returns (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).

Firms aiming to enhance their chances of innovative success can adopt a parallel path strategy that utilizes various approaches to creative problem-solving (Abernathy and Rosenbloom 1969, Leiponen and Helfat 2010, Nelson 1961). One effective approach to increase the number of parallel paths is through crowdsourcing, which involves engaging multiple independent problem solvers possessing diverse knowledge and alternative methods that may lead to effective solutions (Boudreau et al. 2011, Jeppesen and Lakhani 2010). The recent advances in generative AI 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, with its computational advantages and interactive conversational characteristics, presents a compelling option for producing a wide array of ideas economically and efficiently. Human collaborators can prompt the models to produce and simulate a diverse range of perspectives 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. In contrast, although crowdsourcing has previously been a viable solution to reduce costs and harness productivity gains 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 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 and value of solutions crowdsourced from humans to those generated by an individual strategically prompting a Large Language Model (LLM). LLMs are a subset of generative AI designed to understand and produce text based on extensive training from published text (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). 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) 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 challenging, 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 individuals, totaling 3,900 evaluator–solution pairs. Moreover, because evaluators were randomly assigned solutions, our estimated relationships between the solution source and the assessed novelty, value, and creativity of the solutions can be interpreted as causal. To demonstrate the variance in the capabilities of an HAI approach for creative problem-solving, we used different human prompt engineering techniques and AI model configurations.

Our findings indicate that HAI solutions exhibit a level of creativity comparable to those produced by the HC. A closer examination reveals that HAI solutions, on average, demonstrate potentially higher environmental and economic value. In contrast, the HC solutions are characterized by a higher level of novelty—both on average and among the statistically rarer ideas in the upper tail of the rating distribution. Moreover, we investigate the impact of model configuration on the novelty, value, and creativity of the responses. Our analysis demonstrates that simple instructions reminding the LLM to produce unique responses effectively enhance the novelty of HAI responses without compromising their value. Furthermore, in terms of creative potential, our observations suggest that using strategic prompt engineering, the outputs from the HAI partnership exhibit marginally higher creativity compared to the HC.

Overall, our study contributes to the strategy and innovation literature by demonstrating how the collective endeavors between humans and AI can augment organizational creative problem-solving. As pursuing multiple problem-solving approaches is likely to increase innovative success (Abernathy and Rosenbloom 1969, Leiponen and Helfat 2010), the potential for workers to generate a multitude of ideas using AI offers unprecedented opportunities for accelerating innovation efforts economically and efficiently. Moreover, emerging techniques such as prompt engineering that effectively guide models while allowing space for human creativity offer an "AI-in-the-loop" strategy, where AI becomes an integral part of the human workflow, to realize the promise of these technologies.

Our study, nevertheless, only illustrates one type of interactive process where one human works with AI through prompt designs. We anticipate the evolution of generative AI technologies, when paired with collective human minds and innovative interaction modes, will further expand the scale, speed, and range of ideas generated. Consequently, this paper aims to provide a framework for integrating generative AI strategically and purposefully to augment creative problem-solving processes.

Creative Problem Solving and The Benefits of Parallel Paths

According to a statistical view of innovation, the success of a firm's problem-solving efforts is closely tied to the ability to find an extreme-value outcome (Boudreau et al. 2011, Dahan and Mendelson 2001, Nelson 1961, Terwiesch and Xu 2008). Whereas most ideas are clustered around the mean, the right tail of the quality distribution corresponds to those that are statistically rare (Dahan and Mendelson 2001, Terwiesch and Ulrich 2009). These rare ideas are distinguished by their high degree of creativity in terms of their novelty and value (Amabile 1983, Fleming et al. 2007, Lingo and O'Mahony 2010, Rindova and Petkova 2007). Consistent with this view, increasing the number of independent approaches, or "parallel paths," can improve overall creative performance when there is uncertainty over the best way to solve a problem (Abernathy and Rosenbloom 1969, Leiponen and Helfat 2010, Nelson 1961). The parallel path effect suggests that incorporating a greater number of problem-solving methods increases the likelihood of achieving extreme outcomes (Boudreau et al. 2011, Dahan and Mendelson 2001). Arguably, utilizing a variety of different strategies is particularly critical when the objective is to maximize the creative performance of a few top ideas as opposed to many average ones (Girotra et al. 2010).

Crowdsourcing contests are one creative problem-solving approach that leverages a diverse pool of solvers with differing backgrounds and experiences to increase the number of alternative perspectives available to address the problem (Jeppesen and Lakhani 2010, Lifshitz-Assaf 2018, Piezunka and Dahlander 2015). By creating many parallel paths, firms can generate a larger set of ideas that also expand the range of their quality. Accordingly, crowdsourcing enhances the odds of identifying a novel and valuable idea that falls on the extreme, right tail of the distribution (Terwiesch and Ulrich 2009). 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 extreme 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, although crowdsourcing has been a highly effective approach for enhancing the parallel path effect, it has some drawbacks.

Using Large Language Models (LLMs) to Advance Human-AI Creative Problem Solving

This section offers an overview of generative AI and its capacity for facilitating creative problem-solving. We start by providing an understanding of generative AI technologies, with a particular emphasis on LLMs. Subsequently, we elaborate on how interactions between the human user and the AI, known as prompt engineering, can lead to the generation of creative outputs. Lastly, we outline the potential advantages and disadvantages of employing LLMs to generate creative ideas.

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 and composing music to creating realistic images or 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 prominent type of generative AI specializing in natural language processing and generation. Analyzing text datasets allows them to produce coherent, contextually relevant text that closely mimics human writing. LLMs commonly use transformer neural network architectures (Vaswani et al. 2017), which excel at handling long-range dependencies in language due to their self-attention mechanisms that dynamically weigh the importance of different words within a sequence to capture context (Ash and Hansen 2023, Bahdanau et al. 2014). LLMs are trained on a wide-ranging compilation of proprietary and internet-sourced datasets, encompassing literature, scientific texts, online forums, and news outlets, that cover an extensive array of topics, genres, and subjects in multiple languages. This diverse training helps them develop a comprehensive linguistic base. They learn via self-supervised learning, where the model refines its language comprehension by predicting future elements or filling in masked parts of the text. Gradient-based algorithms play a critical role in the training of LLMs. These optimization algorithms function by iteratively updating the model's internal parameters to minimize the discrepancy between predicted outcomes and target values. The gradient serves as a directional indicator, guiding the optimization process toward a solution that progressively reduces errors.

LLMs can undergo further refinement through supervised learning methodologies. This involves providing the model with human-annotated datasets that exemplify desired behaviors or task-specific outputs. Supervised learning enables the LLM to fine-tune its responses, improving its performance in targeted domains. Finally, alignment techniques (Bai et al. 2022, Ouyang et al. 2022) are often employed to promote helpful, safe, and human-aligned responses. This multi-stage training process underpins LLMs' versatility and high performance across numerous language-based tasks. LLMs have already demonstrated remarkable capabilities in a variety of organizational tasks. This includes enhancing performance on knowledge-intensive tasks for certain employees or activities (Brynjolfsson et al. 2023, Dell'Acqua et al. 2023), offering mentorship and advice to entrepreneurs (Otis et al. 2023), and augmenting creativity in well-defined domains, such as consumer products (Girotra et al. 2023).

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 B for a detailed overview of the inference processes of LLMs). Its primary strength lies not only in its ability to generate fluent and coherent text efficiently with a relatively small marginal cost (Brand et al. 2023, Horton 2023) but also in understanding and manipulating task instructions in various ways, such as summarizing, translating, or answering an extremely broad set of questions (Bubeck et al. 2023), and producing responses that encompass a spectrum of linguistic styles and perspectives (McCoy et al. 2023).

Strategic Prompt Engineering. Prompt engineering plays a crucial role in effecting the full potential of LLMs, emphasizing the interplay between the human user and AI capabilities (Saravia 2022). LLMs currently lack independent agency, so the quality and relevance of their outputs hinge significantly on the skill with which humans craft prompts (Zamfirescu-Pereira et al. 2023). This collaborative process between the user and the model is essential (Zamfirescu-Pereira et al. 2023). Using strategic prompt engineering, a user can direct the model to leverage specific aspects of its training, revealing the nuanced impact of input phrasing on the resulting outputs, even when the prompts contain similar objectives or semantic meaning (Brown et al. 2020). The strategic formulation of prompts can lead to a variety of creative outcomes by exploiting the LLM's ability to parse and process data in ways that a human mind alone might not conceive (Battle and Gollapudi 2024, OpenAI 2024). For instance, by adjusting the prompt to emphasize certain aspects of a problem or to explore it from unconventional perspectives, an LLM can develop solutions combining elements from disparate knowledge domains, thus fostering out-of-the-box thinking (Girotra et al. 2023, Meincke et al. 2024). Hence, prompt engineering may enable the user to leverage the LLM's knowledge base and pattern recognition abilities to enable a synergistic collaboration that can push the boundaries of traditional problem-solving approaches, such as human crowdsourcing, to produce novel and valuable solutions.

Anticipated Cost-Benefit Implications. The advanced capabilities of LLMs, such as GPT-4, indicate a strong potential for application in creative problem-solving. Notably, LLMs may streamline the idea generation and innovation processes, making them both more cost-effective and efficient (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 contribute to creative tasks without additional incentives. Moreover, LLMs can rapidly generate solutions at a larger scale, substantially enriching the idea pool in much less time compared to conventional human crowdsourcing methods. They also provide the benefit of delivering consistently high quality outputs, addressing the issue of variable submission quality often encountered with human solutions (Dahlander and Piezunka 2020, Piezunka and Dahlander 2015). These attributes position LLMs as a powerful tool in the creative process.

Although they possess remarkable abilities, LLMs also present specific limitations that can hinder their effectiveness in creative problem-solving applications. One notable issue is "confabulation," where the model generates plausible but factually incorrect or fabricated information, which can reduce the value or appropriateness of proposed solutions (Ji et al. 2023). Additionally, LLMs sometimes struggle with context, failing to maintain coherence over extended conversations or challenging problem scenarios. They may lack domain-specific expertise or access to private information, both critical for excelling in creative problem-solving tasks (Amabile 1983)—particularly as such tasks span multiple domains and are characterized by a high degree of uncertainty in their formulation (Boudreau et al. 2011, Leiponen and Helfat 2010). These limitations mean that while LLMs can assist in the ideation phase by generating a breadth of ideas, little is known about the relevance, accuracy, and appropriateness of the solutions, particularly when addressing nuanced, context-sensitive challenges. The impact of LLMs on creative problem-solving is a rapidly evolving area of research, given the recent emergence of these applications.

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 four criteria: *Novelty*, *Environmental Value*, *Financial Value*, and *Feasibility and Scalability of Implementation*.

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. Including a $75 platform fee, the crowdsourcing challenge had a total cost of $2,555. Of the 148 submissions, the research team deemed 125 eligible after filtering through solutions that were on-topic and sufficiently detailed to offer viable solutions.

LLM Idea Generation with Prompt Engineering. We use GPT-4 to generate various solutions in response to the same crowdsourcing challenge of developing sustainable, circular economy business ideas. Prompt engineering, which involves creating effective input prompts for the LLMs (Brown et al. 2020), greatly affects the AI's output quality and relevance (Battle and Gollapudi 2024, OpenAI 2024). The field of prompt engineering is rapidly developing, and our methods are based on early work: as of mid-2023, when this research took place, techniques like one-shot or few-shot prompting, Chain-of-Thought processes, and role-playing prompts were gaining traction. One-shot or few-shot prompting introduces a small set of examples to an LLM before it completes a task (Brown et al. 2020). This incremental addition of context helps anchor the model's understanding, providing clarity on the expected output format and the nature of the task. The "chain-of-thought" technique asks the model to break down complex problems into smaller steps and provide intermediate reasoning (Wei et al. 2023). Moreover, LLMs demonstrate intriguing role-playing abilities, where they can adopt a human persona or represent various entities (Shanahan et al. 2023). Early research suggests that role-playing prompts could implicitly trigger Chain-of-Thought processes and potentially enhance LLM's reasoning capabilities across various benchmarks (Kong et al. 2023). For all prompt engineering, we used the default GPT-4 temperature of 1.0.

Considering the evolving nature and understanding of prompt engineering, we use three alternative prompt engineering approaches to produce HAI solutions. 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 aims to simulate the contextual richness of HC, potentially enhancing the model's alignment with human-produced solutions and stimulating different creative answers. Finally, we role-play with expert, famous personas from 23 circular economy-relevant industries. This aims to steer the model toward mimicking diverse expert knowledge bases, fostering deeper reasoning and the potential for innovative, industry-specific solutions grounded in practical applications. Appendix A details the specific prompts used to generate the HAI solutions.

As recent work suggests that LLMs may produce homogenized outputs, potentially reducing the diversity of ideas (Dell'Acqua et al. 2023, Doshi and Hauser 2023, Stevenson et al. 2022), we implemented our three prompt engineering approaches through two alternative configurations of GPT-4 aimed at diversifying outputs: (1) multiple instance solutions and (2) single instance solutions with differentiation instruction. To the best of our knowledge, we are the first to report on the use and impact of these alternative configurations.

For the first configuration, each distinct instance of GPT-4 generates its solution independently from the same input prompt. While the model and prompt remain identical across instances, the sampling methods intrinsic to LLMs mean 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 B.2 for sampling methods of LLMs).

The second configuration leverages an iterative prompting scheme. In this process, the human engages in back-and-forth interactions with the LLM, refining and editing the prompts in multiple rounds to arrive at a desired output. We use a single instance of GPT-4 to generate multiple solutions successively, one at a time, adding the following sentence in the context prompt:

> *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 in the user prompt between each new generation while also including all the previously generated solutions:

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

By introducing a differentiation instruction between successive responses, a single instance of GPT-4 will attempt to diversify its successive responses from previous ones, enabling a potentially broader exploration of the search space compared to multiple instances, as LLMs tend to generate similar patterns with the same initial prompt. We aim to promote solution diversity, reduce redundancy, encourage originality, and stimulate creativity. Our technique is inspired by “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). This mechanism guides the LLM toward the desired outcome more accurately and efficiently.

Intuitively, the first configuration more closely mimics the concept of independent crowd solvers, as each instance of GPT-4 operates independently, starting from a different initialization and possibly exploring different areas of the problem and solution space. In contrast, the second configuration resembles an individual solver who iteratively proposes distinct ideas. 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. Table A1 provides sample HC and HAI prompt-engineered solutions.

Evaluator Recruitment and Procedures

Our study (approved under Harvard University IRB23-0770) uses human evaluators to judge the novelty and value of human and GPT-4 solutions. First, as shown in Figure 1, 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 E 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. For the HAI solutions, we randomly selected a mix of HAI-generated responses, instructed with three alternative prompts, and evenly allocated between multiple and single instance configurations (see Appendix A).

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 level. As shown in Figure 1, 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?*), environmental value (*How much does it benefit the planet?*), and financial value (*What financial value can it create for businesses?*). Overall, each block was evaluated 16.67 times on average (min = 15, max = 18, s.d. = 0.88).

--- Insert Figure 1 here ---

As a motivation to exert effort and to ensure thoroughness, we offered each evaluator $12 for completing the survey, with a bonus of $1 for each solution where they matched the consensus quality rating (evaluators were asked to give an overall quality for each assigned solution, and the consensus was defined using the mode). 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 three main dependent variables, corresponding to the evaluator's *Novelty rating*, *Value rating*, and *Creativity rating* of each solution. We computed the *Value rating* by taking the average of the evaluator's environmental and financial value ratings and the *Creativity rating* as the *Novelty rating* multiplied by the *Value rating* (*Novelty rating* x *Value rating*) (Poetz and Schreier 2012). To examine extreme outcomes, we created binary variables for *Top novelty rating, Top value rating*, and *Top creativity rating*. Each of these binary variables is set to one if the solution received the top rating, and zero otherwise.

Independent Variables. Our main independent variable, *HAI solution*, 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 between the HAI solutions as either *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, HC achieves a statistically higher average novelty score than HAI Multiple instance ($p = 0.013$) but becomes statistically indifferent from HAI Single instance. In terms of average creativity, HAI and HC are comparable (Multiple instance: $p = 0.624$; Single instance: $p = 0.317$). At the top decile, HC exhibits higher novelty and creativity, but lower value compared to HAI (Novelty: $p < 0.001$; Value: $p = 0.106$; Creativity: $p < 0.001$).

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

Estimation Approach

We analyzed 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 (OLS) in the presence of nested data (Gelman and Hill 2006).

Results

Mixed Effects Models

Tables 3–8 report the mixed model results of *Novelty rating* (Table 3) and *Top novelty rating* (Table 4), *Value rating* (Table 5), and *Top value rating* (Table 6), as well as *Creativity rating* (Table 7) and *Top creativity rating* (Table 8) on solution source. In all tables, the main independent variable in Models 1–3 is *HAI solution*, and the main independent variable in Models 4–6 is *HAI instance*. Both Models 1 and 4 report the main effect of the solution source. Models 2 and 5 add the evaluator screening criteria, and Models 3 and 6 control for additional evaluator attributes, cohort, and solution word count. For *Top novelty rating*, *Top value rating*, and *Top creativity rating*, we additionally report mixed effects logistic regression model results in Appendix C.

Estimated Relationships Between Solution Sources and Solution Novelty. In Table 3, Model 1 indicates that compared to HC solutions, the HAI solutions receive a lower novelty rating on average (Model 1: -0.124, $p < 0.001$). Models 2 and 3 indicate that the estimated coefficient remains stable and robust after adding the evaluator screening criteria (Model 2: -0.124, $p < 0.001$) as well as the evaluator attributes, cohort, and solution word count controls (Model 3: -0.140, $p < 0.001$).

In Model 4, we model the solution source as a categorical variable to differentiate between HC solutions and HAI multiple and single instance solutions. Compared to the HC solutions, we observe that the HAI solutions generated with multiple instances of GPT-4 are rated as significantly less novel (Model 4: -0.209, $p < 0.001$), but there is no difference between the HC and single instance HAI solutions (Model 4: -0.039, *ns*). 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 4 are significantly different from each other ($p < 0.001$). Next, Models 5 and 6 indicate that the estimated relationships remain significant after adding the evaluator screening criteria (HAI Multiple instance: -0.209, $p < 0.001$; HAI Single instance: -0.039, *ns*) and the evaluator demographic attributes, cohort, and solution word count controls (HAI Multiple Instance: -0.217, $p < 0.001$; HAI Single Instance: -0.056, *ns*).

Next, we turn to Table 4 to investigate the relationships between the *most* novel solutions, which achieved the *Top novelty rating*, and the solution source. In Model 1, we observe that, compared to HC responses, HAI solutions are 7.4 percentage points (pp) less likely to receive the top novelty rating (Model 1: -0.074, $p < 0.001$). Models 2 and 3 show that the estimated relationships remain robust after adding the evaluator covariates, cohort, and solution word count controls. Model 4 splits the HAI solutions into m*ultiple* and *single instance* solutions and indicates that both configurations are less likely than the HC solutions to receive the top novelty rating (HAI Multiple Instance: -0.088, $p < 0.001$; HAI Single Instance: -0.059, $p < 0.001$). Once again, Models 5 and 6 indicate that the reported coefficients remain consistent and robust with the evaluator screening criteria and other controls.

Although Tables 3 and 4 highlight the higher perceived novelty of the HC solutions compared to the HAI ones, instructing GPT-4 to differentiate its responses within a single instance configuration demonstrates its potential to produce notably more novel outputs on average than prompting with multiple or parallel instances.

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

Estimated Relationships Between Solution Sources and Solution Value. Turning to Table 5, Model 1 indicates that HAI solutions are rated as more valuable than HC solutions (Model 1: 0.171, $p < 0.001$). We observe that this estimated relationship remains statistically significant in Models 2 and 3, which add the evaluator screening criteria (Model 2: 0.171, $p < 0.001$) and evaluator and solution controls (Model 3:

0.152, $p < 0.001$), respectively. Next, in Model 4, we use the categorical variable, *HAI instance*, to differentiate between the HC, HAI multiple and HAI single instance solutions. We observe that compared to HC responses, both the HAI multiple and single instance solutions are rated as more valuable (HAI Multiple instance: 0.160, $p < 0.001$; HAI Single instance: 0.182, $p < 0.001$). A post hoc pairwise comparison of coefficients indicates that the multiple and single instance coefficients are not significantly different from each other ($p = 0.622$). We note that the estimated relationships are unchanged in Models 5 and 6, which add evaluator screening criteria (HAI Multiple Instance: 0.160, $p < 0.001$; HAI Single Instance: 0.182, $p < 0.001$) and evaluator demographic covariates, cohort, and solution word count (HAI Multiple Instance: 0.148, $p < 0.001$; HAI Single Instance: 0.156, $p < 0.001$).

Next, we investigate the relationships between achieving the *Top value rating* and the solution source. Table 6 Model 1 shows no significant difference between the HC and HAI regarding their likelihood of generating a highly valuable solution (Model 1: 0.019, *ns*). Models 2 and 3 indicate that there is once again no difference in the top value rating between the HC and HAI. Turning to Model 4, we observe that, compared to the HC responses, there is no difference between the HAI multiple and single instance configurations and the likelihood of generating a highly valuable solution (HAI Multiple Instance: 0.020, *ns*; HAI Single Instance: 0.017, *ns*). Models 5 and 6 indicate no change in the estimated relationships.

In summary, in Tables 5 and 6, we find that the HAI responses achieved higher value ratings on average than the HC solutions. However, there is no difference in top value between the solutions produced by the HC and HAI, and the multiple and single instance HAI configurations do not have a meaningful effect on the solution's value. One possible explanation is that the differentiation prompt in the single instance configuration will likely force different or unique answers that push the model towards greater novelty without changing the value of their outputs. An important insight of the single instance configuration is that we can achieve more novel responses (see Tables 2 and 3) without compromising the perceived value of the responses. We note that in supplementary analyses (Tables C5-C8), the reported results are robust across alternative specifications of the *Value rating* as separate dimensions corresponding to *Environmental* and *Financial value*.

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

Estimated Relationships Between Solution Sources and Solution Creativity. Turning to Table 7, Model 1 indicates no difference in creativity between the HC and HAI solutions (Model 1: 0.066, *ns*). These patterns remain consistent in Models 2 and 3. Next, in Model 4, we observe that there is a negative but not significant difference between the HC and HAI multiple instance solutions (Model 4: -0.232, *ns*), and a positive and marginally significant difference between the HC and HAI single instance solutions (Model 4: 0.363, $p < 0.10$). The post hoc pairwise comparison of coefficients indicates that the HAI multiple and single instance coefficients are significantly different from each other ($p < 0.001$). These patterns remain consistent in Models 5 and 6 which add the evaluator screening criteria and the evaluator demographic covariates, cohort, and solution word count controls, respectively. As the HAI single instance configuration differentiates successive responses from previous ones, the observed patterns in the data corroborate that it allows for a potentially deeper exploration of the solution space compared to the HAI multiple instance configuration. Consequently, the HAI single instance configuration may achieve higher creativity levels on average, suggesting that it is a viable, cost-efficient, and productive alternative to creative problem-solving that is comparable to the multiple perspectives of the HC.

Next, we investigate the relationships between the most creative solutions, achieving the *Top creativity rating*, and the solution source. Table 8 Model 1 shows no significant difference between HC and HAI regarding their likelihood of generating a highly creative solution (Model 1: -0.004, *ns*). These patterns remain robust in Models 2 and 3. Turning to Model 4, we observe that, compared to the HC responses, there is no difference between the HAI multiple and single instance configurations and the likelihood of generating a highly creative solution (HAI Multiple Instance: -0.003, *ns*; HAI Single Instance: -0.005, *ns*). There is no change in the reported relationships in Models 5 and 6.

In summary, in Tables 7 and 8, we find no meaningful difference between the level of creativity in the HC and HAI solutions. The data suggest that HAI collaboration, enabled by prompt engineering, could yield creative outputs comparable to those generated by human solvers. In Tables C9-C14, we demonstrate that the different prompt engineering approaches applied in this study can uncover subtle differences in the solutions, potentially steering the LLM responses towards greater novelty or value. Overall, our findings underscore the valuable role that HAI collaborations may play in creative problem-solving.

--- Insert Tables 7 and 8 here ---

Discussion

We began this paper with the following question: In the era of generative AI, how might humans use this technology effectively to advance creative problem-solving? In particular, it remains unclear whether an individual worker's use of generative AI can result in the development of creative outputs to address open-ended and challenging organizational and societal problems (Ivcevic and Grandinetti 2024). To investigate this question, we partnered with Continuum Lab, an AI firm, to launch a crowdsourcing challenge to identify sustainable, circular economy business opportunities, and compared the novelty and value of their outputs to those generated with HAI collaboration, facilitated with different prompt engineering approaches. We subsequently invited human evaluators to assess the novelty and value of the submitted solutions without revealing their sources as HC or HAI generated.

We find that HAI solutions are capable of producing outputs that achieve comparable levels of creativity to the HC. This finding aligns with recent research, indicating that outputs generated through HAI collaboration are nearing human levels of creativity (Doshi and Hauser 2023, Franceschelli and Musolesi 2023, Girotra et al. 2023, Gómez-Rodríguez and Williams 2023, Guzik et al. 2023). We extend this work to provide a detailed look at the distribution of the solutions, which suggests that HAI solutions are on average, higher in value. In contrast, the HC solutions are higher in novelty—both on average and among those that are statistically rare at the upper right tail of the outcome distribution. Additionally, we assess the impact of model configurations on the novelty, value, and creativity of the responses. We find that a simple instruction reminding GPT-4 to produce unique responses can effectively elevate the novelty of the HAI responses without compromising their value. Our findings suggest HAI outputs possess greater creative potential than HC outputs on the margins under the single instance configuration. 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 indicates that generative AI presents a promising, time- and cost-effective alternative to creative problem-solving.

The Future of Human-AI Creative Problem Solving

Our study offers evidence that an individual working with AI can produce creative outputs that are comparable to those produced by HC solvers. These findings have several implications for revolutionizing the status quo of the creative problem-solving process with generative AI. The process of identifying multiple parallel paths to find an effective solution can entail considerable costs, particularly when the problem to solve goes outside a firm's core competencies (Henderson and Cockburn 1994, Katila and Ahuja 2002, Tushman and Anderson 1986). In this study, we closely examined the capabilities of human-guided AI collaboration in creative problem-solving and compared it to the outputs of the HC. We chose crowdsourcing as a context for our study because it is a proven strategy to economize on internal resources and to broaden the range of independent solutions to a problem (Boudreau et al. 2011, Dahan and Mendelson 2001). However, crowdsourcing is not without limitations. This idea generation model often necessitates providing incentives to encourage participation (Jeppesen and Lakhani 2010), and the need to manage a large volume of submissions, including low-quality ones (Bell et al. 2024). This process has been shown to overburden the cognitive capacities of human decision-makers and contest organizers, leading to a less efficient selection process (Piezunka and Dahlander 2015). Notably, in our study, the distribution of HC outputs would have a more pronounced left tail if not for the initial filtering of incomplete and off-topic solutions. In comparison, our model of HAI creative problem-solving offers a cost-effective and scalable approach to generate solutions while guaranteeing a minimum threshold of quality without the need to incentivize and manage a large number of participants.

It is important to note that HAI achieved levels of creativity comparable to the HC, despite using relatively simple prompt engineering approaches. We recognize that organizations are likely to employ more sophisticated methods to enhance the performance of LLMs. Consequently, it is reasonable to anticipate improvements in the quality of HAI solutions over time as prompt engineering or similar interaction approaches become more refined. Therefore, rather than replacing human creativity, our findings imply that generative AI could be a powerful tool within the idea-generation phase. This may allow organizations to strategically focus resources on later stages of innovation, such as solution refinement and implementation (Perry-Smith and Mannucci 2017). Yet, consistent 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). More specifically, our findings suggest that collaborations between humans and AI can steer responses toward more valuable outcomes, which may be attributable to the fine-tuning and alignment techniques used to train the LLMs. This observation raises the concern of a decrease in the generation of unique and innovative ideas. This is especially worrying if evaluators primarily value ideas that align with established success patterns, potentially leading to a bias toward replicating past successful solutions–namely, incremental innovations (Dewar and Dutton 1986), rather than groundbreaking breakthroughs.

Another concern with the application of generative AI in creative problem-solving involves navigating the complexities of intellectual property (IP) rights and addressing environmental impacts. The evolving landscape of international copyright laws regarding AI-generated content necessitates novel approaches to determining ownership and fair use. From an environmental perspective, the substantial computing power required for generative AI underlines the need for a proactive stance on energy consumption, focusing on sustainability and carbon emissions reduction (Kumar and Davenport 2023). These legal and environmental complications demand a re-evaluation of IP regimes, the development of environmentally sustainable AI practices, and an ongoing ethical dialogue to guide responsible AI deployment.

Moreover, as AI technologies become integrated into the creative process, humans may experience shifts in their roles and responsibilities, prompting reflections on the nature of creativity and humans' unique contributions. These potential paradigm shifts underscore the importance of balancing between human agency and AI augmentation. Despite its promising potential, it is critical to integrate AI as a support tool within the human creative problem-solving process, to enhance rather than replace our capabilities.

Methodological Considerations, Limitations, and Future Directions

Although our study utilized a highly practical approach to study HAI creative problem-solving in a rapidly advancing field, our study is nevertheless subject to limitations that open the door 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 different international contexts. Second, another limitation is 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 understanding that experts bring, potentially leading to an underestimation or overestimation of the solutions' novelty and applicability. Both these limitations offer promising directions for future research–expanding the evaluators pool to a more globally diverse representation of crowds and experts.

Third, the creativity of the HAI outputs in our study may have been influenced by the training data, the model setting, and a limited number of prompt engineering strategies. 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). Furthermore, the effectiveness of different prompt engineering techniques, such as few-shot learning, step-by-step guidance ("chain-of-thought"), and iterative prompting ("prompt-chaining"), significantly influences the novelty and value of the solutions generated by LLMs (Meincke et al. 2024, Zhou et al. 2022). Prompts designed to elicit depth, contrarian views, and creative thinking are particularly effective in enhancing the quality of LLM outputs, underscoring the importance of sophisticated prompts in leveraging AI for creative problem-solving (Meincke et al. 2024). Moreover, the training data cut-off for GPT-4, set in September 2021 during our period of data collection, indicates that it might not encompass the latest developments, emerging trends, and knowledge. This lag in the training data could have affected the LLM's responses in terms of perceived originality and applicability relative to the HC responses.

Lastly, in our research, we focused on the capabilities of a single LLM. One possibility is to incorporate 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 (RAG) systems could enable LLMs to access and process external knowledge bases, enhancing factual accuracy and enriching their responses (Lewis et al. 2020). 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, their collaborative output could offer more creative recombinations than a single response from GPT-4.

Despite these limitations, our findings have important implications for creative problem-solving, as they demonstrate the feasibility of HAI models of interaction and idea generation. By providing a proof of concept, our study lays the groundwork for leveraging HAI collaboration to generate multiple parallel paths for approaching solutions effectively and efficiently. This approach holds promise for enhancing the creative problem-solving process and unlocking new avenues for innovative activities. Looking forward, the rapid advancement in the capabilities of generative AI holds tremendous promise for enhancing collaborations between humans and AI. This synergistic integration, known as AI-in-the-loop, has the potential to transform creative problem-solving at scale.

References

Abernathy WJ, Rosenbloom RS (1969) Parallel strategies in development projects. *Manag. Sci.* 15(10):B-486-B-505.
Agrawal A, Gans J, Goldfarb A (2018) *Prediction Machines: The Simple Economics of Artificial Intelligence* (Harvard Business Review Press).
Amabile TM (1983) The social psychology of creativity: A componential conceptualization. *J. Pers. Soc. Psychol.* 45(2):357.
Ash E, Hansen S (2023) Text algorithms in economics. *Annu. Rev. Econ.* 15:659–688.
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.
Battle R, Gollapudi T (2024) The Unreasonable Effectiveness of Eccentric Automatic Prompts.
Becker GS (1994) *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Educatio* 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.
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.
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.
Dahlander L, Piezunka H (2020) Why crowdsourcing fails. *J. Organ. Des.* 9:1–9.
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*.
Fleming L, Mingo S, Chen D (2007) Collaborative brokerage, generative creativity, and creative success. *Adm. Sci. Q.* 52(3):443–475.
Franceschelli G, Musolesi M (2023) On the creativity of large language models. *ArXiv Prepr. ArXiv230400008*.

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.
Henderson R, Cockburn I (1994) Measuring competence? Exploring firm effects in pharmaceutical research. *Strateg. Manag. J.* 15(S1):63–84.
Horton JJ (2023) *Large language models as simulated economic agents: What can we learn from homo silicus?* (National Bureau of Economic Research).
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 creative performance 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.
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*
Kim H, Glaeser EL, Hillis A, Kominers SD, Luca M (2023) Decision authority and the returns to algorithms. *Strateg. Manag. J.* n/a(n/a).
Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S (2018) Human decisions and machine predictions. *Q. J. Econ.* 133(1):237–293.
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.
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).
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.
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. *Evid. Large-Scale Exam. Bio-Pharma FirmsMarch 15 2021 MISQ Forthcom.*
McCoy RT, Yao S, Friedman D, Hardy M, Griffiths TL (2023) Embers of autoregression: Understanding large language models through the problem they are trained to solve. *ArXiv Prepr. ArXiv230913638*.
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.
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.
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*
Shanahan M, McDonell K, Reynolds L (2023) Role play with large language models. *Nature* 623(7987):493–498.
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*.
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
1	Novelty	3.412	3	1.047	1	5	1.000												
2	Value	3.482	3.5	0.851	1	5	0.476	1.000											
3	Creativity	12.306	12	5.606	1	25	0.878	0.811	1.000										
4	HAI	0.769	1	0.421	0	1	-0.050	0.085	0.005	1.000									
5	HAI M/S*	1.154	1	0.769	0	2	0.000	0.076	0.031	0.822	1.000								
6	Experience	4.680	3.5	5.172	0	45.5	0.001	0.022	0.018	0.000	0.000	1.000							
7	Interest	3.913	4	0.16	3	5	0.061	0.124	0.108	0.000	0.000	0.134	1.000						
8	Score	2.447	2.5	1.158	0	5	-0.069	-0.119	-0.108	0.000	0.000	-0.156	-0.058	1.000					
9	Female	0.370	0	0.483	0	1	0.039	0.069	0.059	0.000	0.000	-0.161	0.039	0.235	1.000				
10	Bachelor’s	0.617	1	0.486	0	1	-0.018	-0.015	-0.021	0.000	0.000	0.034	0.076	0.038	0.008	1.000			
11	STEM	0.467	0	0.499	0	1	0.001	0.004	0.000	0.000	0.000	0.142	0.009	-0.084	-0.080	0.064	1.000		
12	Employed	0.860	1	0.347	0	1	0.026	0.010	0.021	0.000	0.000	0.127	0.063	-0.102	-0.208	0.235	0.012	1.000	
13	Cohort	0.517	1	0.499	0	1	0.101	0.128	0.135	0.000	0.000	0.160	0.085	-0.151	-0.074	-0.008	0.022	0.129	1.000
14	Word Count	237.769	238	114.243	35	1049	0.030	0.063	0.056	0.164	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 Length of Solutions	204 words	231 words	265 words
Average Novelty (out of 5)	3.508	3.230	3.469
Standard Deviation of Novelty	1.127	1.040	0.993
Average Value (out of 5)	3.351	3.510	3.533
Standard Deviation of Value	0.917	0.837	0.815
Average Creativity (out of 25)	12.300	12.000	12.600
Standard Deviation of Creativity	5.900	5.590	5.360
Average Novelty of Top Decile	4.360	3.900	4.000
Average Value of Top Decile	3.880	3.950	3.930
Average Creativity of Top Decile	16.500	15.200	15.400
P-value		**vs. HC**	**vs. HC**
(Is the average novelty different?)		0.013	0.657 vs. multiple instance 0.002
(Is the average value different?)		0.002	0.001 vs. multiple instance 0.55
(Is the average creativity different?)		0.624	0.317 vs. multiple instance 0.025

Note: This table style is adapted from Girotra et al. (2023).

Table 3. Nested Mixed Effects Models of Evaluator Ratings of Novelty Rating on Solution Source (Human Crowd or Human-AI)

	Dependent Variable: Novelty Rating					
	(1)	(2)	(3)	(4)	(5)	(6)
HAI Solution	-0.124***	-0.124***	-0.140***			
	(0.035)	(0.035)	(0.035)			
HAI Multiple Instance				-0.209***	-0.209***	-0.217***
				(0.038)	(0.038)	(0.039)
HAI Single Instance				-0.039	-0.039	-0.056
				(0.038)	(0.038)	(0.039)
Intercept	3.508***	3.381***	3.242***	3.508***	3.381***	3.262***
	(0.046)	(0.180)	(0.215)	(0.046)	(0.180)	(0.215)
N	3900	3900	3900	3900	3900	3900
# blocks	18	18	18	18	18	18
# evaluators	300	300	300	300	300	300
Screening criteria	N	Y	Y	N	Y	Y
Other controls	N	N	Y	N	N	Y
Log-Likelihood	-5430.30 df = 5	-5434.96 df = 8	-5439.89 df = 15	-5419.78 df = 6	-5424.45 df = 9	-5430.94 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, with 300 evaluators nested in eighteen solution blocks. Models 2-3 and 5-6 include the screening criteria: Work Experience, Level of Interest, and Knowledge Test Score. Models 3 and 6 include the following covariates: Gender, Highest Level of Education, Major, Employment Status, Cohort Session, and Solution Word Count. Standard errors are in parentheses.

Table 4. Nested Mixed Effects Models of Evaluator Top Novelty Ratings on Solution Source (Human Crowd or Human-AI)

	Dependent Variable: Top Novelty Rating (0/1)					
	(1)	(2)	(3)	(4)	(5)	(6)
HAI Solution	-0.074***	-0.074***	-0.079***			
	(0.012)	(0.012)	(0.012)			
HAI Multiple Instance				-0.088***	-0.088***	-0.091***
				(0.014)	(0.014)	(0.014)
HAI Single Instance				-0.059***	-0.059***	-0.065***
				(0.014)	(0.014)	(0.014)
Intercept	0.208***	0.189***	0.176**	0.208***	0.189***	0.179**
	(0.015)	(0.057)	(0.068)	(0.015)	(0.057)	(0.068)
N	3900	3900	3900	3900	3900	3900
# blocks	18	18	18	18	18	18
# evaluators	300	300	300	300	300	300
Screening criteria	N	Y	Y	N	Y	Y
Other controls	N	N	Y	N	N	Y
Log-Likelihood	-1329.62 df = 5	-1338.23 df = 8	-1356.98 df = 15	-1330.04 df = 6	-1338.65 df = 9	-1358.06 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, with 300 evaluators nested in eighteen solution blocks. Models 2-3 and 5-6 include the screening criteria: Work Experience, Level of Interest, and Knowledge Test Score. Models 3 and 6 include the following covariates: Gender, Highest Level of Education, Major, Employment Status, Cohort Session, and Solution Word Count. Standard errors are in parentheses.

Table 5. Nested Mixed Effects Models of Evaluator Value Ratings on Solution Source (Human Crowd or Human-AI)

	Dependent Variable: Value Rating					
	(1)	(2)	(3)	(4)	(5)	(6)
HAI Solution	0.171***	0.171***	0.152***			
	(0.025)	(0.025)	(0.026)			
HAI Multiple Instance				0.160***	0.160***	0.148***
				(0.028)	(0.028)	(0.028)
HAI Single Instance				0.182***	0.182***	0.156***
				(0.028)	(0.028)	(0.029)
Intercept	3.351***	3.080***	3.017***	3.351***	3.080***	3.018***
	(0.038)	(0.172)	(0.200)	(0.038)	(0.172)	(0.200)
N	3900	3900	3900	3900	3900	3900
# blocks	18	18	18	18	18	18
# evaluators	300	300	300	300	300	300
Screening criteria	N	Y	Y	N	Y	Y
Other controls	N	N	Y	N	N	Y
Log-Likelihood	-4294.88 df = 5	-4293.91 df = 8	-4288.06 df = 15	-4297.25 df = 6	-4296.28 df = 9	-4290.79 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 value, with 300 evaluators nested in eighteen solution blocks. Models 2-3 and 5-6 include the screening criteria: Work Experience, Level of Interest, and Knowledge Test Score. Models 3 and 6 include the following covariates: Gender, Highest Level of Education, Major, Employment Status, Cohort Session, and Solution Word Count. Standard errors are in parentheses.

Table 6. Nested Mixed Effects Models of Evaluator Ratings of Top Value Rating on Solution Source (Human Crowd or Human-AI)

	Dependent Variable: Top Value Rating (0/1)					
	(1)	(2)	(3)	(4)	(5)	(6)
HAI Solution	0.019	0.019	0.012			
	(0.014)	(0.014)	(0.014)			
HAI Multiple Instance				0.020	0.020	0.016
				(0.016)	(0.016)	(0.016)
HAI Single Instance				0.017	0.017	0.007
				(0.016)	(0.016)	(0.016)
Intercept	0.268***	0.071	0.065	0.268***	0.071	0.064
	(0.019)	(0.085)	(0.102)	(0.019)	(0.085)	(0.102)
N	3900	3900	3900	3900	3900	3900
# blocks	18	18	18	18	18	18
# evaluators	300	300	300	300	300	300
Screening criteria	N	Y	Y	N	Y	Y
Other controls	N	N	Y	N	N	Y
Log-Likelihood	-1975.78 df = 5	-1977.58 df = 8	-1990.34 df = 15	-1979.13 df = 6	-1980.93 df = 9	-1993.51 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 value, with 300 evaluators nested in eighteen solution blocks. Models 2-3 and 5-6 include the screening criteria: Work Experience, Level of Interest, and Knowledge Test Score. Models 3 and 6 include the following covariates: Gender, Highest Level of Education, Major, Employment Status, Cohort Session, and Solution Word Count. Standard errors are in parentheses.

Table 7. Nested Mixed Effects Models of Evaluator Creativity Ratings (Novelty x Value) on Solution Source (Human Crowd or Human-AI)

Dependent Variable: Creativity Rating						
	(1)	(2)	(3)	(4)	(5)	(6)
HAI Solution	0.066	0.066	−0.064			
	(0.171)	(0.171)	(0.173)			
HAI Multiple Instance				−0.232	−0.232	−0.304
				(0.190)	(0.190)	(0.190)
HAI Single Instance				0.363+	0.363+	0.201
				(0.190)	(0.190)	(0.194)
Intercept	12.256***	10.775***	10.022***	12.256***	10.775***	10.087***
	(0.244)	(1.111)	(1.299)	(0.244)	(1.111)	(1.300)
N	3900	3900	3900	3900	3900	3900
# blocks	18	18	18	18	18	18
# evaluators	300	300	300	300	300	300
Screening criteria	N	Y	Y	N	Y	Y
Other controls	N	N	Y	N	N	Y
Log-Likelihood	-11721.69 df = 5	-11716.86 df = 8	-11699.42 df = 15	-11716.04 df = 6	-11711.21 df = 9	-11695.66 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 creativity (solution novelty x solution value), with 300 evaluators nested in eighteen solution blocks. Models 2-3 and 5-6 include the screening criteria: Work Experience, Level of Interest, and Knowledge Test Score. Models 3 and 6 include the following covariates: Gender, Highest Level of Education, Major, Employment Status, Cohort Session, and Solution Word Count. Standard errors are in parentheses.

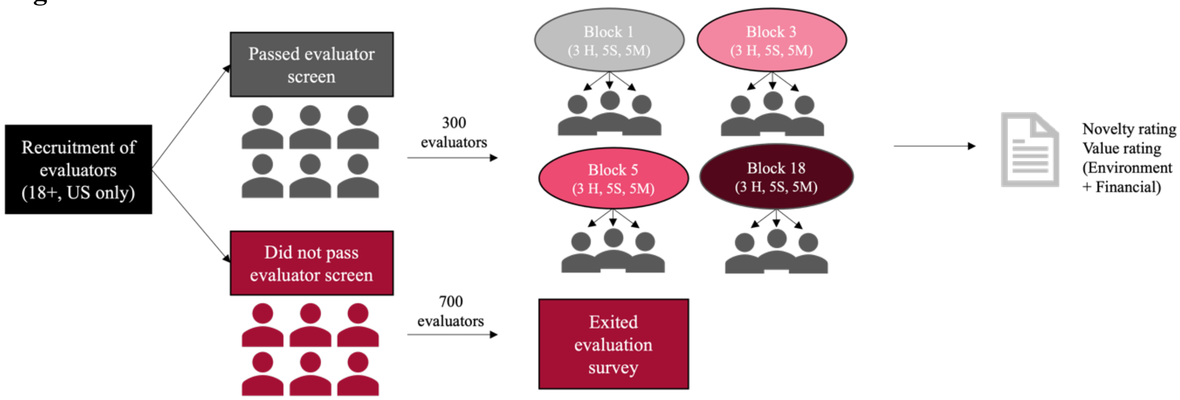
Table 8. Nested Mixed Effects Models of Evaluator Top Creativity Ratings (Novelty x Value) on Solution Source (Human Crowd or Human-AI)

Dependent Variable: Top Creativity Rating (0/1)						
	(1)	(2)	(3)	(4)	(5)	(6)
HAI Solution	−0.004	−0.004	−0.004			
	(0.006)	(0.006)	(0.006)			
HAI Multiple Instance				−0.003	−0.003	−0.003
				(0.007)	(0.007)	(0.007)
HAI Single Instance				−0.005	−0.005	−0.005
				(0.007)	(0.007)	(0.007)
Intercept	0.038***	0.030	0.030	0.038***	0.030	0.029
	(0.007)	(0.028)	(0.033)	(0.007)	(0.028)	(0.033)
N	3900	3900	3900	3900	3900	3900
# blocks	18	18	18	18	18	18
# evaluators	300	300	300	300	300	300
Screening criteria	N	Y	Y	N	Y	Y
Other controls	N	N	Y	N	N	Y
Log-Likelihood	1236.12 df = 5	1226.68 df = 8	1202.46 df = 15	1231.99 df = 6	1222.56 df = 9	1198.35 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 creativity, with 300 evaluators nested in eighteen solution blocks. Models 2-3 and 5-6 include the screening criteria: Work Experience, Level of Interest, and Knowledge Test Score. Models 3 and 6 include the following covariates: Gender, Highest Level of Education, Major, Employment Status, Cohort Session, and Solution Word Count. Standard errors are in parentheses.

Figure 1. 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 crowd (HC) and 10 human-AI (HAI) solutions. Of the HAI 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 the novelty, environmental, and financial value of each assigned solution. The value rating is an average of the environmental and financial ratings.

Generative AI and Creative Problem Solving

Appendix

Appendix A: Detailed Explanations Prompt Building Mechanisms and Model

As shown in Table A1, we use three prompt engineering levels to generate the AI solutions, comprising both multiple (M) and single (S) instance configurations for each level. This resulted in six distinct configurations by level and instance, corresponding to Levels 1S, 1M, 2S, 2M, 3S, and 3M.

- Level 1 uses the initial problem description for baseline comparison of human crowd (HC) and human-AI (HAI) solutions.
- Level 2 adds individual characteristics of the 125 human solvers, simulating the context of the original human crowd.
- Level 3 introduces expert persona prompting, leveraging GPT-4's vast data to generate diverse, industry-specific solutions by mimicking expert personas from 23 industries.

Table A1. Prompt Engineering Configurations Used to Generate HAI Solutions

Prompt Engineering Configurations	Description	Rationale
Level 1	The AI model receives the same problem description given to human participants or solvers.	This baseline allows for a direct comparison between HAI and HC responses, as both parties receive identical initial conditions.
Level 2	In addition to the original problem description, the AI model is given individual solver characteristics reflecting the human crowd (i.e., job title, geographical location, industry, solution maturity).	By adding individual-level characteristics, the AI model's context becomes more similar to the human crowd's, potentially generating solutions closer to those produced by HC.
Level 3	The AI model receives the original problem description and individual-level personas of experts from 23 industries relevant to the circular economy.	Incorporating expert personas from various industries encourages the AI model to emulate the diverse perspectives of a knowledgeable crowd. Additionally, it encourages the generation of creative industry-specific solutions.
Multiple Instance	Each distinct instance of GPT-4 generates its solution independently from the same input prompt.	Intuitively, a multiple instance configuration mimics the concept of independent crowd solvers, as each instance of GPT-4 operates independently.

Single Instance	A single instance of GPT-4 generates multiple solutions successively, one at a time, with a differentiation instruction between successive responses.	Intuitively, a single instance configuration resembles an individual solver who iteratively proposes distinct ideas.

Table A2. Examples of HC and HAI Problems and Solutions for Top Value by Configuration

Example Problem and Solution	Configuration
Problem: A significant portion of food waste is generated by restaurants, supermarkets, and households, which not only squanders resources but also adds to the growing problem of landfills. In particular, the disposal of food waste in landfills results in the generation of methane, a greenhouse gas twenty times more potent than carbon dioxide. Hence, finding an efficient and economic method to manage food waste reduces the need for new landfills and helps mitigate climate change. Solution: To address this issue, I propose the creation of local bio-energy centers. The idea is to partner with restaurants, markets, and households to collect their organic waste. This waste will then be converted into fertilizer for community gardens and parks, as well as biogas, which can be used to generate electricity. First, there would be waste collection bins at strategic locations so businesses and households can conveniently dispose of their organic waste. The waste would be regularly collected and brought to the bio-energy centers for processing. The organic waste, once in the bio-energy center, will be treated in a process called anaerobic digestion. Anaerobic digestion is basically a series of biological processes in which microorganisms break down biodegradable material in the absence of oxygen. This process results in the production of biogas, which consists mainly of methane. The methane generated can be used as a renewable source of energy, to generate heat and electricity, while nutrient-rich digestate (a byproduct of the process) can be used as a low-cost, efficient fertilizer to promote local agriculture and horticulture. This solution provides multiple benefits: reduction in waste that would have gone to landfills; production of clean, renewable energy; providing organic, chemical-free fertilizers for agricultural purposes; and contributing to the local economy. The feasibility and scalability of this model have been proven in several European countries. The financial impact is expressed in both cost savings (as waste management costs are reduced) and income sourcing (sale of energy, fertilizer). The environmental benefits are also clear, as this process drastically reduces the release of harmful greenhouse gases that contribute to climate change.	1M
Problem: The rapid depletion of natural resources, exacerbated by traditional linear economic practices, is a pressing issue. Consider the automobile industry,	2M

industrial equipment, metal products, and the construction sector. These industries, though economically significant, are notorious for contributing significantly to environmental degradation. The issues span from resource mining, energy-intensive manufacturing processes, and end-product disposal, often in landfills. There's a pressing need for innovative strategies to extend the lifecycle of products, reduce waste, and conserve natural resources. Solution: My solution proposes the 'Resource Lifecycle Extension and Re-utilization Platform' (ReLEaRP). This would function as a multi-sector digital marketplace combining the aspects of sharing, leasing, reusing, repairing, refurbishing and recycling of products and materials, thus encapsulating the concept of a Circular Economy. In this ecosystem, businesses can list their waste materials or old equipment, which might be useful to other industries. For instance, the 'waste' metal from an automobile company could be valuable raw material for a construction firm that uses metal extensively. Businesses can also offer their idle assets for lease, fostering resource sharing. The platform would also facilitate repair and refurbish services. Businesses with impaired equipment can link with repair service providers to get them fixed rather than disposed of. Similarly, used products can be refurbished and re-introduced to the market. The platform can also engage recycling companies to convert waste materials to useful products. The effectiveness of such a system can create a ripple effect across businesses. It encourages waste minimization, promotes resource conservation, reduces production cost by using reclaimed materials and expands the potential for new revenue streams from leasing and selling refurbished products. Moreover, implementation appears feasible. The concept leverages existing digital marketplace models, can start in small local circles and be gradually scaled to bigger markets. Initially, the model can start with pilot sectors like automobile and construction, and once successful, the model can be replicated to other industries, promoting a circular economy at a larger scale. From an environmental standpoint, the extended product lifecycles, decreased waste, and reduced resource mining will lessen the negative environmental impact. Financially, businesses stand to save on raw material costs, generate revenues from idle assets and waste, and reduce landfill costs. Overall, the concept of 'Resource Lifecycle Extension and Re-utilization Platform' significantly contributes to ushering in a more sustainable and economically viable circular economy.	
Problem: Despite emerging advancements in green technology and renewable energy, transportation and logistics remain heavily reliant on traditional, non-renewable energy sources. This is a pressing problem, particularly in a world where climate change is increasingly urgent. More specifically, the process of shipping and freight transportation continues to emit a significant amount of greenhouse gases. Today's non-circular economy largely rests upon the use of single-use containers and extended freight delivery routes, often crisscrossing	3M

locations unnecessarily due to suboptimal planning. Consequently, heavy costs for businesses, inefficient processes, and environmental damage proliferate. In essence, the problem relates to the emission-intensive, expensive and inefficient ways of transporting goods. <u>Solution:</u> I propose a combination of intelligent route planning, reusable packaging and better use of data as a circular economy solution. Firstly, advanced AI systems can optimize delivery routes in real-time, increasing efficiency while reducing fuel consumption and emissions. Key to achieving this is incorporating machine learning algorithms and AI to effectively use data related to current traffic, weather, and delivery locations. Such systems can lead to fewer miles driven and a reduction in unnecessary idling, therefore reducing costs and emissions. Secondly, reusable packaging solutions can be adopted, directly minimizing waste produced by the industry. These durable containers can be recycled and used multiple times with their life cycle significantly extended. Also, returnable packing materials frequently occupy less space than their disposable counterparts when empty, leading to expanded transport efficiency. Lastly, using data to forecast demand and match it with supply in the industry can reduce the number of empty kilometers traveled by vehicles. By applying predictive analytics, we can determine demand trends which help in optimizing cargo load and reducing unnecessary trips. This model is not only adaptable and scalable to varying business sizes and needs but also capable of economic pay-off due to reduced fuel consumption and packing costs. Plus, it bears the potential to have a significant positive environmental impact. Utterly congruous with the circular economy principles, this proposition ensures longer product utility, optimized resource usage, efficient supply, and lesser wastage.	
<u>Problem:</u> Electronic waste, or 'e-waste', is a growing concern. With new devices continuously entering the market, old electronics are too often discarded, often improperly. These devices can contain toxic materials hazardous to both environmental and human health. The existing recycling rate for e-waste is low, and valuable, finite resources present in these devices, such as rare metals, are not sufficiently recovered. <u>Solution:</u> The proposed 'Electronics Lifecycle Optimization Initiative (ELOI)' would be based on these principles: 1. Extended Producer Responsibility (EPR): Enforce stringent EPR regulations that require electronic manufacturers to manage the lifecycle of their products, including end-of-life disposal and recycling. 2. Design for Disassembly and Recycling: Encourage electronics manufacturers to design devices with their entire life cycle in mind. This means making it easy to disassemble devices for recycling and to use materials that can be efficiently	1S

recovered. 3. E-Waste Recycling Kiosks: Establish e-waste recycling kiosks in public spaces, stores, and community centers to facilitate convenient e-waste disposal. 4. Urban Mining: Promote urban mining, is the process of reclaiming compounds and elements from products, buildings, and waste, which includes e-waste recycling as a chief component, helping recover valuable and precious metals. The initiative's environmental benefits would come from reducing toxic e-waste, reducing the need for new raw material extraction, and reclaiming valuable resources. It creates financial value by generating a market for second-hand components and reclaimed materials. EPR regulations provide an incentive for manufacturers to become involved in the second-hand market, opening up new opportunities for revenue. With strong regulations and increasing awareness about e-waste, this initiative's feasibility is high. Its scalability extends to every locale with electronic consumers, effectively making it a global strategy. It satisfies all evaluation criteria, making it a strong contender for the circular economy challenge.	
Problem: The construction industry in Asia produces a vast quantity of surplus materials - such as bricks, cement, wood, and metal - that are discarded after projects, leading to significant waste. The automobiles and industrial equipment sector generate a significant amount of scrap metal and used parts. With the existing linear 'create-use-discard' model, these valuable resources are often wasted, causing environmental harm and financial loss. The challenge here is creating a sustainable, circular solution that optimally utilizes these resources and minimizes waste. Solution: I propose the 'Relove & Rebuild' initiative, a platform that connects construction companies, automakers, and industrial houses with smaller businesses or individuals who need these surplus materials or parts. 'Builders Bay' in the platform would list excess materials from construction sites. Smaller builders, artists, DIY enthusiasts, can purchase them at discounted prices, preventing these materials from going to waste. 'Revamp Garage' would list the scrap metal and used parts from automobiles and industrial equipment. Local repair shops, artisans, hobbyists can buy these parts to refurbish or to create new products. 'Green Points' would be given for each transaction, encouraging participation and reinforcing sustainable behavior. By reducing waste, the initiative significantly lowers the environmental impact. Financially, savings are made by both parties - the seller recovers some costs	2S

from surplus, while the buyers gain access to cheaper resources. The solution requires a user-friendly digital platform and engagement from the industries. Once it gains traction, the model can be scaled across different regions, increasing its environmental and financial impacts. 'Relove & Rebuild' presents a novel way to approach surplus materials, harnessing the power of the circular economy to transform waste into wealth, and fostering an ecosystem of sustainable practices in construction and manufacturing industries.	
Problem: In the Construction Materials industry, concrete production poses a significant environmental challenge. Concrete's primary ingredient, cement, is produced from limestone, which requires high-temperature kilns and emits vast amounts of CO_2. Combined with the large volumes of concrete used worldwide, the cement industry accounts for about 7% of global CO_2 emissions. Managing these emissions is imperative for combating climate change and fostering a circular economy within the construction sector. Solution: my proposition encapsulates an innovative approach termed "Green Concrete Revolution," composed of "Low-Carbon Concrete Production," "Concrete Recycling," and "Carbon Capture and Usage." The first strategy, "Low-Carbon Concrete Production," revolves around producing concrete with less dependence on Portland Cement, the main cause of high emissions in concrete production. This goal could be accomplished by replacing a part of cement with industrial by-products like fly ash or slag. Researchers are also exploring new types of cement that require lower kiln temperatures, thereby reducing energy use and emissions. New concrete compositions, like self-healing concrete or ultra-high-performance concrete, despite their higher initial cost, require less frequent replacement and repair, extending their lifecycle and reducing emissions in the long run. Secondly, "Concrete Recycling" becomes imperative. When demolishing concrete structures, recycling and reusing the concrete waste in new projects should be standardized. It can reduce the demand for virgin aggregate and cement, translating into lesser resource extraction and emissions. Lastly, "Carbon Capture and Usage" refers to the integration of carbon capture technology in cement plants. The captured CO_2 can be injected into fresh concrete, where it becomes locked once the concrete hardens, effectively making this material a carbon sink. The aggregate effect of these solutions produces a substantial environmental benefit, significantly lowering the industry's carbon emissions. The financial impact is two-fold; although low-carbon concrete and carbon capture technology may be more expensive initially, they offer the potential for operational cost savings due to lesser energy use, and the long-term lifecycle cost savings provide an economic edge. Moreover, with climate policies tightening worldwide, embracing these green initiatives can help avoid future regulatory restrictions or penalties. In terms of feasibility, the technology for these solutions is available and is being progressively improved. The scalability of this approach	3S

is global due to the universal use of concrete in construction. Adopting the "Green Concrete Revolution" can fundamentally shift the Construction Materials industry towards a sustainable and circular economic model.	
Problem: Roughly 1/3 of all food produced for human consumption goes to waste. 87% is caused by perishable items. Waste caused by perishable products in supermarkets and online grocery stores has a negative effect on the environment in addition to the financial loss. Solution: To help reduce the waste caused by perishable products in supermarkets and online grocery stores, a food waste solution has been developed by Wasteless. It is a proprietary dynamic pricing algorithm for products based on their expiration date. This allows retailers to deploy AI to mark off prices on perishable products. The system is managing the inventory of fresh products in relation to expiring dates and the consumers' attitudes towards them. Wasteless's pricing engine employs a branch of machine learning called ¬´Reinforcement Learning¬ª. This allows the pricing engine to quickly learn how consumers respond to dynamic pricing so it can then find the optimal discounting policy. The retailers will achieve a remarkable increase in revenue by optimizing pricing and incentivizing their shoppers to buy products that would otherwise go to waste. Suppliers will leverage prices to consumers to better match supply with demand, reduce overstock situations, and ensure constant product availability at the point of sale.	HC

Table A3. Examples of Human and AI Problems and Solutions for Top Novelty by Configuration

Example Problem and Solution	Configuration
Problem: A significant portion of food waste is generated by restaurants, supermarkets, and households, which not only squanders resources but also adds to the growing problem of landfills. In particular, the disposal of food waste in landfills results in the generation of methane, a greenhouse gas twenty times more potent than carbon dioxide. Hence, finding an efficient and economic method to manage food waste reduces the need for new landfills and helps mitigate climate change. Solution: To address this issue, I propose the creation of local bio-energy centers. The idea is to partner with restaurants, markets, and households to collect their organic waste. This waste will then be converted into fertilizer for community gardens and parks, as well as biogas, which can be used to generate electricity. First, there would be waste collection bins at strategic locations so businesses and households can conveniently dispose of their organic waste. The waste would be regularly collected and brought to the bio-energy centers for processing. The organic waste, once in the bio-energy center, will be treated in a process called anaerobic digestion. Anaerobic digestion is basically a series of biological processes in which microorganisms break down biodegradable material in the absence of oxygen. This process results in the production of biogas, which consists mainly of methane. The methane generated can be used as a renewable source of energy, to generate heat and electricity, while nutrient-rich digestate (a byproduct of the process) can be used as a low-cost, efficient fertilizer to promote local agriculture and horticulture.	1M

This solution provides multiple benefits: reduction in waste that would have gone to landfills; production of clean, renewable energy; providing organic, chemical-free fertilizers for agricultural purposes; and contributing to the local economy. The feasibility and scalability of this model have been proven in several European countries. The financial impact is expressed in both cost savings (as waste management costs are reduced) and income sourcing (sale of energy, fertilizer). The environmental benefits are also clear, as this process drastically reduces the release of harmful greenhouse gases that contribute to climate change.	
Problem: Africa is fraught with challenges pertaining to food and beverage waste, inefficient packaging methods, and lack of waste management techniques. Every year, million tons of plastic and other packaging materials end up in landfills, deteriorating the environment and posing a threat to human health. Simultaneously, a significant percentage of food and beverages are wasted due to inadequate storage and distribution systems. This overall inefficiency leads to economic losses and emissions of greenhouse gases contributing to climate change. Solution: We propose a solution titled "Zero Waste and Nutrient Circularity in Food, Beverages & Packaging through Bio-Conversion Technology." The idea centers around the use of black soldier flies (Hermetia illucens), which can consume different types of organic waste, including food and beverage waste. This process will convert waste into larvae, which in turn can be used as a protein source for animal feed. On the other hand, the residual waste can be used as a nutrient-rich biofertilizer to replenish agricultural lands. In terms of packaging, we envision transitioning to zero-waste packaging solutions. One such approach is the introduction of edible, biodegradable packaging material made from natural substances such as seaweed. In conjunction, a widespread ,"Return, Reward, and Recycle." initiative would encourage consumers to return their used packaging for responsible recycling or composting, incentivized through discounts or other benefits. The solution's novelty lies in integrating biological conversion and eco-friendly packaging into a unified waste management system. It has a substantial environmental impact, reducing plastic waste and emissions, while the financial impact is realized through cost savings in waste management, creation of saleable products, and sustainable brand image enhancement. It is feasible with public participation and scalable through public-private partnerships and legislative support.	2M
Problem: The beverages industry, in which I specialize, makes a substantial contribution to global waste – particularly plastic waste – since packaging is predominantly single-use. Despite recycling efforts, a substantial percentage of these bottles, cans, and containers end up in landfills or the natural environment. The problem is multifaceted and contributes to several global challenges: resource depletion due to virgin plastic production, waste management difficulties, and pollution, not to mention the carbon emissions associated with production, transportation, and recycling. Additionally, this linear production model puts financial pressure on businesses as fluctuations in raw material prices can heavily impact costs. The industry is in need of a transformative, circular solution.	3M

<u>Solution:</u> The solution lies in a "Beverage-as-a-Service" (BaaS) model that leverages IoT technology for the reuse of beverage containers and drives customer engagement in the circular economy. Instead of the traditional, linear 'produce-consume-dispose' model, containers (bottles, cans) would be designed for reusability, embedded with a chip that tracks their lifecycle. Upon purchase, customers pay a small deposit – they are essentially leasing the container. When the container is empty, customers return it to a collection point (grocery stores, vending machines, collection stations). The IoT chip then communicates the container's return, and the deposit is refunded to the customer digitally. Cleaned and refilled, the containers then loop back into the market – no need for recycling or waste treatment. Simultaneously, the digital trail created by the IoT chips will enable data collection on consumption habits. For businesses, this means data-driven decisions and personalized marketing. Even more, cost savings as fluctuations in raw material prices will have less impact, reducing financial risk. Plus, the deposit system will encourage a high return rate, ensuring a steady supply of reusable containers. This is a scalable solution that can start with select locations and products before expanding. It creates a closed-loop system, thus minimizing environmental impact by eliminating waste and reducing production need. While certain infrastructure investments are necessary, such as cleaning and refill systems and IoT technology, the long-term benefits in cost savings and sustainability make this a feasible solution for the beverage industry.	
<u>Problem:</u> Single-use meal containers and packaging in food delivery and take-away services generate massive amounts of waste. Most of these containers are made from non-recyclable or difficult-to-recycle materials due to food contamination issues. This ultimately exacerbates problems related to waste disposal and resource depletion. While some delivery services and restaurants have started to provide more sustainable packaging options, the scale of the issue requires more comprehensive and effective solutions. <u>Solution:</u> A 'Reusable Food Container Service (RFCS)' could offer a sustainable and innovative approach to this. 1. Durable Containers: Collaborate with restaurants and food delivery services to introduce durable, reusable, and standardized meal containers. These containers, made of safe and long-lasting materials, would come in different standard sizes that can accommodate various types of food. 2. Deposit Model: Customers would pay a small refundable deposit for the reusable container at the point of ordering their meal. Alternatively, subscribers to the RFCS could obtain a set of containers to use for multiple food orders. 3. Collection and Cleaning: After use, dishes could be returned in designated collection points or directly to partnered restaurants/delivery services during the next order. The containers would then be professionally cleaned in a central location, ready for reuse in line with hygiene standards. 4. Awareness: Promote the importance of sustainability in food delivery habits and	1S

personalize it - show users how many single-use containers they have saved from waste by choosing RFCS. The RFCS offers a unique approach that extends the idea of reusable shopping bags and water bottles to the widespread practice of food take-away and delivery. It contributes to reducing packaging waste, which is a significant environmental problem in urban areas worldwide. It also provides a model for delivering economic benefits through saving on disposable packaging costs and creating revenue from deposits and subscription fees. The main challenges for feasibility include establishing efficient cleaning and distribution logistics, ensuring hygiene standards, and getting buy-in from businesses and customers. However, the popularity of take-away services and the increasing awareness around packaging waste suggests that the scalability of this concept could be wide-reaching. The RFCS has the potential to revolutionize the way we package and consume take-out food, turning a linear process into a cyclical one that safeguards our planet and resources.	
Problem: In a number of African cities and towns, a significant amount of food and beverages is wasted as a result of mismanagement, lack of proper storage facilities, and logistical challenges in the supply chain. Simultaneously, the vast usage of non-reusable and non-recyclable packaging materials contributes to growing littering issues and landfills. Many households and small businesses lack the knowledge or resources to manage these issues effectively, thereby leading to considerable environmental and socio-economic challenges, such as pollution, waste of valuable resources, added stress on landfills and waste management infrastructure, and lost economic opportunities. Solution: To address these challenges, I propose the development of an innovative 'Smart Food Waste & Packaging Management System', focusing on the principles of the circular economy. This system would essentially comprise an integrated network of IoT-enabled 'Smart Bins' for households and businesses, a centrally coordinated collection and redistribution system, a waste-to-energy micro-scale facility, and an educational mobile app platform. The Smart Bins connected to Wi-Fi, would segregate food waste and packaging materials and alert the central system when full, enabling planned and timely pickups. The collected food waste would be redirected through two paths based on their conditions - safely edible items could be donated to local food banks, while spoiled food items would be processed at a micro-scale waste-to-energy facility to generate bioenergy. The collected packaging waste would be sorted and recycled. Simultaneously, an interactive mobile application would educate users about proper waste segregation, the concept of 'food rescue', recycling options for packaging materials, and the benefits of reducing food and packaging waste. Consumers could also be incentivized through a reward system within the app to promote proactive participation. This solution would massively reduce food waste and packaging littering, curtail the load on landfills, generate bioenergy, and foster sustainable behavior among the consumer base. Financially, it has potential to create revenue from the sale of bioenergy and recycled packaging materials, while the reduction in waste collection and landfilling costs for municipalities also makes it an attractive prospect. This system might require initial investment, partnerships with tech companies for app development and IoT setup, and collaboration with municipalities for implementation. However,	2S

once proven successful, it could be scaled up across different cities throughout Africa. The proposed system stands out in its novelty due to its integrated approach towards managing food waste and packaging, its use of smart technology for effective waste management, and its focus on consumer education - all working synergistically to revolutionize waste management in Africa.	
Problem: The 'beyond Big 4' metals sector faces challenges related to the tailings issue. Mining for metals like cobalt, nickel, or rare earths often result in large quantities of leftover rock and processed material, known as tailings. These tailings are stored in dams but pose severe environmental risks, including groundwater contamination and dam failures, leading to catastrophic spills. It's critical to manage these tailings both for environmental protection and for repurposing these waste materials into valuable resources. Solution: I suggest the implementation of a comprehensive "Tailings Transformation Strategy" based on "Innovative Tailings Reprocessing" and "Stable Storage Measures." "Innovative Tailings Reprocessing" involves developing technologies to efficiently extract residual valuable metals from tailings. Advances in extraction techniques, including bioleaching or hydrometallurgical methods, can make it feasible to recover these metals that would otherwise go to waste. Additionally, tailings can also be repurposed for other uses such as construction material, thereby minimizing waste. "Stable Storage Measures" focus on investing in the infrastructure and techniques to securely store tailings. This might involve more secure dam designs, real-time monitoring systems, or even elimination of tailings dams by shifting towards dry stacking methods. The "Tailings Transformation Strategy" would dramatically mitigate the environmental risks associated with tailings while also creating new sources of metallic resources, reducing the need for further raw material extraction. On the financial side, the recovered metals could add to revenue streams, potentially offsetting the costs of improved storage measures. As for feasibility, it relies on advancing technologies, some of which are already in use to a certain extent. Considering the global mining industry's scale, the potential for scalability is substantial. In these ways, the "Tailings Transformation Strategy" can enable the metals sector's significant strides towards a circular economy.	3S
Problem: Since the construction industry is the largest user of coal (used to fire the bricks) in the country and a significant cause of air pollution, it also leads to the loss of fertile topsoil. Therefore, I came up with an ingenious solution to this pertinent problem by creating bricks, made from foundry dust and waste plastic! Solution: To overcome this ecological problem, we came up with the bricks, made from foundry dust and waste plastic. There is no use of water in the making of these bricks. Also, more water is saved during the construction process, as the walls built with these bricks do not need to be cured with water. Also, the best part of this idea is that while 70 percent of the total sand can be reused, the other 30 percent, in the form of foundry dust, is too fine to be used again. Also, with enhanced technology we evolve this idea in making interlocking bricks, which essentially work like Lego blocks.	HC

A.1 Detailed Explanation of Prompt Building Mechanism

We elaborate on the specific mechanism used to construct the prompts for the AI-generated solutions in the study. Our implementation uses the Python programming language on GoogleColab and leverages OpenAI's GPT-4 model. We interact with the GPT-4 model using OpenAI's ChatCompletion API, designed for conversational tasks and allows for multi-turn exchanges with the model by including a series of structured messages as inputs.

Each message included in the API request is categorized by role and content. The "role" attribute is assigned as either "system" or "user." The "system" role provides high-level instructions or context for the conversation, while the "user" role prompts the model to generate specific outputs based on the given task.

The specific Python function utilized for the API request was openai.ChatCompletion.create(), which accepts several parameters. The "model" parameter specifies the AI model being used, which in this case was set to "gpt-4." The "messages" parameter is a list of structured messages to be delivered to the model. Each message in the list is a dictionary containing two keys: "role" and "content." We set the "temperature" parameter to the default value of 1, max_tokens to the maximum capacity of 8191, top_p to the default of 1, frequency_penalty to the default of 0, and presence-penalty to the default of 0.

In our study, the content of the "system" message was set to the "context," a string that provides the general context of the problem to be solved exactly as it was shared with humans, potentially augmented with additional information such as solver characteristics or persona details based on the prompt level. The "user" message's content was set to a specific "content" string, containing the template for the answer.

The function call in our code is thus:

```
response = openai.ChatCompletion.create(
                model="gpt-4",
                messages=[
                {"role": "system", "content": default_context},
                {"role": "user", "content": content}
                ],
     temperature=1,
     max_tokens=8191,
     top_p=1,
     frequency_penalty=0,
     presence_penalty=0)
```

Note that as of May 2023, the default temperature was 0.7, which is equivalent to 1.0 as of March 2024 after OpenAI recently rescaled the parameter.

Upon execution, the function returns a "response," which contains the AI-generated solution. This solution is then used for further evaluation and comparison in the context of our study. This iterative process of prompt creation and HAI response generation was carried out 750 times to match the 125 HC solutions for each level and configuration.
The code used to generate the solutions is publicly available at https://github.com/leobix/creative.

A.2 Modifying Prompt Qualifying Adjectives to Match Distribution of Human-Generated Answers

Given the diversity and variation inherent in HC-generated responses, we aimed to replicate this natural dispersion by slightly adjusting each ChatGPT prompt. Specifically, we incorporated a range of qualifying adjectives indicative of the expected answer length in the prompts.

We defined a set of adjectives: "highly detailed and elaborate", "succinct", "brief", "concise", "short", "comprehensive", "long", "5-paragraph", "3-paragraph", "medium-length", "very precise and elaborate", "20-sentence". The Python code iteratively selected adjectives from these sets in a predefined random sequence and injected them into the prompts.

By employing a variety of qualifying adjectives for the expected length of the problem and solution, we generated a range of AI responses that mirrored the distribution of human-generated solutions more closely. This enhancement further refined our experiment, offering a richer comparison of the capabilities between human-generated and AI-generated solutions.

A.3 Prompts Used

We provide the prompts we used for each level below:

[Level 1M-2M-3M]

System prompt (context):
We are excited to announce an opportunity for freelancers to collaborate with researchers at the Digital, Data, and Design Institute at Harvard to source the most innovative and cutting-edge circular economy solutions for the business world.

Circular Economy is a simple idea.

Basically it involves sharing, leasing, reusing, repairing, refurbishing and recycling existing materials and products as long as possible.

We would like you to submit your circular economy idea, which can be a unique new idea or an existent idea that is used in the industry.

Here is an example: Car Sharing in order to reduce the carbon footprint associated with driving.

Submit your real-life use cases on how companies can implement the circular economy in their businesses. New ideas are also welcome, even if they are 'moonshots'. Your suggestions will help Harvard researchers understand the impact of the circular economy on business. Let's get creative and revolutionize the world through the circular economy!

Your goal is to win the top monetary prizes. Judges will use the following evaluation criteria:
* Novelty (How different is it from existing solutions?)
* Environmental Impact (How much does it benefit the planet?)
* Financial Impact (What financial value can it create for businesses?)
* Feasibility and Scalability of Implementation (How likely is it to succeed and how scalable is it?)

[Level 2M adds the following to the above prompt]
To answer the question, you will take the perspective of the following persona:
You are a [Job Title] located in [Continent]. You propose a solution that applies to [Industry of Solution]. The maturity of your solution is [Maturity].

We provide below two examples of such persona:

- You are a Executives, Managers, and Entrepreneurs, located in Africa. You propose a solution that applies to Food, Beverages, Packaging and Waste Management. The maturity of your solution is Proof of Concept.
- You are a Technical and Creative Professionals, located in Europe. You propose a solution that applies to Transportation and Logistics. The maturity of your solution is Ideation.

[Level 3M adds the following]
To answer the question, you will take the perspective of the following persona:
[Expert Name] who has expertise in [Expert Field].

User Prompt:

Answer the following two questions to propose a circular economy idea that could win the challenge according to the evaluation criteria.

Problem: Tell us about the problem your solution is meant to solve.

Solution: Describe the solution in your own words.

Use the following template to answer:

[Level 1M]
Problem: <Write a high quality, *ADJECTIVE1* answer.>
Solution: <Write a high quality, *ADJECTIVE2* solution.>

[Level 2M-3M]
Problem: <Write a high quality, *ADJECTIVE1* answer, **corresponding to the personality, inspiration, and knowledge of your persona**.>
Solution: <Write a high quality, *ADJECTIVE2* solution, **corresponding to the personality, inspiration, and knowledge of your persona**.>

[Level 1S-2S-3S]

System prompt (context):

We are excited to announce an opportunity for freelancers to collaborate with researchers at the Digital, Data, and Design Institute at Harvard to source the most innovative and cutting-edge circular economy solutions for the business world.

Circular Economy is a simple idea.

Basically it involves sharing, leasing, reusing, repairing, refurbishing and recycling existing materials and products as long as possible.

We would like you to submit your circular economy idea, which can be a unique new idea or an existent idea that is used in the industry.

Here is an example: Car Sharing in order to reduce the carbon footprint associated with driving.

Submit your real-life use cases on how companies can implement the circular economy in their businesses. New ideas are also welcome, even if they are 'moonshots'. Your suggestions will help Harvard researchers understand the impact of the circular economy on business. Let's get creative and revolutionize the world through the circular economy!

Your goal is to win the top monetary prizes by satisfying to the maximum the following evaluation criteria.
* Novelty (How different is it from existing solutions?)
* Environmental Impact (How much does it benefit the planet?)
* Financial Impact (What financial value can it create for businesses?)
* Feasibility and Scalability of Implementation (How likely is it to succeed and how scalable is it?)

Answer the following two questions to propose a circular economy idea that could win the challenge according to the evaluation criteria.

Problem: Tell us about the problem your solution is meant to solve.

Solution: Describe the solution in your own words.

Use the following template to answer:

[Level 1S]
Problem: <Write a high quality answer.>
Solution: <Write a high quality solution.>

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.

[Level 2S-3S]
Problem: <Write a high quality answer, corresponding to the personality, inspiration, and knowledge of your persona.>
Solution: <Write a high quality solution, corresponding to the personality, inspiration, and knowledge of your persona.>

We will give you a series of different personas, and make sure each new answer from each persona tackles a different problem than the previous ones and proposes a different solution. Make sure each answer corresponds to the perspective, characteristics, and knowledge of your persona.

User prompt:

Give a new high quality, *ADJECTIVE1* Problem and high quality, *ADJECTIVE2* Solution. Make sure to tackle a different problem than the previous ones and propose a different solution. Make also sure your answers satisfy the evaluation criteria (novelty, environmental impact, financial impact, feasibility and scalability).

[Level 2S adds the following]
The persona you embody for this answer:
You are a [Job Title] located in [Continent]. You propose a solution that applies to [Industry of Solution]. The maturity of your solution is [Maturity].

[Level 3S adds the following]
The persona you embody for this answer:

[Expert Name] who has expertise in [Expert Field].

A.4 Expert Personas from the Randomly Sampled Level 3 Answers Evaluated

Level 3M:
(Mette Hay, Home Furnishings), (Hubertus Muehlhaeuser, Electrical Equipment), (Ramon Laguarta, Containers & Packaging), (Kenichiro Yoshida, Consumer Electronics), (David Steiner, Waste Management), (Michael Green, Building Products), (William L. McComb, Household Appliances), (Howard Schultz, Food), (David Abney, Transportation & Logistics), (Lisa P. Jackson, Software & IT Services), (Dave Lennard, Building Products), (Simon Segars, Electrical Equipment), (Stephen Kieran, Buildings), (Akio Toyoda, Automobiles & Tires), (Jensen Huang, Technology/ Hardware Products), (Tom Linebarger, Construction Machinery), (Emmanuel Faber, Food), (Catherine Howarth, Financials), (James Timberlake, Buildings), (Ren Zhengfei, Consumer Electronics), (Alex Gorsky, Health Care Products), (Stella McCartney, Apparel & Textiles), (Paul Polman, Forest Products), (Elon Musk, Automobiles & Tires), (Bill Browning, Construction Materials), (Andrew Martin, Home Furnishings), (Tim Cook, Technology/ Hardware Products), (Alex Keith, Cosmetics), (David Greenberg, Cosmetics), (Jean-Paul Agon, Cosmetics)

Level 3S:
(Satya Nadella, Software & IT Services), (Kenichiro Yoshida, Consumer Electronics), (Andrew Liveris, Construction Materials), (Akio Toyoda, Automobiles & Tires), (Mark Eames, Metals: Beyond Big 4), (Michael Dell, Technology/ Hardware Products), (Larry Fink, Financials), (Lisa Su, Technology/ Hardware Products), (Catherine Howarth, Financials), (Tim Cook, Technology/ Hardware Products), (Pat Gelsinger, Technology/ Hardware Products), (Richard Adkerson, Metals: Beyond Big 4), (Ramon Laguarta, Containers & Packaging), (Paul Polman, Forest Products), (Tom Linebarger, Construction Machinery), (Thomas Rau, Home Furnishings), (Uday Yadav, Electrical Equipment), (Leif Johansson, Health Care Products), (Berry Wiersum, Forest Products), (Jan Jenisch, Construction Materials), (Marc Benioff, Software & IT Services), (Stephen Kieran, Buildings), (John Hayes, Containers & Packaging), (Rick Fedrizzi, Buildings), (Mark Bitzer, Household Appliances), (Ivan Glasenberg, Metals: Beyond Big 4), (John Elkington, Construction Materials), (Lance Fritz, Transportation & Logistics), (Denise Morrison, Food), (Bill Browning, Construction Materials)

A.5 Initial Formulation of the Challenge for Humans
(after removing the administrative details to participate in the challenge)

We are excited to announce an opportunity for freelancers to collaborate with researchers at the Digital, Data, and Design Institute (D^3) (https://d3.harvard.edu/) at Harvard to source the most innovative and cutting-edge circular economy solutions for the business world.

Circular Economy is a simple idea.

Basically it involves sharing, leasing, reusing, repairing, refurbishing and recycling existing materials and products as long as possible.

We would like you to submit your circular economy idea, which can be a unique new idea or an existent idea that is used in the industry.

Here is an example: Car Sharing in order to reduce the carbon footprint associated with driving.

Here is more information on circular economy:
https://ellenmacarthurfoundation.org/topics/circular-economy-introduction/overview

Submit your real-life use cases on how companies can implement the circular economy in their businesses. New ideas are also welcome, even if they are 'moonshots'. Your suggestions will help Harvard researchers understand the impact of the circular economy on business. Let's get creative and revolutionize the world through the circular economy!

Our team will be evaluating your entries using the following criteria:
* Novelty (How different is it from existing solutions?)
* Environmental Impact (How much does it benefit the planet?)
* Financial Impact (What financial value can it create for businesses?)
* Feasibility and Scalability of Implementation (How likely is it to succeed and how scalable is it?)

The best overall solution will receive a $1,000 prize and be presented with an official trophy from the D^3 Institute at Harvard.

Top 500 best solutions will receive a cash prize of $10 each. The same freelancer could potentially win multiple prizes if they had submitted more than one winning entries! Apart from the cash prize, the winning entries will receive a letter from a Harvard faculty member and an official certificate from Harvard D^3 Lab.

Some creators of the best solutions will also receive an invitation to participate in an exclusive two-day virtual Circular Economy Catalyst Event at Harvard Business School in April 2023 to learn how leading companies, startups, and investors are engaging in the circular economy.

Appendix B: Technical Details of Inference Mechanism of LLMs

The training process of GPT-4 has never been made publicly available. We refer to the GPT-4 technical report for extended reading: https://cdn.openai.com/papers/gpt4.pdf. The inference stage in LLMs is the phase where the model generates text based on the input provided. This stage follows the pre-training and fine-tuning phases and relies on the Transformer architecture that underpins the model.

The inference process encompasses the following steps:

1. Input Tokenization: The input text is tokenized into subwords or tokens using a tokenizer trained on the same corpus as the language model.
2. Token Embedding: Tokens are converted into numerical vectors, known as embeddings, which capture semantic and syntactic information.
3. Positional Encoding: To provide information about the sequential order of the tokens, positional encodings are added to the embeddings since the Transformer architecture does not inherently understand the sequential nature of the text data.
4. Transformer Processing:
    a. The embeddings pass through multiple layers of the Transformer, each consisting of self-attention mechanisms (see B.1 for details) and feed-forward neural networks.
    b. Self-attention (see B.1 for details) allows the model to weigh the importance of different parts of the input sequence when generating each token in the output sequence.
    c. The feed-forward networks apply further transformations to the attention-weighted embeddings.
5. Output Token Generation: After tokenizing and encoding a prompt, this leaves a block of data representing our input as the machine understands it, including meanings, positions, and relationships between words. The model uses the final layer's output to estimate the probability distribution over the next token. GPT-4 employs a sampling strategy that calculates the probability

distribution of the next token based on the context provided by the input sequence and the internal representations learned during training.

B.1 Self-Attention Mechanism

The benefits of self-attention for language processing increase as the model scales. Simply put, it allows LLMs to take context from beyond sentence boundaries, giving the model a greater understanding of how and when a word is used. The idea is formalized with a *self-attention function*, which takes as input a sequence of initial token embeddings and outputs a sequence of new token embeddings that allow the initial embeddings to interact. Let $(p^0{}_{d,1}, p^0{}_{d,2}, \dots, p^0{}_{d,N})$ be the initial embeddings that make up a document. The new embedding at each position n is given by,

$$p^1{}_{d,n} = \sum_{n'=1}^{N_d} w_{(d,n),n'}\, p^0{}_{d,n'} \text{ where } \sum_{n'=1}^{N_d} w_{(d,n),n'} = 1.$$

.

That is, each embedding in the transformed sequence is itself a weighted average of the embeddings in the initial sequence. The non-negative attention weights $w_{(d,n),n'}$, which are estimated during model training, determine which pairs of (potentially distant) tokens interact to form each context-sensitive word embedding in the final document representation. We thank Ash & Hansen (2023) for their clearly explained embedding sequences with attention in their paper: https://www.annualreviews.org/content/journals/10.1146/annurev-economics-082222-074352.

B.2 Decoding Strategies and Sampling Methods

Several sampling methods can be used to select the next word from the probability distribution. It is unclear which method GPT-4 uses precisely:

- Greedy Sampling: Chooses the token with the highest probability, leading to deterministic outputs.
- Random Sampling: Selects a token randomly based on the probability distribution, allowing for varied outputs.
- Top-k Sampling: Restricts the sampling pool to the top-k most likely tokens, balancing variety and coherence.
- Top-p (Nucleus) Sampling: Chooses from a subset of tokens that cumulatively make up to a certain probability p, focusing on high-probability tokens while maintaining diversity.

Once the next token is sampled, the model continues the process autoregressively, generating one token at a time and feeding the updated sequence back into the model until a termination condition is met, such as the end-of-sequence token or a specified maximum length. It is worth noting that previous studies have shown that custom decoding methods could significantly improve a language model's output for a specific task. Despite the potential shown in tailored decoding techniques, there seems to be a declining trend in their popularity. This may be attributed to two primary factors: 1) the increasing propensity for close-sourced models, which are less amenable to user-defined decoding adaptations, and 2) the improvements in baseline performance of pre-trained LLMs, which diminish the perceived need for such customizations.

Nevertheless, decoding adjustments could serve as an avenue for enhancing the generation of less common, or "long-tail," solutions. Two ways to generate more nuanced and varied outputs from LLMs is using simple modifications, such as adjustment of temperature settings or the manipulation of top-k/top-p parameters. These exploratory steps could pave the way for potentially unlocking new capabilities within these systems.

The "temperature" hyperparameter influences the randomness or "creativity" of the model's outputs. When generating text, the model calculates a probability distribution over possible following words. The temperature modifies the sampling from this distribution through a softmax calculation. At a higher temperature value, the distribution is flatter, rendering the output more random as it considers a broader range of word possibilities. Conversely, a lower temperature results in sharper distribution, leading to more deterministic outputs, wherein the model is more inclined to pick the most probable next word. For this study, we set the temperature at GPT-4 API's default value.

Top-k sampling is a method where the model restricts its choice of the next word to the k most likely options. The value of k determines the breadth of the model's consideration set; a smaller k leads to a higher probability that the model will select a more common or expected word, resulting in text that is typically more coherent but less varied. A larger k value gives the model leeway to pick less probable words, thereby increasing novelty and variation in the output but potentially at the cost of coherence and predictability.

Top-p, or nucleus sampling, takes a different approach by choosing from a dynamic set of options. Rather than selecting from the top k possibilities, it selects from the smallest set of words whose cumulative probability exceeds the threshold p. This means the model considers a range of words just enough to sum up to the probability p, allowing for dynamic and context-dependent variation in the number of words considered. With a lower p value, the model's outputs are more focused and less random, similar to having a lower temperature or smaller k. As p increases, the model can sample from a wider array of words, thus injecting more randomness and diversity into the generated text, akin to increasing the temperature or k value.

In practice, adjusting top-k and top-p can help balance between the generation of common, highly probable text and more diverse or surprising outputs. By tuning these parameters, one can calibrate the LLM to produce outputs that range from safe and predictable to novel and wide-ranging, thus enhancing the generation of less common, "long-tail" solutions.

B.3 Output Generation

The output generation is the culmination of the inference stage, where the sequence of predicted tokens is converted back into human-readable text. The model's ability to generate fluent and contextually appropriate text is a direct result of the complex interaction between its learned parameters and the inference-time sampling strategies.

Appendix C: Regression Analysis of Human-AI (HAI) and Human Crowd (HC) Solutions

Table C1. Nested Mixed Effects Logistic Models of Evaluator Ratings of Top Novelty Rating on Solution Source (Human Crowd or Human-AI)

	Dependent Variable: Top Novelty Rating (Logistic)					
	(1)	(2)	(3)	(4)	(5)	(6)
HAI Solution	-0.624***	-0.624***	-0.644***			
	(0.107)	(0.107)	(0.108)			
HAI Multiple Instance				-0.778***	-0.779***	-0.785***
				(0.125)	(0.125)	(0.125)
HAI Single Instance				-0.484***	-0.484***	-0.510***
				(0.120)	(0.120)	(0.121)
Work experience		-0.035+	-0.035		-0.035+	-0.035
		(0.021)	(0.021)		(0.021)	(0.021)
Level of interest		0.174	0.180		0.175	0.180
		(0.116)	(0.115)		(0.116)	(0.115)
Knowledge test score		-0.144+	-0.142+		-0.144+	-0.142+
		(0.082)	(0.084)		(0.082)	(0.084)
Intercept	-1.603***	-1.772***	-1.897**	-1.605***	-1.775***	-1.873**

	(1)	(2)	(3)	(4)	(5)	(6)
	(0.133)	(0.518)	(0.613)	(0.133)	(0.519)	(0.614)
N	3900	3900	3900	3900	3900	3900
# blocks	18	18	18	18	18	18
# evaluators	300	300	300	300	300	300
Controls	N	N	Y	N	N	Y
Log-Likelihood	-1483.24 df = 4	-1479.09 df =7	-1474.08 df = 14	-1479.86 df = 5	-1475.71 df = 8	-1471.22 df = 15
R2 Marg.	0.015	0.031	0.045	0.018	0.035	0.048
R2 Cond.	0.370	0.372	0.372	0.374	0.375	0.374

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

Notes. This table presents mixed effects logistic model results from evaluator ratings of solution top novelty dummy, with 300 evaluators nested in eighteen solution blocks. Models 3 and 6 include the following covariates: Gender, Highest Level of Education, Major, Employment Status, Cohort Session, and Solution Word Count. Standard errors are in parentheses. The number of points per axis for evaluating the adaptive Gauss-Hermite approximation to the log-likelihood is set to zero for convergence.

Table C2. Nested Mixed Effects Logistic Models of Evaluator Ratings of Top Value Rating on Solution Source (Human Crowd or Human-AI)

	Dependent Variable: Top Value Rating (Logistic)					
	(1)	(2)	(3)	(4)	(5)	(6)
HAI Solution	0.128	0.129	0.091			
	(0.101)	(0.101)	(0.102)			
HAI Multiple Instance				0.140	0.140	0.119
				(0.111)	(0.111)	(0.112)
HAI Single Instance				0.117	0.117	0.061
				(0.111)	(0.111)	(0.113)
Work experience		-0.023	-0.024		-0.023	-0.024
		(0.022)	(0.022)		(0.022)	(0.022)
Level of interest		0.480***	0.484***		0.480***	0.484***
		(0.133)	(0.132)		(0.133)	(0.132)
Knowledge test score		-0.182+	-0.179+		-0.182+	-0.179+
		(0.095)	(0.096)		(0.095)	(0.096)
Intercept	-1.368***	-2.700***	-2.710***	-1.368***	-2.700***	-2.718***
	(0.136)	(0.594)	(0.702)	(0.136)	(0.594)	(0.702)
N	3900	3900	3900	3900	3900	3900
# blocks	18	18	18	18	18	18
# evaluators	300	300	300	300	300	300
Controls	N	N	Y	N	N	Y
Log-Likelihood	-1924.69 df = 4	-1914.09 df = 7	-1904.31 df = 14	-1924.67 df = 5	-1914.06 df = 8	-1904.12 df = 15
R2 Marg.	0.001	0.042	0.066	0.001	0.042	0.067
R2 Cond.	0.476	0.481	0.483	0.485	0.482	0.484

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

Notes. This table presents mixed effects logistic model results from evaluator ratings of solution top value dummy, with 300 evaluators nested in eighteen solution blocks. Models 3 and 6 include the following covariates: Gender, Highest Level of Education, Major, Employment Status, Cohort Session, and Solution Word Count. Standard errors are in parentheses. The number of points per axis for evaluating the adaptive Gauss-Hermite approximation to the log-likelihood is set to zero for convergence.

Table C3. Nested Mixed Effects Logistic Models of Evaluator Ratings of Top Creativity Rating on Solution Source ((Human Crowd or Human-AI)

	Dependent Variable: Top Creativity Rating (Logistic)					
	(1)	(2)	(3)	(4)	(5)	(6)
HAI Solution	−0.136	−0.136	−0.135			

	(1)	(2)	(3)	(4)	(5)	(6)
	(0.216)	(0.216)	(0.217)			
HAI Multiple Instance				−0.102	−0.102	−0.102
				(0.240)	(0.240)	(0.240)
HAI Single Instance				−0.172	−0.172	−0.171
				(0.243)	(0.243)	(0.246)
Work experience		−0.003	−0.004		−0.003	−0.003
		(0.030)	(0.030)		(0.030)	(0.030)
Level of interest		0.204	0.214		0.204	0.215
		(0.182)	(0.179)		(0.182)	(0.179)
Knowledge test score		−0.334*	−0.334*		−0.334*	−0.334*
		(0.135)	(0.135)		(0.135)	(0.135)
Intercept	−3.820***	−3.802***	−3.765***	−3.820***	−3.802***	−3.773***
	(0.223)	(0.811)	(0.961)	(0.223)	(0.811)	(0.961)
N	3900	3900	3900	3900	3900	3900
# blocks	18	18	18	18	18	18
# evaluators	300	300	300	300	300	300
Controls	N	N	Y	N	N	Y
Log-Likelihood	-517.12 df = 4	-511.75 df = 7	-505.03 df = 14	-517.07 df = 5	-511.69 df = 8	-504.98 df = 15
R2 Marg.	0.001	0.042	0.066	0.001	0.042	0.067
R2 Cond.	0.476	0.481	0.483	0.485	0.482	0.484

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

Notes. This table presents mixed effects logistic model results from evaluator ratings of solution top creativity dummy, with 300 evaluators nested in eighteen solution blocks. Models 3 and 6 include the following covariates: Gender, Highest Level of Education, Major, Employment Status, Cohort Session, and Solution Word Count. Standard errors are in parentheses. The number of points per axis for evaluating the adaptive Gauss-Hermite approximation to the log-likelihood is set to zero for convergence.

Table C4. Ordinary Least Squares Models of Top Decile Novelty Ratings on Solution Source (Human Crowd or Human-AI with prompt engineering levels)

	Dependent Variable: Top Decile Novelty Rating (0/1)					
	(1)	(2)	(3)	(4)	(5)	(6)
HAI Solution	−0.180***			−0.202***		
	(0.046)			(0.044)		
HAI Multiple Instance		−0.207***			−0.207***	
		(0.051)			(0.049)	
HAI Single Instance		−0.152**			−0.182***	
		(0.051)			(0.049)	
HAI Level 1			−0.124*			−0.203***
			(0.055)			(0.058)
HAI Level 2			−0.191***			−0.200***
			(0.055)			(0.060)
HAI Level 3			−0.224***			-0.204**
			(0.055)			(0.064)
Intercept	0.241***	0.241***	0.241***	−0.875	−0.860	−0.869
	(0.040)	(0.040)	(0.040)	(0.622)	(0.622)	(0.695)
N	234	234	234	234	234	234
Screening criteria	N	N	N	Y	Y	Y
Other controls	N	N	N	Y	Y	Y
R-squared	0.062	0.069	0.077	0.185	0.188	0.185

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

Notes. This table displays the results of ordinary least squares (OLS) analyses, focusing on the novelty ratings of solutions in the top decile. These ratings are derived from the average novelty score across all pairs of evaluators and solutions. In Models 3-6, we the following variables: Cohort Session and Solution Word Count, along with the average evaluator ratings for screening criteria such as Work Experience, Level of Interest, and Knowledge Test Score. Additionally, these models account

	(1)	(2)	(3)	(4)	(5)	(6)

for the average values of demographic and background covariates, namely Gender, Highest Level of Education, Major, and Employment Status. Standard errors are in parentheses.

Table C5. Nested Mixed Effects Models of Evaluator Environmental Value Ratings on Solution Source (Human Crowd or Human-AI)

Dependent Variable: Environmental Value Rating						
	(1)	(2)	(3)	(4)	(5)	(6)
HAI Solution	0.180***	0.180***	0.160***			
	(0.029)	(0.029)	(0.029)			
HAI Multiple Instance				0.148***	0.148***	0.136***
				(0.032)	(0.032)	(0.032)
HAI Single Instance				0.212***	0.212***	0.186***
				(0.032)	(0.032)	(0.033)
Intercept	3.616***	3.354***	3.375***	3.616***	3.354***	3.382***
	(0.042)	(0.191)	(0.224)	(0.042)	(0.191)	(0.224)
N	3900	3900	3900	3900	3900	3900
# blocks	18	18	18	18	18	18
# evaluators	300	300	300	300	300	300
Screening criteria	N	Y	Y	N	Y	Y
Other controls	N	N	Y	N	N	Y
Log-Likelihood	-4798.23 df = 5	-4801.19 df = 8	-4798.07 df = 15	-4798.25 df = 6	-4801.21 df = 9	-4799.14 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 environmental value, with 300 evaluators nested in eighteen solution blocks. Models 2-3 and 5-6 include the screening criteria: Work Experience, Level of Interest, and Knowledge Test Score. Models 3 and 6 include the following covariates: Gender, Highest Level of Education, Major, Employment Status, Cohort Session, and Solution Word Count. Standard errors are in parentheses.

Table C6. Nested Mixed Effects Models of Evaluator Top Environmental Value Ratings on Solution Source (Human Crowd or Human-AI)

Dependent Variable: Top Environmental Value Rating (0/1)						
	(1)	(2)	(3)	(4)	(5)	(6)
HAI Solution	0.021	0.021	0.017			
	(0.013)	(0.013)	(0.014)			
HAI Multiple Instance				0.016	0.016	0.014
				(0.015)	(0.015)	(0.015)
HAI Single Instance				0.026+	0.026+	0.021
				(0.015)	(0.015)	(0.015)
Intercept	0.228***	0.050	0.085	0.228***	0.050	0.086
	(0.019)	(0.082)	(0.098)	(0.019)	(0.082)	(0.098)
N	3900	3900	3900	3900	3900	3900
# blocks	18	18	18	18	18	18
# evaluators	300	300	300	300	300	300
Screening criteria	N	Y	Y	N	Y	Y
Other controls	N	N	Y	N	N	Y
Log-Likelihood	-1797.71 df = 5	-1799.13 df = 8	-1814.50 df = 15	-1797.87 df = 6	-1802.30 df = 9	-1817.79 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 environmental value, with 300 evaluators nested in eighteen solution blocks. Models 2-3 and 5-6 include the screening criteria: Work Experience, Level of Interest, and Knowledge Test Score. Models 3 and 6 include the following covariates: Gender, Highest Level of Education, Major, Employment Status, Cohort Session, and Solution Word Count. Standard errors are in parentheses.

Table C7. Nested Mixed Effects Models of Evaluator Financial Value Ratings on Solution Source (Human Crowd or Human-AI)

Dependent Variable: Financial Value Rating						
	(1)	(2)	(3)	(4)	(5)	(6)
HAI Solution	0.162***	0.162***	0.143***			
	(0.032)	(0.032)	(0.033)			
HAI Multiple Instance				0.172***	0.172***	0.160***
				(0.036)	(0.036)	(0.036)
HAI Single Instance				0.153***	0.153***	0.126***
				(0.036)	(0.036)	(0.037)
Intercept	3.086***	2.805***	2.658***	3.086***	2.805***	2.654***
	(0.043)	(0.185)	(0.217)	(0.043)	(0.185)	(0.217)
N	3900	3900	3900	3900	3900	3900
# blocks	18	18	18	18	18	18
# evaluators	300	300	300	300	300	300
Screening criteria	N	Y	Y	N	Y	Y
Other controls	N	N	Y	N	N	Y
Log-Likelihood	-5173.57 df = 5	-5170.42 df = 8	-5169.74 df = 15	-5175.95 df = 6	-5172.79 df = 9	-5171.67 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 financial value, with 300 evaluators nested in eighteen solution blocks. Models 2-3 and 5-6 include the screening criteria: Work Experience, Level of Interest, and Knowledge Test Score. Models 3 and 6 include the following covariates: Gender, Highest Level of Education, Major, Employment Status, Cohort Session, and Solution Word Count. Standard errors are in parentheses.

Table C8. Nested Mixed Effects Models of Evaluator Top Financial Value Ratings on Solution Source (Human Crowd or Human-AI)

Dependent Variable: Top Financial Value Rating (0/1)						
	(1)	(2)	(3)	(4)	(5)	(6)
HAI Solution	0.007	0.007	0.005			
	(0.010)	(0.010)	(0.010)			
HAI Multiple Instance				0.014	0.014	0.013
				(0.011)	(0.011)	(0.011)
HAI Single Instance				0.000	0.000	−0.004
				(0.011)	(0.011)	(0.011)
Intercept	0.092***	0.034	0.005	0.092***	0.034	2.654***
	(0.012)	(0.047)	(0.056)	(0.012)	(0.047)	(0.217)
N	3900	3900	3900	3900	3900	3900
# blocks	18	18	18	18	18	18
# evaluators	300	300	300	300	300	300
Screening criteria	N	Y	Y	N	Y	Y
Other controls	N	N	Y	N	N	Y
Log-Likelihood	-593.51 df = 5	-594.81 df = 8	-614.87 df = 15	-596.09 df = 6	-597.39 df = 9	-617.08 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 financial value, with 300 evaluators nested in eighteen solution blocks. Models 2-3 and 5-6 include the screening criteria: Work Experience, Level of Interest, and Knowledge Test Score. Models 3 and 6 include the following covariates: Gender, Highest Level of Education, Major, Employment Status, Cohort Session, and Solution Word Count. Standard errors are in parentheses.

Table C9. Nested Mixed Effects Models of Evaluator Novelty Ratings on Solution Source (Human Crowd or Human-AI)

Dependent Variable: Novelty Rating			
	(1)	(2)	(3)
HAI Level 1	-0.135*	-0.132*	-0.138*
	(0.054)	(0.054)	(0.054)
HAI Level 2	-0.070	-0.071	-0.093+

	(1)	(2)	(3)
	(0.054)	(0.054)	(0.054)
HAI Level 3	-0.168**	-0.170**	-0.190***
	(0.054)	(0.054)	(0.545)
Intercept	3.508***	3.389***	3.238***
	(0.049)	(0.181)	(0.215)
N	3900	3900	3900
# blocks	18	18	18
# evaluators	300	300	300
Screening criteria	N	Y	Y
Other controls	N	N	Y
Log-Likelihood	-5432.99 df = 7	-5437.65 df = 10	-5442.61 df = 17

+ 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, with 300 evaluators nested in eighteen solution blocks. Model 2 includes the screening criteria: Work Experience, Level of Interest, and Knowledge Test Score. Model 3 includes the following covariates: Gender, Highest Level of Education, Major, Employment Status, Cohort Session, and Solution Word Count. Standard errors are in parentheses.

Table C10. Nested Mixed Effects Models of Evaluator Top Novelty Ratings on Solution Source (Human Crowd or Human-AI with prompt engineering levels)

Dependent Variable: Top Novelty Rating (0/1)

	(1)	(2)	(3)
HAI Level 1	-0.091***	-0.090***	-0.090***
	(0.019)	(0.019)	(0.019)
HAI Level 2	-0.038*	-0.038*	-0.046*
	(0.019)	(0.019)	(0.019)
HAI Level 3	-0.093***	-0.094***	-0.100***
	(0.019)	(0.019)	(0.019)
Intercept	0.208***	0.192***	0.183**
	(0.017)	(0.058)	(0.069)
N	3900	3900	3900
# blocks	18	18	18
# evaluators	300	300	300
Screening criteria	N	Y	Y
Other controls	N	N	Y
Log-Likelihood	-1332.84 df = 7	-1341.40 df = 10	-1360.45 df = 17

+ 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, with 300 evaluators nested in eighteen solution blocks. Model 2 includes the screening criteria: Work Experience, Level of Interest, and Knowledge Test Score. Model 3 includes the following covariates: Gender, Highest Level of Education, Major, Employment Status, Cohort Session, and Solution Word Count. Standard errors are in parentheses.

Table C11. Nested Mixed Effects Models of Evaluator Value Ratings on Solution Source (Human Crowd or Human-AI with prompt engineering levels)

Dependent Variable: Value Rating

	(1)	(2)	(3)
HAI Level 1	0.180***	0.184***	0.179***
	(0.040)	(0.040)	(0.040)
HAI Level 2	0.101*	0.099*	0.074+
	(0.040)	(0.039)	(0.040)
HAI Level 3	0.233***	0.230***	0.203***
	(0.040)	(0.040)	(0.040)
Intercept	3.351***	3.066***	2.986***

	(1)	(2)	(3)
	(0.038)	(0.172)	(0.201)
N	3900	3900	3900
# blocks	18	18	18
# evaluators	300	300	300
Screening criteria	N	Y	Y
Other controls	N	N	Y
Log-Likelihood	-4295.93 df = 7	-4294.91 df = 10	-4288.82 df = 17

+ 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 value, with 300 evaluators nested in eighteen solution blocks. Model 2 includes the screening criteria: Work Experience, Level of Interest, and Knowledge Test Score. Model 3 includes the following covariates: Gender, Highest Level of Education, Major, Employment Status, Cohort Session, and Solution Word Count. Standard errors are in parentheses.

Table C12. Nested Mixed Effects Models of Evaluator Top Value Ratings on Solution Source (Human Crowd or Human-AI with prompt engineering levels)

Dependent Variable: Top Value Rating (0/1)

	(1)	(2)	(3)
HAI Level 1	0.018	0.020+	0.019
	(0.012)	(0.012)	(0.012)
HAI Level 2	0.012	0.011	0.012
	(0.012)	(0.012)	(0.012)
HAI Level 3	−0.002	−0.003	−0.002
	(0.012)	(0.012)	(0.012)
Intercept	0.052***	0.017	0.030
	(0.009)	(0.039)	(0.047)
N	3900	3900	3900
# blocks	18	18	18
# evaluators	300	300	300
Screening criteria	N	Y	Y
Other controls	N	N	Y
Log-Likelihood	-1978.05 df = 7	-1979.49 df = 10	-1991.76 df = 17

+ 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 value, with 300 evaluators nested in eighteen solution blocks. Model 2 includes the screening criteria: Work Experience, Level of Interest, and Knowledge Test Score. Model 3 includes the following covariates: Gender, Highest Level of Education, Major, Employment Status, Cohort Session, and Solution Word Count. Standard errors are in parentheses.

Table C13. Nested Mixed Effects Models of Evaluator Creativity Ratings (Novelty x Value) on Solution Source (Human Crowd or Human-AI with prompt engineering levels)

Dependent Variable: Creativity Rating

	(1)	(2)	(3)
HAI Level 1	0.029	0.063	0.028
	(0.268)	(0.267)	(0.265)
HAI Level 2	0.077	0.058	−0.125
	(0.265)	(0.264)	(0.266)
HAI Level 3	0.091	0.077	−0.099
	(0.267)	(0.266)	(0.268)
Intercept	12.256***	10.772***	10.031***
	(0.245)	(1.115)	(1.303)
N	3900	3900	3900
# blocks	18	18	18
# evaluators	300	300	300

	(1)	(2)	(3)
Screening criteria	N	Y	Y
Other controls	N	N	Y
Log-Likelihood	-11722.06 df = 7	-11717.25 df = 10	-11699.72 df = 17

+ 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 creativity, with 300 evaluators nested in eighteen solution blocks. Model 2 includes the screening criteria: Work Experience, Level of Interest, and Knowledge Test Score. Model 3 includes the following covariates: Gender, Highest Level of Education, Major, Employment Status, Cohort Session, and Solution Word Count. Standard errors are in parentheses.

Table C14. Nested Mixed Effects Models of Evaluator Top Creativity Ratings (Novelty x Value) on Solution Source (Human Crowd or Human-AI with prompt engineering levels)

	Dependent Variable: Top Creativity Rating (0/1)		
	(1)	(2)	(3)
HAI Level 1	0.004	0.005	0.005
	(0.009)	(0.009)	(0.009)
HAI Level 2	−0.006	−0.006	−0.006
	(0.009)	(0.009)	(0.009)
HAI Level 3	−0.011	−0.011	−0.011
	(0.009)	(0.009)	(0.009)
Intercept	0.038***	0.033	0.033
	(0.007)	(0.027)	(0.033)
N	3900	3900	3900
# blocks	18	18	18
# evaluators	300	300	300
Screening criteria	N	Y	Y
Other controls	N	N	Y
Log-Likelihood	1229.69 df = 7	1220.51 df = 10	1196.20 df = 17

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

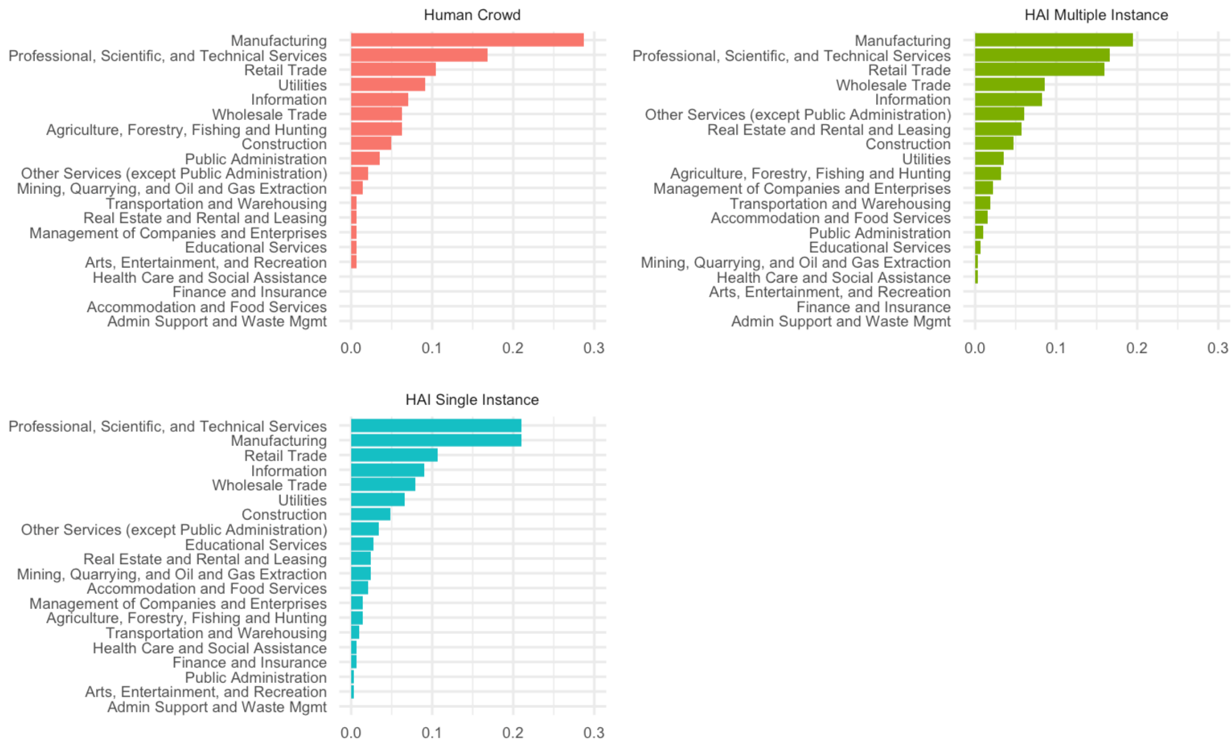
Notes. This table presents mixed-model (hierarchical linear modeling) results from evaluator ratings of top solution creativity, with 300 evaluators nested in eighteen solution blocks. Model 2 includes the screening criteria: Work Experience, Level of Interest, and Knowledge Test Score. Model 3 includes the following covariates: Gender, Highest Level of Education, Major, Employment Status, Cohort Session, and Solution Word Count. Standard errors are in parentheses.

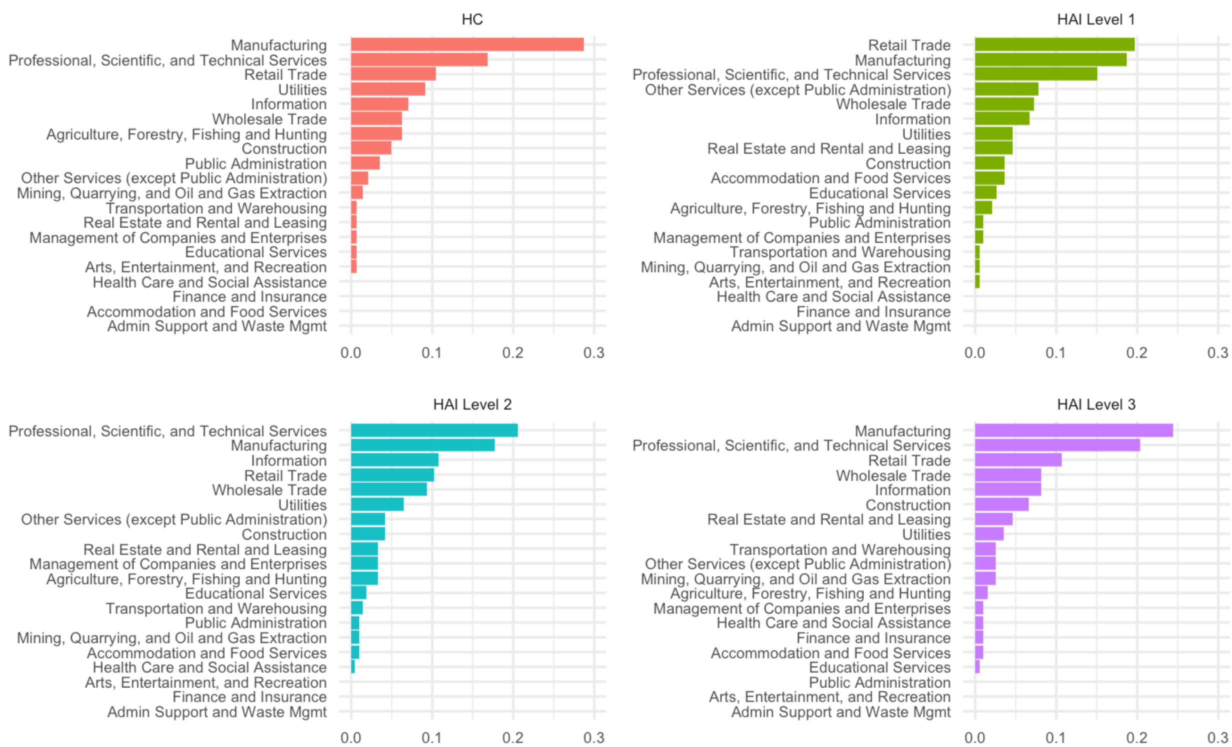
Appendix D: Additional Text Analysis of Human Crowd (HC) and Human-AI (HAI) Solutions

D.1 ChatGPT Results of Industry Classifications

To further explore the rich-text content of our HC and HAI generated solutions, we used OpenAI's GPT-4 to classify them into distinct industry groups. We chose the standardized 2-digit NAICS (North American Industry Classification System) sectors from 2022. For these classification tasks, we did not restrict the number of classifications per solution, meaning that a given solution could fall into one or more industries of application. Figure C1 displays the diversity of industry applications, sorted by the aggregated frequency of industry classes. Here, we see "Manufacturing" and "Professional, Scientific, and Technical Services" are consistently represented as the top two industries. This indicates that these are common bases across ideas generated by HC and HAI followed by other industries such as "Retail Trade" or "Information." Below we plot the share representation of industry classification by solution source (HC, HAI Multiple Instance, HAI Single Instance) as well as by prompt engineering levels (HAI Level 1, HAI Level 2, HAI Level 3).

Figure D1. Comparisons of Industry Classifications by Solution Source and HC and HAI Levels





We observe that within solution source or prompt engineering level, both HAI and HC have a wide range of distribution in terms of application industries. HAI Multiple and Single Instance are slightly more diverse as they are less concentrated in "Manufacturing". However, the overall share representations between HC and HAI groups resemble each other.

To further assess the diversity of each prompting level, we used the Shannon diversity index—a metric often used to measure the diversity of species in a specific space. The Shannon index H, can be normalized to the Shannon equitability index which takes values between 0 to 1, denoted as E_H. The Shannon equitability index within a level that contains k total sectors composing of proportion p of sector i is measured as:

$$E_H = \frac{-\sum_{i=1}^{k} p_i * log(p_i)}{log(k)}.$$

We calculated Level HC = 0.793, HAI Level 1 = 0.820, HAI Level 2 = 0.824, and HAI Level 3 = 0.799, which suggests that AI and human levels generated similar degrees of diversity of industry applications.

D.2 Principal Component Analysis

To further exploit the rich text of our solutions, we transformed the text into BERT word embeddings and performed principal component analysis (PCA), a statistical technique for dimensionality reduction, to project the text of the solutions onto 2-dimensional space. PCA reduces the dimensionality of the text data, which can assist with plotting, visualizing, and identifying patterns in the data. Figures D2-D4 showcase that the AI solutions appear to cluster more centrally, while human solutions exhibit a broader spread, suggesting greater variance in the PCA space. This spatial distribution indicates the underlying diversity intrinsic to HC, compared to HAI solutions. Moreover, although there is a degree of overlap, each level occupies a relatively distinguishable region in the PCA space, hinting at underlying differences in the characteristics across sources of the solutions.

Figure D2. PCA projections by HC and HAI

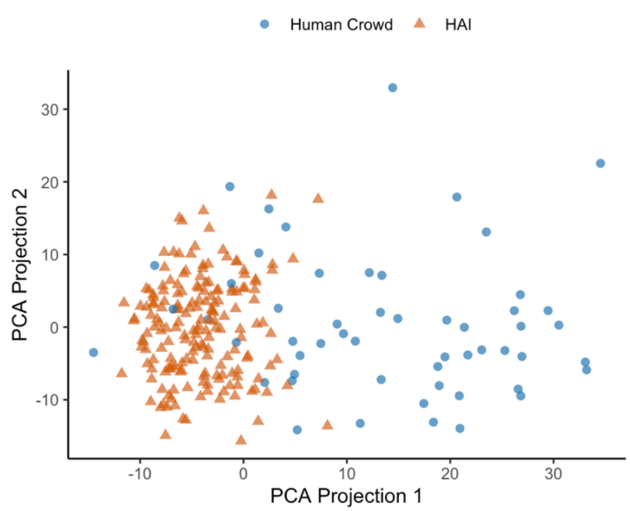


Figure D3. PCA projections by HC and HAI instances (M/S)

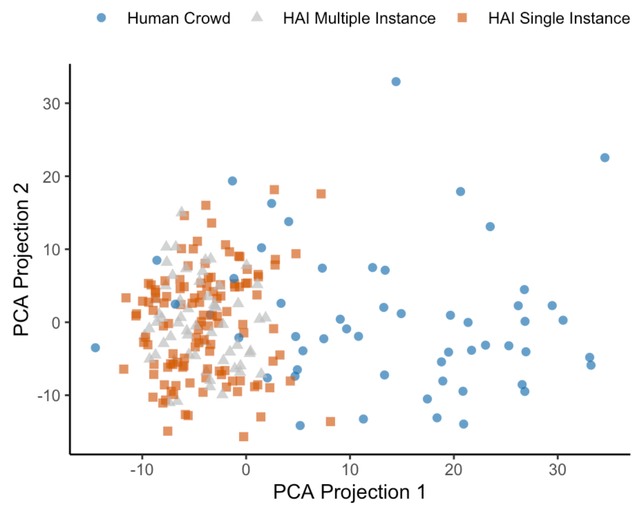
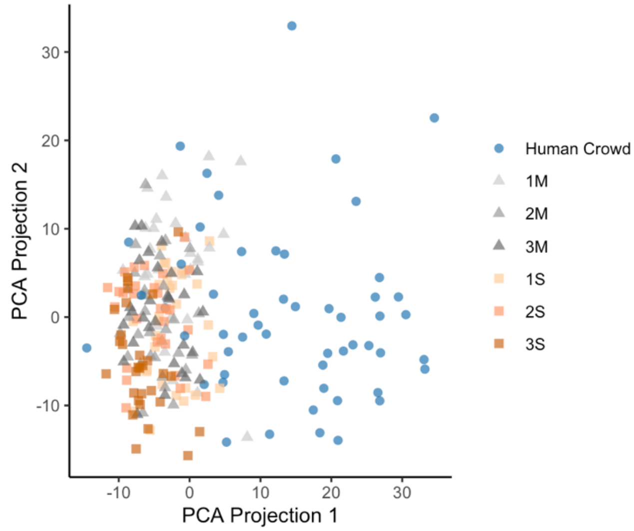


Figure D4. PCA projections by all HC and HAI levels and model configurations



Lastly, we color-coded the aggregated novelty and value ratings of each HC- and HAI-generated solutions on the PCA plots. The spread and density of points in Figure D5 suggest that highly novel ideas are scattered, arising from more diverse regions, most of which originate from HC. Figure D6, on the other hand, color-codes aggregated value ratings and suggests that high-value solutions are more confined to a specific area within the PCA projections, indicating that value may tend to concentrate more in certain regions of the solution space than novelty.

Figure D5. PCA projections colored by novelty

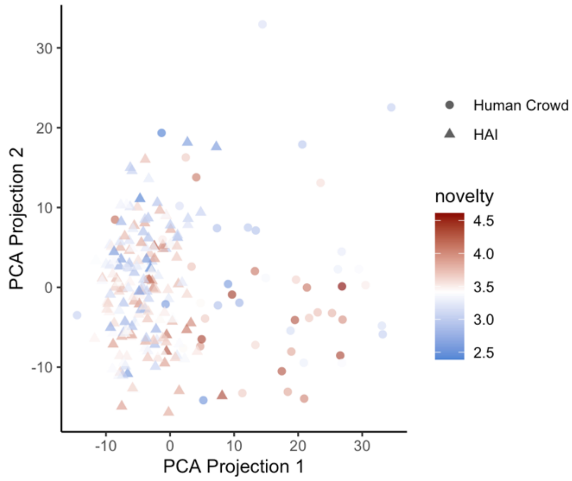
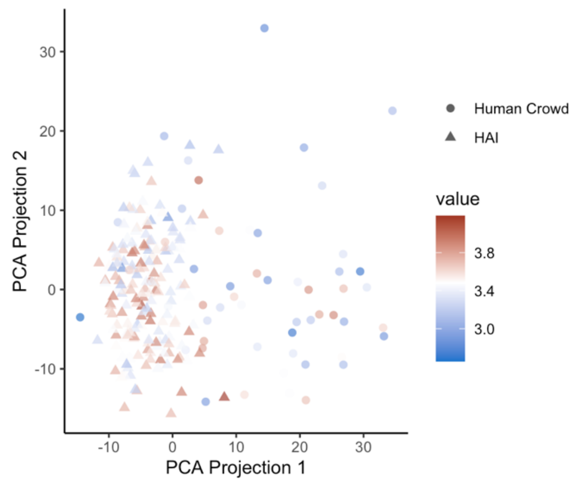


Figure D6. PCA projections colored by value



Appendix E: Survey Materials

E.1 Screening Survey Materials

Q0 Welcome to this 5-minute screening survey. We will ask you a few questions about yourself and your domain knowledge in the circular economy. You will be paid $1 for completion of the screening.

Depending on your responses, we will determine your eligibility to participate in the follow-up evaluation task.

First, what is your Prolific ID?

____

Q1 Where are you located?

o United States (1)

o Outside United States (2)

Q2 How old are you?

o under 18 (1)

o 18-24 (2)

o 25-34 (3)

o 35-44 (4)

o 45-54 (5)

o 55-64 (6)

o 65 or older (7)

End of Block: Welcome

Start of Block: Screening1 - self-claimed interest + expertise

Q3 How interested are you in the problem of circular economy?

	1 (little to no interest) (1)	2 (2)	3 (moderate interest) (3)	4 (4)	5 (very much interest) (5)
Interest (1)	o	o	o	o	o

Q4 List all industries you have previously worked (outside of educational experience)?

	Have you worked in this field?		If yes, for how many years?
	Yes (1)	No (2)	
Apparel & Textiles (1)	o	o	▼ less than 1 year (1 ... > 20 years (6)
Automobiles & Tires (2)	o	o	▼ less than 1 year (1 ... > 20 years (6)
Beverages (3)	o	o	▼ less than 1 year (1 ... > 20 years (6)
Building Products (4)	o	o	▼ less than 1 year (1 ... > 20 years (6)
Buildings (5)	o	o	▼ less than 1 year (1 ... > 20 years (6)
Construction Machinery (6)	o	o	▼ less than 1 year (1 ... > 20 years (6)

Construction Materials (7)	o	o	▼ less than 1 year (1 ... > 20 years (6)
Consumer Electronics (8)	o	o	▼ less than 1 year (1 ... > 20 years (6)
Containers & Packaging (9)	o	o	▼ less than 1 year (1 ... > 20 years (6)
Cosmetics (10)	o	o	▼ less than 1 year (1 ... > 20 years (6)
Electrical Equipment (11)	o	o	▼ less than 1 year (1 ... > 20 years (6)
Financials (13)	o	o	▼ less than 1 year (1 ... > 20 years (6)
Food (14)	o	o	▼ less than 1 year (1 ... > 20 years (6)
Forest Products (15)	o	o	▼ less than 1 year (1 ... > 20 years (6)
Health Care Products (16)	o	o	▼ less than 1 year (1 ... > 20 years (6)
Home Furnishings (17)	o	o	▼ less than 1 year (1 ... > 20 years (6)
Household Appliances (18)	o	o	▼ less than 1 year (1 ... > 20 years (6)
Industrial Machinery (19)	o	o	▼ less than 1 year (1 ... > 20 years (6)

Metals: beyond big 4 (20)	o	o	▼ less than 1 year (1 ... > 20 years (6)
Software & IT Services (21)	o	o	▼ less than 1 year (1 ... > 20 years (6)
Technology/ Hardware Products (22)	o	o	▼ less than 1 year (1 ... > 20 years (6)
Transportation & Logistics (23)	o	o	▼ less than 1 year (1 ... > 20 years (6)
Waste Management (24)	o	o	▼ less than 1 year (1 ... > 20 years (6)
Other (25)	o	o	▼ less than 1 year (1 ... > 20 years (6)

End of Block: Screening2 - work exp

Start of Block: Screening3 - skills

Q5 Welcome to the skills test! Below you will find 5 multiple choice questions related to your claimed area of expertise. Please answer them carefully. *To avoid plagiarism, we will be grading not only based on your accuracy but also on your time of completion.*

Which of the following principles is NOT associated with a circular economy?

o Waste as a resource (1)

o System effectiveness (2)

o Long-term usage (3)

o Linear consumption (4)

o User of renewable energy (5)

Q6 The circular economy envisions waste as:

o A necessary byproduct of production (1) o A resource that should be minimized (2)

o A cost to be managed and reduced (3)

o An unavoidable aspect of human activity (4)

o An indicator of inefficient resource use (5)

Q7 In the circular economy, what does "product as a service" mean?

o Renting out products as services (1)

o Transforming products into services (2)

o Charging for the service a product provides, rather than the product itself (3)

o Offering complimentary services with the product (4)

o Selling services instead of products (5)

Q8 What is one potential challenge of transitioning to a circular economy?

o Initial investment cost (1)

o Increased product durability (2)

o Reduced reliance on non-renewable resources (3)

o Reduction in waste production (4)

o Increased use of renewable energy (5)

Q9 Which of the following industries has commonly adopted the circular economy model?

o Fossil fuel energy production (1)

o Single-use plastic manufacturing (2)

o Furniture manufacturing (3)

o Fast-fashion clothing (4)

o Lead-acid battery production (5)

End of Block: Screening3 - skills

Start of Block: Congrats

Q10 Congratulations! You have been selected to participate in the evaluation task.

We expect the follow-up evaluation survey to take ~30 minutes. We will first ask you some demographic information about yourself, followed by evaluations of 13 solutions provided for a large research institution to understand the impact of circular economy on business.

We will pay you $12 for your time and effort. Additionally, you will have the opportunity to receive up to $13 in bonuses depending on your performance, for a maximum compensation of $25.

Are you willing to participate in the follow-up evaluation task?

o Yes (1)

o No (2)

End of Block: Congrats

Display This Question:

If Congratulations! You have been selected to participate in the evaluation task. We expect the foll... = Yes

Q11-0 Thanks for your interest! The follow-up evaluation survey will be active on Prolific soon. It will pop up in your Prolific feed after we custom-invite everyone who is eligible to participate based on the screening survey. The title of the study will be something like "Evaluating Circular Economy Solutions."

E.2 Evaluation Survey Instructions and Demographic Information

Start of Block: Consent

Q00 **Information**

The following is a short summary of this study to help you decide whether to be a part of this study. More detailed information is listed later in this form.

Why am I being invited to take part in a screening for this research study?

We invite you to take part in this study screening because you are over 18 years old and reside in the United States.

What should I know about a research study?

Someone will explain this research study to you.
Whether or not you take part is up to you.
Your participation is completely voluntary.
You can choose not to take part.
You can agree to take part and later change your mind.
Your decision will not be held against you.
Your refusal to participate will not result in any consequences or any loss of benefits that you are otherwise entitled to receive.
You can ask all the questions you want before you decide.

Why is this research being done?

The goal of this study is to understand how people evaluate the creativity of crowdsourced solutions.

Crowdsourcing leverages many diverse perspectives to improve the likelihood of getting high-value solutions. We want to understand how people evaluate these solutions.

How long will the research last and what will I need to do?

We expect that you will be in this research study for up to 30 minutes. You will be asked to evaluate 13 solutions to a crowdsourcing challenge using an evaluation framework. You will be asked to evaluate these solutions based on multiple criteria including novelty, feasibility, and impact). After this task, you will be asked to answer a short demographics questionnaire.

Is there any way being in this study could be bad for me?

We don't believe there are any risks from participating in this research.

Will being in this study help me in any way?

There are no benefits to you from your taking part in this research. We cannot promise any benefits to others from your taking part in this research. However, possible benefits to others include creating frameworks to more successfully evaluate high-value solutions to problems.

What happens if I do not want to be in this research?

Participation in research is completely voluntary. You can decide to participate, not participate, or discontinue participation at any time without penalty or loss of benefits to which you are otherwise entitled. Your alternative to participating in this research study is to not participate.

Detailed Information

The following is more detailed information about this study in addition to the information listed above.

What happens if I say yes, but I change my mind later?

You can leave the research at any time; it will not be held against you. Any data you created will be destroyed and not used for research.

Is there any way being in this study could be bad for me? (Detailed Risks)

We will do our best to protect your data during storage and when they are shared. However, there remains a possibility that someone could identify you. There is also the possibility that people who are not supposed to might access your data and samples. In either case, we cannot reduce the risk to zero.

If I take part in this research, how will my privacy be protected? What happens to the information you collect?

Efforts will be made to limit the use and disclosure of your Personal Information, including name and email if provided, to people who have a need to review this information. We cannot promise complete secrecy. Organizations that may inspect and copy your information include the IRB and other representatives of this organization.

Some of your data from the screening survey will be merged with your data from the main study upon completion of the main study tasks. If you do not agree to take part in the main study or leave the study early, we will destroy your screening data and any data created for the main study.

If identifiers are removed from your identifiable private information that are collected during this research, that information could be used for future research studies or distributed to another investigator for future research studies without your additional informed consent.

Compensation

If you agree to take part in this research study, you will receive a base pay of $12. Additionally, you will have the opportunity to earn $13 in bonuses for each solution where your rating is aligned with the consensus or the mode rating among all evaluators rating the same solution. In other words, for each solution where your rating aligns with the mode rating of the other evaluators, you will receive an extra $1 of compensation, for a total of $13 across the 13 solutions you will be asked to rate. The maximum total compensation you will receive is $25.

Who can I talk to?

If you have questions, concerns, or complaints, or think the research has hurt you, talk to the research

team at <redacted>.
This research has been reviewed and approved by [redacted] Institutional Review Board ("IRB"). You may talk to them at [redacted] if:

- Your questions, concerns, or complaints are not being answered by the research team.
- You cannot reach the research team.
- You want to talk to someone besides the research team.
- You have questions about your rights as a research subject.
- You want to get information or provide input about this research.

Do you consent to participate in this study?

- ○ Yes (1)
- ○ No (2)

End of Block: Consent

Start of Block: Instructions

QID1 Thank you for participating in our research study!

What is your Prolific ID?

____

QID2 Instructions

Today you will be asked to evaluate 13 circular economy ideas to help a large research institution understand the impact of circular economy on business. We expect this survey to take ~30 minutes. We will first ask you some demographic information about yourself, followed by the evaluation task.

Your task is to rate the solutions based on their performance across four criteria:
Novelty: How different is it from existing solutions?
Environmental Impact: How much does it benefit the planet?
Financial Impact: What financial value can it create for businesses?
Feasibility and Scalability of Implementation: How likely is it to succeed and how scalable is it?

Based on these four criteria, you will then assess the overall quality of the solution. Please rate each solution on a scale from 1 to 5, with 1 being poor and 5 being outstanding. All solutions are of the format 'Problem' (problem identified) followed by 'Solution' (proposed solution).

After completing the survey, don't forget to click the Prolific URL at the end of the survey to claim your base payment of $12. We will manually review your solutions to assess your bonus payment amount (up to $13).

End of Block: Instructions

Start of Block: evaluation_0

[13 problem and solution pairs each followed by the matrix box below]

	1 (Poor) (1)	2 (Below average) (2)	3 (Average) (3)	4 (Above average) (4)	5 (Excellent) (5)
Novelty (How different is it from existing solutions?) (1)	○	○	○	○	○
Feasibility and Scalability of Implementation (How likely is it to succeed and how scalable is it?) (2)	○	○	○	○	○
Environmental Impact (How much does it benefit the planet?) (3)	○	○	○	○	○
Financial Impact (What financial value can it create for businesses?) (4)	○	○	○	○	○
Quality (Based on the four criteria above, what is the overall quality of the solution?) (5)	○	○	○	○	○

End of Block: evaluation_0

Start of Block: evaluation_1

End of Block: evaluation_1

….

….

Start of Block: evaluation_17

End of Block: evaluation_17

Start of Block: Demographics

QID106 What is the highest level of education you have completed?

- o Some high school, no diploma (1)
- o High school graduate, diploma or the equivalent (2)
- o Trade/technical/vocational training (3)
- o Associate degree (4)
- o Bachelor's degree (5)
- o Master's degree (6)
- o Professional degree (7)
- o Doctorate degree (8)

QID107 In which field did you complete your highest level of education?

- o Arts and Humanities (1)
- o Social Sciences (2)
- o Business (3)
- o Life Sciences (4)
- o Physical Sciences (5)
- o Engineering (6)
- o Technology/Computer Science (7)
- o Education (8)
- o Health and Medicine (9)
- o Other (10) ____

QID108 What is your current employment status?

- o Employed (1)
- o Self-employed (2)
- o Unemployed (3)
- o Student (4)
- o Retired (5)

QID111 What is your gender?

o Male (1)

o Female (2)

o Prefer not to say (3)

End of Block: Demographics

Start of Block: End

QID112 Thanks for your participation! Please click the button below to be redirected back to Prolific and register your submission.

If you have any additional comments or feedback, please feel free to leave them below.

____

____

____

____

End of Block: End