Using Large Language Models for Idea Generation in Innovation

Lennart Meincke

Operations, Information and Decisions, The Wharton School, University of Pennsylvania, lennart@wharton.upenn.edu

Karan Girotra

Cornell Tech and Johnson College of Business, Cornell University, girotra@cornell.edu

Gideon Nave

Marketing, The Wharton School, University of Pennsylvania, gnave@wharton.upenn.edu

Christian Terwiesch

Operations, Information and Decisions, The Wharton School, University of Pennsylvania, terwiesch@wharton.upenn.edu

Karl T. Ulrich

Operations, Information and Decisions, The Wharton School, University of Pennsylvania, ulrich@wharton.upenn.edu

This research evaluates the efficacy of large language models (LLMs) in generating new product ideas. To do so, we compare three pools of ideas for new products targeted toward college students priced at $50 or less. The first pool of ideas was created by university students in a product design course before the availability of LLMs. The second and third pools of ideas were generated by OpenAI's GPT-4 using zero-shot and few-shot prompting, respectively. We evaluated idea quality using standard market research techniques to predict average purchase intent probability. We used text mining to assess idea similarity and human raters to evaluate idea novelty. We find that AI-generated ideas outperform human-generated ideas in terms of average purchase intent, with few-shot prompting yielding slightly higher intent than zero-shot prompting. However, AI-generated ideas are perceived as less novel and exhibit higher pairwise similarity, particularly with few-shot prompting, indicating a less diverse solution landscape. When focusing on the quality of the best ideas (rather than on the average ideas), we find that AI-generated ideas are seven times more likely to rank among the top 10% of ideas, demonstrating a significant advantage over human-generated ideas. We propose that this 7:1 advantage is a conservative estimate, as it does not account for AI's greater productivity. Our findings suggest that despite some drawbacks, AI creativity presents a substantial benefit in generating high-quality ideas for new product development.

Funding: Funding was provided by the Mack Institute for Innovation Management at the Wharton School of the University of Pennsylvania.

Key words: innovation; idea generation; creativity; creative problem solving; LLM; large-scale language models; AI; artificial intelligence; ChatGPT; GPT

1. Introduction

Generative artificial intelligence (GenAI) has remarkably advanced in creating life-like images and coherent, fluent text. Open AI's ChatGPT chatbot, based on the Generative Pre-trained Transformer (GPT) series of large language models (LLM), can equal or surpass human performance in academic examinations and tests for professional certifications (OpenAI et al. 2023). Moreover, LLMs can provide valuable professional advice in fields like software development, medicine, and law.

Despite their remarkable performance, LLMs sometimes produce text that is semantically or syntactically plausible but is, in fact, factually incorrect or nonsensical, a phenomenon often referred to as "hallucinations." This outcome is a byproduct of how LLMs are designed, as they are optimized to generate the most statistically likely sequences of words with an intentional injection of randomness. In most applications, this randomness and the associated hallucinations and inconsistencies create problems that limit the use of LLM-based solutions to low-stakes settings, or they require extensive human supervision.

But are there applications in which we can leverage the weaknesses of hallucinations and inconsistent quality and turn them into a strength? We propose that the domain of creativity and innovation provides such an application. This domain operates quite differently than most management settings, where we commonly expect to use each unit of work produced. As such, consistency is prized and is, therefore, the focus of contemporary performance management. Erratic and inconsistent behavior is to be eliminated. An airline would rather hire a pilot that executes a within-safety-margins landing 10 out of 10 times rather than one that makes a brilliant approach five times and an unsafe approach another five. But, when it comes to creativity and innovation, say finding a new opportunity to improve the air travel experience or launching a new aviation venture, the same airline would prefer an ideator that generates one brilliant idea and nine nonsense ideas over one that generates ten decent ideas.

The reason for this difference is that when it comes to creativity and innovation, the performance of the process is not determined by the sum or the average of all ideas created. Instead, each idea is seen as a real option that the decision maker can decide to execute (Huchzermeier and Loch 2001). Thus, the performance of the process is determined by the quality of the best idea(s) (Dahan and Mendelson 2001, Terwiesch and Xu 2008, Terwiesch and Ulrich 2009, Girotra et al. 2010). The process of innovation thereby can be thought of as a search process that generates ideas with random quality values by drawing from an underlying stochastic distribution until the cost of creating one additional draw from the distribution (e.g., creating one more product concept or building one more prototype) exceeds the marginal benefit (Weitzman 1979).

Prior research in product development and innovation has modeled various aspects of this search process, including the pros and cons of parallel search (Loch et al. 2001), the tension between sampling from very different regions of the pay-off distributions ("selectionism") versus locally improving a given project (Sommer and Loch 2004), and the need for building balanced portfolios that consist of different

types of projects (Chao and Kavadias 2008).

We follow this line of research and consider a setting in which ideas of unknown quality are created, and the quality of the best few ideas determines the overall performance. This could be a setting of corporate portfolio planning in a large established organization as described by Si et al. (2022). However, to facilitate our experimental design, we focus on the idea generation in the product developement process for a newly formed venture. Specifically, we look for a product idea that targets the college student market and can be sold for $50 or less. This innovation challenge is similar to the study settings used in prior work (e.g., Osborn 1953, Connolly et al. 1990, Sutton and Hargadon 1996, Girotra et al. 2010) to evaluate and compare various brainstorming methods (e.g., group vs. individual; nominal groups vs. hybrid groups).

In contrast to this prior work, we consider ideas generated by humans *and* ideas generated by artificial intelligence (AI) in the form of Open AI's GPT-4. As discussed above, LLMs are designed to generate new content, and in the domain of brainstorming, their stochastic (if not outright erratic) behavior might turn a bug into a feature. Thus, we hypothesize that LLMs have the potential to be excellent ideators. The purpose of this paper, therefore, is to formally test this hypothesis by comparing the performance of LLMs in generating new ideas to that of human idea generators.

Specifically, we compare three pools of ideas for new products targeted toward college students at a price of USD 50 or less. The first pool of ideas was created by students at an elite university enrolled in a course on product design before the availability of LLMs. The second pool was generated by OpenAI's GPT-4 with the same prompt as that given to the students and no other guidance (zero-shot prompting). The third pool was generated by prompting GPT-4 with the same prompt as that given to the students and a sample of highly rated ideas to enable some in-context learning (few-shot prompting). We evaluate the quality of the ideas using standard market research techniques and survey human respondents to predict an average purchase intent probability for each product, which we use as our measure of idea quality. We use text mining techniques to evaluate the similarities of ideas and rely on human raters to assess idea novelty.

This comparison between human idea generation and AI-based idea generation allows us to contribute to the innovation literature by establishing the following novel results.

First, AI-generated ideas are, on average, significantly better (average purchase intent of 0.48 relative to 0.40 for human-generated ideas), especially in the case of few-shot prompting (average purchase intent of 0.49 relative to 0.46 for zero-shot prompting), as shown in Study 1.

Second, despite this success, consumers perceive AI-generated ideas as less novel (perceived novelty of 0.36 relative to 0.41). Moreover, AI-generated ideas are more likely to overlap: text mining reveals that the average pairwise similarity of ideas is higher among AI-generated ideas and further increases when using few-shot prompting. As a result, the underlying solution landscape is less likely to be fully explored (Study 2).

Finally, we show that for a given number of ideas, the quality of the best ideas generated by AI is significantly greater than that of the best ideas generated by humans (Study 3). Specifically, we show that AI-generated ideas are seven times more likely to be among the top 10% of ideas generated in our experiment. This is significant given the context. What matters for innovation is the quality of the *best* idea. The objective of idea generation is to generate at least a few truly exceptionally great ideas. In most innovation settings, we would rather have 10 great ideas and 90 terrible ideas than 100 ideas of solid quality. Holding the number of ideas constant, we need to trade off the advantageous effect of higher average idea quality (Study 1) with the disadvantages of less novelty, more overlapping ideas, and fewer ideas that can be discovered (Study 2). Study 3 clearly establishes AI's supremacy over humans in this respect.

A quarter of a century ago, Goldenberg et al. (1999) asked the question "Can AI-generated ideas finally compete with human ones, long after researchers first considered the possibility?". We believe that the three studies presented in this article provide empirical support for an affirmative answer to this question. From a practical perspective, we see the 7:1 advantage of AI creativity over human creativity as a conservative estimate, as we did not credit AI for its substantially greater productivity.

The remainder of the article is organized as follows. After reviewing some recent work on GenAI and creativity (Section 2), we introduce our theoretical framework and our hypotheses (Section 3), followed by the technical set-up of our experiments (Section 4). We conducted three studies to assess the creativity of human- and AI-generated ideas. First, in Study 1 we ask human participants to rate ideas from both sources (human- and AI-generated) and compare the results (Section 5). Second, in Study 2 we use text-based analysis to calculate how many unique ideas can be created by humans and LLMs in our specific domain as well as ask human participants to rate the novelty of ideas from both sources and compare the results (Section 6). Third, in Study 3 we look at the extreme distributions of idea quality to identify possible advantages for the best ideas by either humans or AI (Section 7). We conclude the paper by discussing potential limitations of our studies, their robustness to alternative specifications (Section 8), and the implications of our findings (Section 9).

2. GenAI applications to creative tasks

Research to date has demonstrated three key findings regarding AI's role in creativity and innovation. First, AI frequently matches or exceeds human performance in creative tasks. Haase and Hanel (2023) found that LLMs have reached human-level performance in divergent thinking tasks such as the Alternative Uses Task (AUT). This is supported by Hubert et al. (2024), who studied GPT-4 responses for the Consequences Task and Divergent Association Tasks, finding that AI is more creative than humans across all its dimensions. While Koivisto and Grassini (2023) find that AI chatbots outperform average human performance in the

AUT, they also note that the most exceptional human ideas still match or exceed those generated by AI.

Second, studies show that AI aids in improving creative outcomes for humans when using it as a tool. Doshi and Hauser (2024) find that AI use helps humans to create more creative and enjoyable short stories. However, the collective diversity decreases and stories become more similar to one another. Similarly, Jia et al. (2024) found that AI assistance boosted employee creativity in a telemarketing company when responding to customer questions, ultimately increasing sales. Zhou and Lee (2024) show that integrating text-to-image AI into creative workflows increased the number of artworks created by 25% and raised the likelihood of receiving the works receiving favorite per view by 50%, highlighting the benefits of LLMs augmenting human workflows ("human in the loop").

Third, studies have explored human preferences for AI-generated versus human-generated creations, often finding that people prefer human involvement. For instance, Hitsuwari et al. (2023) found that survey participants cannot distinguish between AI-generated and human-generated haikus, but rated poems co-created by humans and AI as the most beautiful with no significant preference for haikus created solely by humans or AI. Bellaiche et al. (2023) provide evidence that humans prefer human involvement in art creation by showing that participants prefer AI-generated art falsely labeled as created by humans to the same art correctly labeled as AI-generated, suggesting a bias for human involvement in the creative process. Similarly, Shank et al. (2023) find comparable results for AI-generated classical music, although no such preference was found for electronic music. However, Zlatkov et al. (2023) found no significant preference for either AI or human-generated music overall.

Taken together, this body of research illustrates the potency of AI in creative tasks. AI not only matches human creativity but also improves human performance when used as a collaborative tool. However, at least when considering artistic outcomes, there remains a human preference for creativity that involves human touch. This growing evidence suggests a natural next step: evaluating AI's efficacy in innovation management in general and in idea generation in particular, where artistic preferences are less important, while carefully examing potential issues such as less diverse ideas.

3. Theoretical Framework and Hypotheses

To understand GenAI's ability to tackle various creative tasks, we must first conceptualize creativity. The literature distinguishes between three dimensions of creativity. *Fluency* is the ability to generate many ideas or solutions to a problem. It reflects the quantity of generated ideas. *Flexibility* is the capacity to produce a variety of ideas or solutions, showing an ability to shift approaches or perspectives. And, *originality* is the ability to produce novel and unique ideas (Guilford 1967, Torrance 1968). In addition, the brainstorming literature often considers *idea quality* as a fourth dimension of creativity. We omit fluency as a performance metric, as comparing the number of ideas or the speed of idea generation between a computer and a human

will lead to the obvious result that the computer displays greater fluency, creating more ideas per unit of time. This leaves us with idea quality, flexibility, and originality as the dimensions of comparison between humans and AI.

The atomic unit of analysis in this comparison is an *idea*. In the context of innovation, we define an idea as a novel match between a solution and a need. As mentioned above, across three studies we will ask students as well as GenAI to come up with new product ideas targeted toward college students that can be sold for $50 or less. To illustrate our unit analysis of an *idea*, consider one of the student-generated ideas:

Convertible High-Heel Shoe: Many prefer high-heel shoes for dress-up occasions, yet walking in high heels for more than short distances is very challenging. Might we create a stylish high-heel shoe that easily adapts to a comfortable walking configuration, say by folding down or removing a heel portion of the shoe?

In this example, the need is the desire of some people to dress up and wear high-heeled shoes for some occasions while still walking comfortably. The proposed solution is to make the heel portion of the shoe so that it can be folded down or removed.

Idea generation, by either individuals or groups, is a process that creates a stream of ideas with varying quality levels. This stream can be the result of either human effort or the use of AI. Each of these ideas can be validated on a quality scale. Our quality scale is based on a purchase intent study. Kornish and Ulrich (2014) show that the best indicator of future value creation is the *average purchase intent expressed by a sample of consumers in the target market*. Furthermore, they show that no single individual, expert or novice, is particularly good at estimating value. Instead, a sample of expressed purchase intent from about 15 individuals in the target market is a reliable measure of idea quality.

Some ideas are likely to be brilliant (high-quality), some are horrible (low-quality), and most will be somewhere in between (medium-quality). We can think of this uncertain quality value as a random variable drawn from an underlying pay-off distribution (Weitzman 1979, Dahan and Mendelson 2001).

Recall that we chose to measure three dimentions of creativity associated with idea generation: quality, flexibility, and originality. Our first hypothesis relates to the first dimention: AI's ability to generate ideas comparable in their average quality to human-generated ideas. In other words, we focus on the mean of the underlying idea-quality distribution. We make two arguments for why GPT-4 would create ideas of higher average quality than humans. First, the training data for GPT-4 includes millions of product reviews revealing unmet user needs, social media posts of excited and frustrated customers alike, and marketing materials for countless products that have been launched more or less successfully in the past. Second, the literature reviewed in Section 2 has established that GPT-4 has tremendous creative capabilities in other domains such as music generation or story writing.

Hypothesis 1 (Idea quality): The average quality of AI-generated ideas is higher than the average quality of human-generated ideas.

Our second hypothesis relates to the second two dimensions: flexibility and originality. We first define these concepts in the context of generating ideas for new products and come up with appropriate measurement scales.

There exists a vast number of possible new product ideas that differ along many dimensions. We can think of ideas as positions in a highly dimensional space. OpenAI's GPT-4 models text as multi-dimensional embedding vectors in this space, where each dimension may represent a distinct attribute or feature of the text. Such vectors have hundreds of dimensions. Similar texts will often lie close to each other while different ones will be far apart. However, interpreting the distances and dimensions is often not straightforward given the high dimensionality.

To illustrate, consider a two-dimensional search space like the map of a territory. For example, consider the exploration of such a territory in the search for fishing spots in the ocean. The (x, y) coordinates capture the geographic locations of schools of fish. Each location has a pay-off corresponding to the amount of fish in the water. The goal of the fisherman is to find the location with the greatest fish density. In such a search process, local adjustments along a gradient of increasing fish density in the water via local search may increase the value of a fishing location. Yet, in rugged solution landscapes, i.e., ones that have multiple local optima, such local search is unlikely to yield the globally optimal solution.

Thereby, the ruggedness of the underlying solution landscape makes it impossible to arrive at the most valuable fishing location (idea) in the ocean (idea space) via local adjustments. Rather, a broad exploration is needed (see Sommer and Loch 2004). Without prior knowledge about the landscape, some new locations that are very different from past locations should be explored. This creates the classic trade-off between exploration and exploitation (March 1991).

With this as our backdrop, we provide two ways of operationalizing flexibility, *overlap* and the *total number of discoverable ideas*, and one way to operationalize originality, *idea novelty*. All three are important properties of a search process in general and of an ideation process in particular.

To explain overlap, let's return to our fishing example. To explore fishing locations in an ocean, the locations should be distinctively different from each other. Even in a rugged solution landscape, some spatial correlations in pay-offs between two adjacent coordinates are likely. In much the same way, in the world of innovation, we want our ideas to be distinct from each other. To determine how distinctly different an idea is relative to other ideas, we measure the cosine similarity of its embedding vector relative to the embedding vectors of the other ideas (following Cox et al. 2021 and Dell'Acqua et al. 2023). Section 8

provides alternative measures to this analytical choice. For a given pool of ideas produced by an idea-generation process, human or AI, we can thus randomly pull out two ideas and compute the angle between two associated embedding vectors. The Cosine of such angles will range from -1 to 1, with 1 indicating identical vectors and 0 indicating no similarity (orthogonal). While negative values are possible in principle, they rarely occur in practice as further discussed in study 2. By performing a pairwise comparison of all ideas and averaging their similarities, we can compute the **average pool similarity**. Next, we define two ideas as **overlapping** if their cosine similarity is above $\theta = 0.8$. That is, we count any new idea added to the pool as overlapping if its cosine similarity exceeds 0.8 compared to any of the existing ideas in the pool. Our first measure of flexibility is based on computing the distribution of pairwise cosine similarities and counting the frequency of overlaps. We discuss this and other assumptions in Section 8 and provide extensive robustness analyses including evaluating alternative model specifications.

Next, imagine a fisherman with no memory looking for fish at random locations. Every period, this fisherman sets out and fishes, yielding an estimate for the payoff of a specific location. How many unique fishing locations will be discovered this way? Early in the exploratory efforts, every fishing spot is an unexplored territory. Yet, as this process goes on, the likelihood of overlap increases, i.e., the fisherman is more likely to revisit a location previously tested. Given our definition of overlapping ideas (cosine similarity exceeding the θ=0.8 threshold), we can observe a stream of incoming ideas, one by one, and determine whether a new idea is unique relative to the pool of ideas created up to this point. Early on, just like in the fisherman's case, each idea is likely unique (non-overlapping with the ideas created so far). However, as the process progresses, the percentage of overlapping ideas will increase as the underlying search space gets exhausted. For a finite sequence of T ideas, we can evaluate the number of overlapping ideas, $N_{overlap,}$ and thus compute the number of unique ideas, $N_{unique}=T-N_{overlap}$. Definitions for how we operationalize this approach are shown in study 2.

In addition to utilizing idea overlap for computing the number of unique ideas in a finite stream of ideas, we can further estimate the total number of discoverable ideas in the search space, even if many were not part of the sequence of T ideas, i.e., the ideas have not (yet) been discovered. To do so, we use what in population ecology is known as a capture-recapture model, used to estimate the number of unique fishing locations based on how frequently a previously visited location is revisited by a fisherman with no memory. With such a model, we simply count the incidents of an idea overlapping with a past idea. The frequency of overlap and its increased occurrence rate over time allows for estimating the **number of ideas that can be discovered** (Kornish and Ulrich 2011). This provides us with our second measure of flexibility.

Next, consider originality. The search for ideas can yield ideas that are more or less novel. We measure idea novelty in the same way we measure idea quality – by directly asking potential customers for its novelty assessment and averaging this value. In summary, we evaluate flexibility by looking at idea overlap (which

can be converted into an estimate for the numbers of ideas that can be discovered) and evaluate originality by directly asking consumers to rate novelty.

How will a pool of AI-generated ideas compare to these human-generated ideas in terms of quality, flexibility and originality? By their very design, GPTs are autoregressive processes. They don't plan ahead but predict one word (or token) at a time based on a context window, including the prompt and the prior words created. Such a "one word at a time" process is unlikely to systematically and exhaustively explore an entire solution landscape. This lack of broad exploration will be further amplified in the presence of a system prompt that illustrates the concept of ideas by providing one or multiple ideas from the past (few-shot prompting) relative to the case in which no past ideas are provided (zero-shot prompting). This should limit both the flexibility and the originality of the creative process.These arguments, taken together with existing research in other domains showing less novelty for AI-generated content versus human-generated content (Doshi and Hauser, 2024), lead to the following two hypotheses:

Hypothesis 2a (Flexibility): The likelihood of two ideas overlapping is higher for a pool of AI-generated ideas than for a pool of human-generated ideas, resulting in fewer discoverable ideas.

Hypothesis 2b (Originality): The average novelty of AI-generated ideas is lower than that of human-generated ideas.

Our third hypothesis returns to the concept of idea quality. This time, however, we are not concerned about the average idea quality but instead focus on the quality of the best ideas. Rather than focusing on the quality of the single best idea (the extreme value, Dahan and Mendelson 2001), we focus on the 90th percentile of idea quality distribution, i.e., the top 10 percent of the ideas. We do so for two reasons.

The first reason is statistical estimation: for a single experiment like ours, there simply does not exist a test that allows us to make statistically significant statements for a single data point. Moving to the 90th percentile, we can compare the mean across larger groups of ideas (Section 8 presents our results for other percentiles).

There also exists a second, managerial reason. In many, if not most, practical settings, the assessment of idea quality is noisy, especially in the early stages of an innovation process when an idea is nothing but a title and a few words. For this reason innovation tournaments don't just advance a single idea to the next round, but a set of the *x* percent of the most promising ideas where *x* can vary widely, but typically ranges between 10 and 50 percent (Terwiesch and Ulrich 2009). We therefore state:

Hypothesis 3 (Top Decile): The quality of the 90th percentile AI-generated ideas is higher than that of the

90th percentile human-generated ideas.

4. Experimental setup

For our experiment, we utilize three different pools of ideas, namely student-generated ideas, GPT-4-generated ideas with zero-shot prompting and GPT-4-generated ideas with few-shot prompting. For the student pool, we rely on data collected in 2021 in a product design and innovation course at an elite university. In this course, 50 students participated in an innovation challenge to come up with ideas for a physical product marketed to college students for $50 or less (this price cap is imposed to limit the complexity of the projects in a one-semester course.). The challenge was organized in a traditional innovation tournament format (Terwiesch and Ulrich 2009, 2023), in which individuals first independently generate many ideas, which are then combined into a pool of several hundred ideas and subsequently evaluated by others in the group (i.e., "crowdsourced" evaluations). Thus, we have access to a large set of ideas generated by humans before AI tools became widely available to enhance ideation.

Speifically, we use a pool of independently aggregated human ideas by randomly selecting 200 entries, each comprising a descriptive title and a paragraph of text, from the student ideas generated in these challenges in 2021 (i.e., at a time prior to the widespread availability of ChatGPT and other LLMs). The set of 200 ideas constitutes our first pool and forms the baseline for comparison with the ideas generated using LLMs. We prompt Open AI's GPT-4 (more specifically, gpt-4-0314) with the same prompt we gave the students. No LLM yet acts entirely autonomously. Rather, they are tools used by humans to complete tasks. For this study, we aim for minimal prompt engineering, thus representing a novice user scenario. However, we acknowledge that many strategies could potentially improve LLM performance. For instance, Mihm and Schlapp (2019) show that providing feedback during ideation contests can further improve performance of human innovators and we expect this to hold for LLMs as well

For our first LLM-generated idea pool we use the system prompt to provide contextual information and subsequent user prompts to ask for ideas, ten at a time. The user prompt includes the additional request that the descriptions be 40-80 words, like the student sample.

System Prompt

"You are a creative entrepreneur looking to generate new product ideas. The product will target college students in the United States. It should be a physical good, not a service or software. I'd like a product that could be sold at a retail price of less than about USD 50. The ideas are just ideas. The product need not yet exist, nor may it necessarily be clearly feasible. Number all ideas and give them a name. The name and idea are separated by a colon."

User Prompt

"Please generate ten ideas as ten separate paragraphs. The idea should be expressed as a paragraph of 40-80 words."

The model used for all work covered in this paper is gpt-4-0314 with the "temperature" parameter at 0.7 to retain randomness and thus greater creativity. The temperature parameter controls the randomness of the output, with lower values leading to more deterministic output and higher values increasing variability. At the time of the experiment, the suggested default value for temperature was 0.7 to strike a balance between coherence and creativity, without possibly sampling highly unlikely tokens (i.e., semantic chunks used for representational efficiency) that lead to undesirable responses.

An obstacle to using GPT-4 for generating hundreds of ideas is its finite memory, typically limited to the number of tokens the underlying LLM can consider in generating its responses. Once the number of tokens in a session exceeds the model's limit, the LLM has no memory of the first ideas generated, and subsequent ideas can become increasingly redundant. The number of tokens in the version of GPT-4 we had access to was about 8,000, roughly 7,000 words or approximately 80 ideas (some tokens are used for the system and user prompt and idea titles).

To generate more than 80 ideas resulting from the limited context window, we asked GPT-4 to "compress" the previously generated ideas into shorter summaries. These summaries were then provided to the model before generating the next batch of ideas, ensuring that the model knows the previously generated ideas while remaining within the context limits. We used the below summarization prompt, followed by the original system prompt and generated summaries, and finally, a user prompt that explicitly asks for different ideas. This constitutes our second pool of comparison.

Summarization Prompt

"Aggressively compress the following ideas so that their original meaning remains but they are much shorter. You can use tags or keywords. : <Ideas generated so far> "

System Prompt

<Original System Prompt> + "Previously you generated the following ideas and should not repeat them: <Summaries> "

User Prompt

<Original User Prompt> + "Make sure they are different from the previous ideas."

For our second pool of LLM-generated ideas, we provide the LLM with examples (few-shot learning) of high-quality ideas generated by students. In particular, we appended our prompts to provide the LLM with six highly rated ideas from a separate student set that completed the same exercise and informed GPT-4 that these ideas had been well-received by students in our class. We used six examples due to context window limitations at the time of the experiment as well as drawing on previous experiments from in-context few-shot learning where too many examples can degrade performance (see Meincke and Carton 2024). This constitutes our third pool of comparison.

Good Ideas Prompt

<*Original System Prompt*> + ”Here are some well received ideas for inspiration: <*Good Ideas*>”

Overall, we generated 100 ideas using zero-shot prompting and another 100 using few-shot prompting. The resulting average word count for GPT-4 generated ideas is 69 and 71 for GPT-4 with provided with examples. The average description is 63 words long for student ideas. We compared the resulting few-shot prompted ideas to the examples provided to ensure that GPT-4 did not simply slightly modify the examples. The average pairwise cosine similarity between the six examples and the 100 generated ideas is 0.33 and the highest similarity between two ideas is 0.51. Thus, we have no reason to believe that GPT-4 repeated the provided ideas.

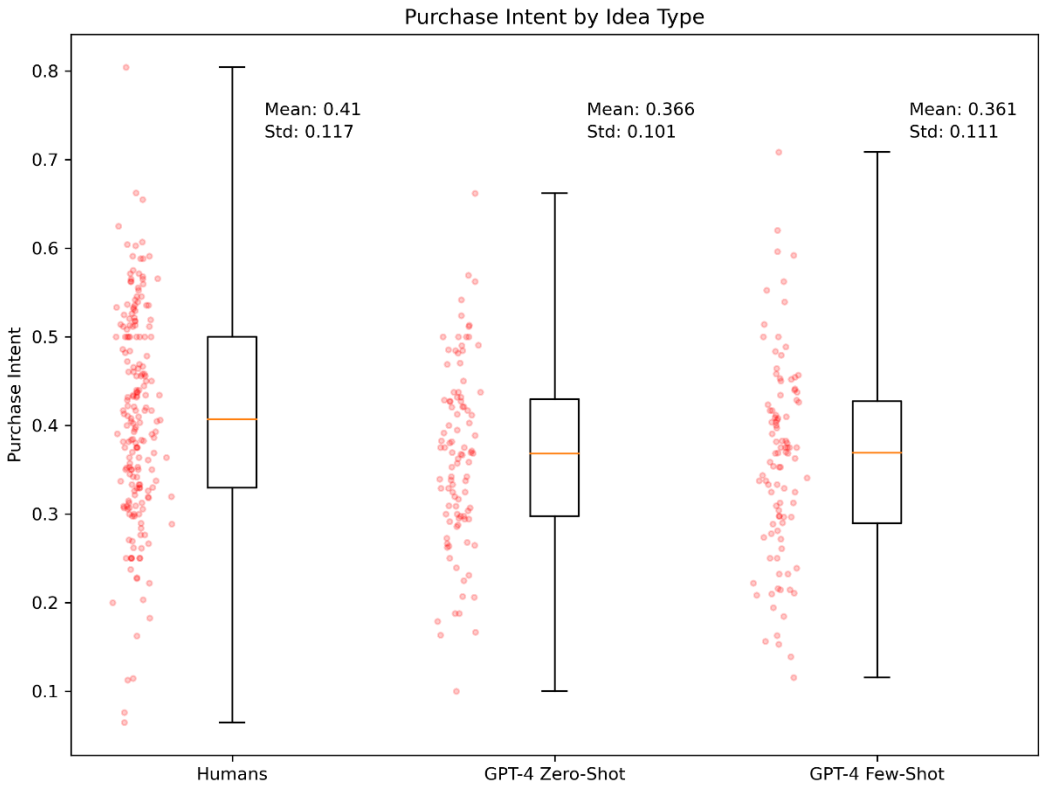
5. Study 1: comparing the quality of ideas generated by humans and AI

The Institutional Review Board (IRB) at the University of Pennsylvania approved the research described in this paper in May 2023, Protocol #853634. We used the online platform Prolific to recruit college-age indiviuals from the United States to evaluate all 400 ideas from the three pools (pool 1 with 200 ideas created by humans, pool 2 with 100 created by GPT-4 with zero-shot prompting, and pool 3 with 100 created by GPT-4 with few-shot prompting) via a purchase intent survey. We presented ideas in random order and randomized at the idea level, meaning that every survey participant could potentially see ideas from multiple sources. Each respondent evaluated an average of 40 ideas. On average, each idea was evaluated 20 times. In the summer of 2023, concerns surfaced that ChatGPT was being used to provide mTurk responses. This practice appears to have been limited to text generation tasks, not to multiple choice tasks like our five-box purchase-intent survey. Indeed, just answering the survey question directly requires less effort than trying to deploy ChatGPT to answer the question. We thus believe that our study participants were humans.

We asked respondents to express purchase intent using the standard “five-box” options: definitely would not purchase, probably would not purchase, might or might not purchase, probably would purchase, and

definitely would purchase. Jameson and Bass (1989) recommend weighting responses for the five possible responses as 0, 0.25, 0.50, 0.75, and 1.00 to develop a single measure of purchase probability, which we use as a measure of idea quality (other weightings are possible, as we discuss in Section 8). Figure 1 shows the full quality distribution of ideas generated by the three pools.

Figure 1 Distribution of idea quality for three sets of ideas



Notes. Purchase intent is the weighted average of the five-box response scale per Jameson and Bass (1989).

Figure 1 shows the quality (purchase probability) of ideas across the three pools. On average, GPT-4 generated ideas with greater purchase intent (46.4% with zero-shot prompting and 49.3% with few-shot prompting) than humans (40.4%). The standard deviation of the quality of ideas is comparable between the three pools. We formally test the impact of idea source on the perceived quality of product ideas via a linear mixed-effects model with purchase intent as the dependent variable. The model included two fixed-effects denoting source (humans are the baseline) and random intercepts and slopes for respondents and ideas. We

find significant differences in the perceived quality of ideas as a function of their source. Ideas generated by GPT-4 with no examples (zero-shot) were rated significantly higher than human-generated ideas (β = 0.059; 95% CI [0.031, 0.088]; t(246) = 4.06, p < 0.001) and ideas generated by GPT-4 provided with positive examples (few-shot) received even higher ratings (β = 0.089; 95% CI [0.060, 0.12]; t(223) = 5.93, p < 0.001). Purchase intent is weakly significantly different between the two pools of LLM-generated ideas (β = 0.03; 95% CI [-0.01, 0.06]; t(199) = 1.892, p = 0.06). These findings indicate that LLM-generated ideas are, on average, more likely to be purchased than human-generated ideas (for additional robustness tests, see Section 8).

6. Study 2: Diversity and Novelty of Ideas

Our second study focuses on how the fraction of overlapping ideas and the resulting estimated total number of ideas the process can generate (idea flexibility, hypothesis 2a) and the perceived novelty of the ideas as assessed by human raters (idea originality, hypothesis 2b) depend on the idea source.

6.1. Overlapping Ideas An idea-generation process creates a sequence of ideas in which each additional idea generated can be compared to the previously created ideas according to its similarity. For a pool of ideas, we can hence compute the average pairwise similarity of one idea compared to all other ideas and then compute the average overall similarity for the entire pool. We can also apply a threshold to pairwise idea similarity to identify at what point the ideas start to become more repetitive, i.e., when we are starting to exhaust the space of new ideas given a particular idea-generation process. A pool of ideas then might have a few overlapping ideas, which informs our second quantitative metric, the total number of ideas the process can generate.

To measure the diversity of the ideas, we calculate the cosine similarity of each idea relative to the rest of the set. We first calculate a vector of text embeddings for each idea. We follow the technical setup in Dell'Acqua et al. (2023) and use Google's Universal Sentence Encoder (USE) model for our idea embeddings, which is specifically optimized for semantic similarity between sentences. Table 1 shows the results.

In geometry, the cosine of the angle between vectors ranges from -1 to 1. However, when using Google USE, negative similarity is rarely encountered, since the overall text structure does not substantially differ between ideas. Ideas follow a similar pattern in terms of text length and style, often leading with the title before the idea description. In our test, a cosine similarity of 1 between two ideas thus indicates that they are very similar (their embedding vectors are aligned), whereas a cosine similarity of 0 implies orthogonal or unrelated ideas. We consider a new idea added to an idea pool to be unique if its pairwise cosine similarity compared to all previously added ideas is never greater than 0.8. Additional robustness checks using different thresholds and measures can be found in Section 8.

Table 1 Summary Statistics for Idea Overlap

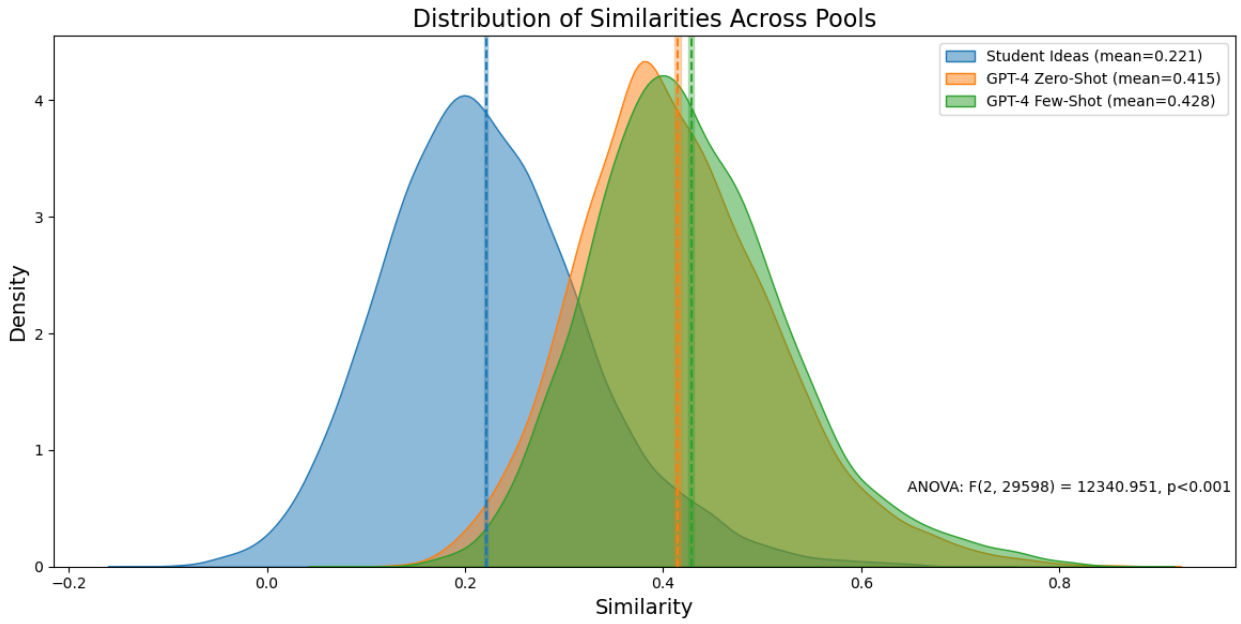
	Student Ideas	GPT-4 zero-shot	GPT-4 few-shot
N Ideas	200	100	100
Average cosine similarity of all ideas	0.221	0.415	0.428
Fraction of ideas in pool with cosine similarity >0.8	0	0.05	0.07

Notes. We compute the fraction as the number of ideas whose average pairwise similarity compared to all other ideas in the pool exceeds 0.8 divided by the total number of ideas in the pool.

For each pool, we compute the average pairwise similarity between all ideas. One-way ANOVA analyses show that the source has a significant effect on the cosine similarity between the three pools. The difference between all three groups is also significant ($\eta^2 = 0.455$, 95% CI [-0.210, -0.204], $F(2, 29598) = 12340.95$, $p < 0.001$). Considering only two groups, human ideas have a significantly smaller cosine similarity than GPT-4-generated ideas ($\eta^2 = 0.358$, 95% CI [-0.197, -0.190], $F(1, 24649) = 13715.82$, $p < 0.001$). Zero-shot GPT-4 ideas exhibit a significantly smaller cosine similarity than few-shot GPT-4 ideas ($\eta^2 = 0.004$, 95% CI [-0.018, -0.010], $F(1, 9898) = 44.24$, $p < 0.001$).

Because there is no overlap among human-generated ideas using cosine similarity, the fraction of ideas would be zero and the number of unique ideas infinitely large, in line with hypothesis 2a. A larger pool of student ideas will eventually contain overlapping ideas (see Kornish and Ulrich 2014 for estimates) but based on our assumptions for similarity, the student sample only contains unique ideas. We perform a binomial test to formally estimate the significance of the differences. We find that the fraction of similar human-generated ideas (95% CI for fraction [0.0, 0.0184]) is significantly smaller than that of the zero-shot GPT-4 ideas (RD = -0.05, 95% CI [-0.093, -0.007], $p < 0.001$) and few-shot GPT-4 ideas (RD = -0.07, 95% CI [-0.120, -0.020], $p < 0.001$), supporting hypothesis 2a. The difference between the two GPT-4 pools is not significantly different (RD = -0.02, 95% CI [-0.086, 0.046], $p = 0.56$). Our findings suggest that a greater number of distinct ideas generated comes from the human-ideation process, as opposed to GPT-4. We calculate the exact numbers in the next section.

Figure 2 **Distribution of cosine similarities across the three pools**



Notes. Density plot of cosine similarities comparing all three pools. The dotted line shows the mean and confidence interval of the estimate for a pool used for the ANOVA. The difference between all three groups is also significant ($\eta^2 = 0.455$, 95% CI [-0.210, -0.204], $F(2, 29598) = 12340.95$, $p < 0.001$).

6.2. Number of Discoverable Ideas Given the fraction of unique ideas, we can estimate the number of unique ideas that could be generated by each of our three processes (pools) – students, LLM (zero-shot), and LLM prompted with examples (few-shot) – using the method of Kornish and Ulrich (2011). This method, which uses the capture-recapture method to analyze the probability that the next idea in a sequence is unique, reportedly originates with Laplace (Cochran 1978), but has been adapted to wildlife ecology and other domains. For illustration, consider again fishing in a lake as a metaphor for the idea-generation process. Each idea is a catch, and the fish is released back into the lake. Sometimes, the same fish will be caught again. The more frequently an individual fish is re-caught, the smaller the estimate of the overall fish population. Thus, the probability that a fish has never been caught previously is a decreasing function of the number of ideas generated.

This probability decay is typically represented by an exponential function.

$$p(n) = e^{-an} \tag{1}$$

We define $p(n)$ as the probability that the next idea is unique given n ideas have been generated already. The expected number of unique ideas out of n generated, $u(n)$, is the integral under this curve.

$$u(n) = (1/a)(1 - e^{-an}) \quad (2)$$

This form of probability decay comes from a specific underlying process, with T unique ideas total (T fish in the pond), and each equally likely to be drawn. This assumption is commonly used in the Lincoln-Peterson method (Lincoln 1930), the standard model for estimating population size in the literature on wildlife ecology. The decay parameter and the total T are linked: $T = 1/a$. This model has only a single parameter, a, which is the inverse of the size of the opportunity space, i.e., an estimate of the total number of unique ideas that an unlimited number of comparable idea generators, each generating an enormous number of ideas, would generate.

Given a set of ideas generated and a count of the number of unique ideas in that set, the model can be used to calculate T, an estimate of the size of the opportunity space. Using the similarity threshold of 0.8 from the cosine similarity metric, we found that 5 of the 100 ideas generated by the LLM with zero-shot prompting were essentially similar to an idea already generated (fish recaptured), and that 7 of the 100 ideas generated via few-shot prompting were redundant. Thus, $u(100)$ is 0.95 in the first case, and $u(100)$ is 0.93 in the second case. This corresponds to an estimate of T of 966 ideas (zero-shot) and of 680 ideas (few-shot) respectively.

In our sample, human-generated ideas were all unique. Thus, as expected from our overlap calculations, and based on the estimates provided by the capture-recapture model, we find support for the second quantitative metric of hypothesis 2a. The number of unique ideas that can be discovered is lower for both pools of AI-generated ideas than for the human idea-generation process. In addition, prompting the LLM with examples seems to further reduce the estimated number of unique ideas available to the process. We perform additional robustness checks in Section 8.

6.3. Perceived Novelty Given that LLMs are designed to generate the statistically most plausible sequence of text based on their training data, perhaps they generate less novel ideas than humans. Novelty is not a goal expressed in the prompt used in this study for either humans or GPT-4 and is typically not a primary objective in commercial product development efforts. Still, to ensure that GPT-generated ideas are not merely lists of existing ideas, we investigate how the novelty of ideas varies between LLM-generated ideas and those generated by humans.

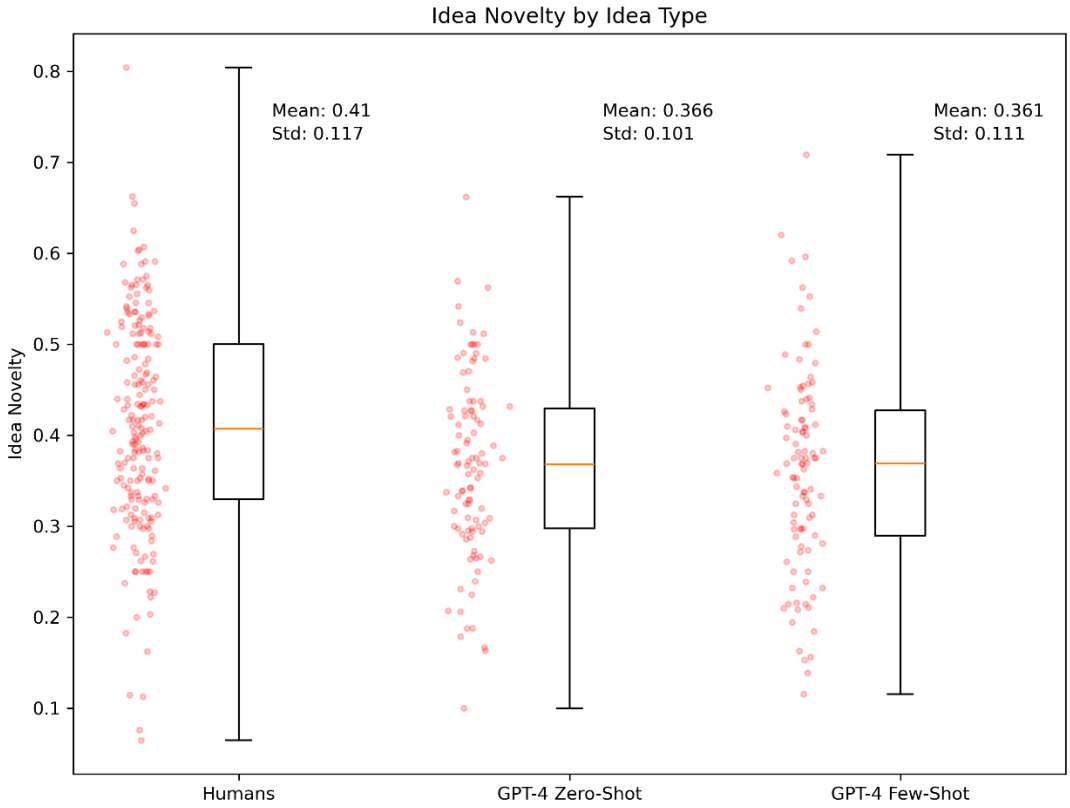
Based on Shibayama et al. (2021), we assessed novelty by asking responders on Prolific the question "Relative to other products you have seen, how novel do you consider the idea for this new product?" [0: Not at all novel, 0.25: Slightly novel, 0.5: Moderately novel, 0.75: Very novel, 1: Extremely novel]. The average novelty of human-generated ideas is 40.6% (SD: 0.117), which is greater than that of zero-shot GPT-4 (36.7%, SD: 0.101), and few-shot GPT-4 (36.1%, SD: 0.111; see Figure 3).

Similar to purchase intent, we estimate a linear mixed-effects model to investigate how the idea source (human ideas, zero-shot GPT-4 and few-shot GPT-4) affects the perceived novelty of product ideas. The model includes two fixed effects for denoting the source (humans are baseline), random intercepts and slopes for both respondents and ideas.

We find significant differences in perceived novelty between human and zero-shot GPT-4-generated ideas (β = -0.038; 95% CI [-0.066, -0.01]; t(269) = -2.67, p = 0.008) at the alpha = 0.05 threshold. Ideas generated by few-shot GPT-4 also receive significantly lower novelty ratings (β = -0.049; 95% CI [-0.078, -0.02]; t(268) = -3.4, p < 0.001) compared to human-generated ideas. These findings suggest that some LLM-generated ideas are perceived as less novel than human-generated ideas.

Perceived novelty is not significantly different between the two pools of LLM-generated ideas (β = -0.01; 95% CI [-0.039, 0.017]; t(195) = -0.757, p = 0.45). Of note, novelty does not appear to be significantly correlated with purchase intent. The correlation coefficient is slightly negative at -0.08 (95% CI [-0.176, 0.016], p=0.12). Additional robustness checks can be found in Section 8.

Figure 3 Distribution of novelty ratings for three samples of ideas



Notes. Novelty based on mTurk assessment per Kwon, Kim, and Lee (2009).

These findings support Hypothesis 2b: AI-generated ideas are, on average, less novel than human-generated ideas. Of note, the average novelty of all ideas, irrespective of source, lies between slightly and moderately novel. While human ideas are around 0.047 points more novel, there is little reason to believe that novelty alone, i.e., being the first to think of an idea, leads to a significant financial advantage. As Terwiesch and Ulrich (2010) and others have argued, the first-mover advantage is a myth. As such, from a commercial point of view, we don't believe that the slightly lower novelty outweighs the productivity and quality benefits of LLMs.

7. Study 3: What is the quality of the best idea(s)?

Table 2 summarizes the titles of the top 40 ideas (10%) in our pool, that is the top 40 out of the 400 ideas used.

Table 2 Top 10% Ideas by Purchase Intent

Title	Source	Purchase Intent	Novelty
Compact Printer	GPT-4 (Few-Shot)	0.76	0.55
Solar-Powered Gadget Charger	GPT-4 (Few-Shot)	0.75	0.44
QuickClean Mini Vacuum	GPT-4 (Zero-Shot)	0.75	0.30
Noise-Canceling Headphones	GPT-4 (Few-Shot)	0.72	0.18
StudyErgo Seat Cushion	GPT-4 (Zero-Shot)	0.72	0.39
Multifunctional Desk Organizer	GPT-4 (Few-Shot)	0.71	0.21
Reusable Silicone Food Storage Bags	GPT-4 (Few-Shot)	0.68	0.34
Portable Closet Organizer	GPT-4 (Few-Shot)	0.67	0.23
Dorm Room Chef [oven, microwave and toaster]*	GPT-4 (Few-Shot)	0.67	0.71
Collegiate Cookware	GPT-4 (Few-Shot)	0.67	0.45
Collapsible Laundry Basket	GPT-4 (Few-Shot)	0.65	0.21
On-the-Go Charging Pouch	GPT-4 (Few-Shot)	0.65	0.33
GreenEats Reusable Containers	GPT-4 (Zero-Shot)	0.65	0.21
HydrationStation [bottle with filter]*	GPT-4 (Zero-Shot)	0.64	0.19
Reusable Shopping Bag Set	GPT-4 (Few-Shot)	0.64	0.19
CollegeLife Collapsible Laundry Hamper	GPT-4 (Zero-Shot)	0.64	0.26
Adaptiflex [cord extension to fit big adapters]*	Student	0.64	0.44
SpaceSaver Hangers	GPT-4 (Zero-Shot)	0.64	0.33
Dorm Room Air Purifier	GPT-4 (Few-Shot)	0.63	0.29
Smart Power Strip	GPT-4 (Few-Shot)	0.63	0.22
CampusCharger Pro	GPT-4 (Zero-Shot)	0.63	0.31
Kitchen Safe Gloves	Student	0.62	0.31
Nightstand Nook [charging, cup holder]*	GPT-4 (Few-Shot)	0.62	0.43
Mini Steamer	GPT-4 (Few-Shot)	0.62	0.41
CollegeCare First Aid Kit	GPT-4 (Zero-Shot)	0.62	0.26
StudySoundProof [soundproofing panels]*	GPT-4 (Zero-Shot)	0.62	0.57
FreshAir Fan	GPT-4 (Zero-Shot)	0.62	0.29
StudyBuddy Lamp [portable, usb charging]*	GPT-4 (Zero-Shot)	0.62	0.43
Bluetooth Signal Merger [share music]*	Student	0.62	0.41

Adjustable Laptop Riser	GPT-4 (Few-Shot)	0.62	0.21
EcoCharge [solar powered charger]*	GPT-4 (Zero-Shot)	0.62	0.43
Smartphone Projector	Student	0.62	0.57
Grocery Helper [hook to carry multiple bags]*	Student	0.62	0.53
FitnessOnTheGo [portable gym equipment]*	GPT-4 (Zero-Shot)	0.62	0.42
Multipurpose Fitness Equipment	GPT-4 (Few-Shot)	0.62	0.37
CollegeCooker	GPT-4 (Zero-Shot)	0.61	0.50
Multifunctional Wall Organizer	GPT-4 (Few-Shot)	0.61	0.31
DormDoc Portable Scanner	GPT-4 (Zero-Shot)	0.61	0.49
Mobile Charging Station Organizer	GPT-4 (Few-Shot)	0.61	0.26
StudyMate Planner	GPT-4 (Few-Shot)	0.61	0.22
DormChef Kitchen Set	GPT-4 (Zero-Shot)	0.61	0.33
LaundryBuddy [laundry basket]*	GPT-4 (Zero-Shot)	0.61	0.30

Notes. The asterisk (*) denotes ideas where the text in square brackets [] is not part of the original title and was added to clarify the idea.

Among the top 40 ideas (top decile) 35 (87.5%) were generated by GPT-4 (see Table 3). In other words, for every human idea in the top 10% we count 7 ideas generated by GPT-4. A Chi-Square Test of independence, with the null hypothesis of equal representation of all sources among the top ideas (20, 10 and 10) rejected the null hypothesis ($x^2 = 26.39$, $p < 0.001$, $df = 2$), thus confirming hypothesis 3.

Table 3 Best Ideas Across Pools

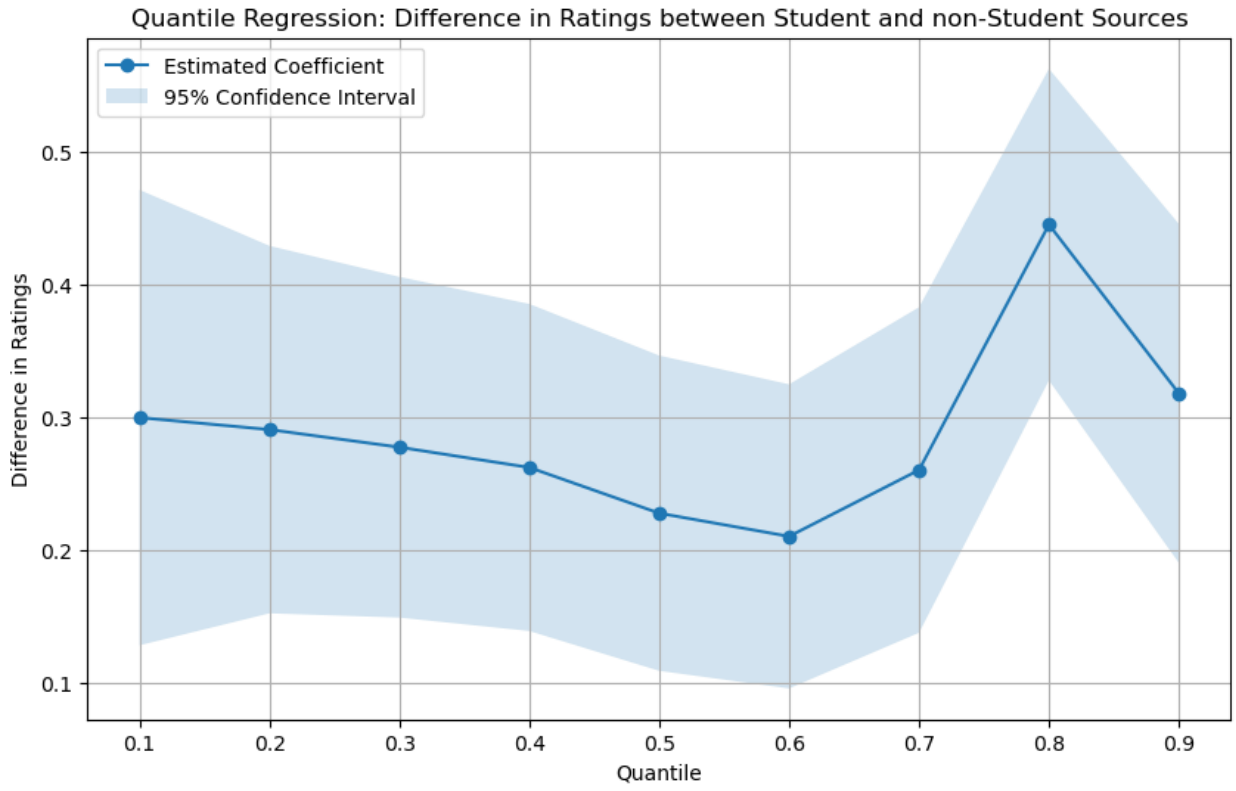
	Student Ideas	GPT-4 zero-shot	GPT-4 few-shot
N Ideas	200	100	100
Average Quality of Top Decile	0.62	0.64	0.66
Average Novelty of Top Decile	0.45	0.35	0.33
Fraction of the top decile of pooled ideas from this source	5/40	15/40	20/40

To better understand how the full distribution of idea qualities is affected by the idea source, we use quantile regression analysis. Quantile regression (Koenker and Hallock 2001) extends traditional regression by computing the relationship between explanatory variables (idea source) and the response variable (idea quality) for different percentiles of the data. As mentioned above, in innovation, the quality of the best ideas is generally more important than the average quality. That is, we prefer a few exceptional ideas to a lot of

mediocre ones. Using quantile regression, we can examine the tails of the distribution instead of the mean, allowing us to test whether GPT-4 excels at generating high-quality ideas only for specific percentiles or whether the effect holds across the entire distribution.

Our analysis follows Girotra et al. (2010). We use the average idea quality ratings as the dependent variable, and our explanatory variable is a binary variable indicating whether the idea is human-generated (baseline level) or AI-generated (GPT-4 zero-shot and GPT-4 few-shot prompting). Figure 4 shows the results. For all percentiles, GPT-4 ideas consistently outperform student ideas. The effect is especially pronounced for the upper tail of the distribution (80% and above), where GPT-4 has the strongest advantage. This implies that not only does GPT-4 generate better ideas on average, but it is also especially adept at producing top-tier ideas compared to students.

Figure 4 **Estimated Difference in Idea Quality Ratings between AI-generated Ideas and Human-generated ideas (baseline), for Different Percentiles**



8. Discussion and Limitations

In this section, we discuss conceptual limitations of our work, limitations related to our research design, as well as data analysis and the robustness of our analysis to a set of alternative specifications and assumptions.

Our findings indicate that GPT-4 produces higher-quality ideas that are more likely to be purchased than humans, though they are perceived as less novel. AI significantly outperforms human creativity in generating top-tier ideas, with GPT-4 ideas being seven times more likely to rank in the top 10%. Given AI's advantage in both quality and productivity, our findings have profound implications for the field of innovation management. For instance, AI can serve as a first step in brainstorming sessions, allowing organizations to rapidly explore a wide variety of ideas with minimal cost and time investment. Human ideators can also provide AI with their own interesting ideas and refine them with the help of AI. Another important implication lies in the potential shift of focus from idea generation to idea evaluation. If LLMs can reliably produce numerous high-quality ideas at very low cost companies might allocate more resources toward assessing and refining those ideas instead of ideating from scratch. This shift could lead to the development of new tools and frameworks specifically designed to help organizations sort, rank, and prioritize AI-generated ideas, further streamlining the innovation process.

However, while the results show that GPT-4 outperforms human creativity in terms of producing top-tier ideas, the reduced novelty and increased similarity among AI-generated ideas point to a limitation. This suggests that a human in the loop is still important to drive the ideation direction and ensure that ideas are as novel as possible. Future research could explore ways to mitigate this issue by enhancing LLMs' ability to generate more diverse and creative solutions through techniques such as fine-tuning.

Investigating whether LLMs can evaluate ideas with the same rigor as human evaluators would help to further improve the ideation process. It would allow an LLM to get immediate feedback on its creations, leaving humans to focus on implementation and strategy.

8.1 Conceptual and Research Design Limitations Conceptually, our prompting approach (i.e., a simple prompt) is not optimized for creativity or novelty. It also follows a single ideator setup instead of approaches such as hybrid brainstorming that lead to more and better ideas (Girotra et al. 2010). A model given more specific instructions on how to ideate effectively might thus perform even better. Different prompting techniques such as Chain-of-Thought (CoT), which asks the model to reason through a problem in multiple steps instead of directly providing an answer (Wei et al. 2023), might also improve performance. Furthermore, providing the model with hundreds of good ideas, either via many-shot learning or fine-tuning could also provide enhanced performance. This suggests that we likely underestimate the true power of AI-based idea generation.

Second, it is possible that professional product innovators would generate better ideas than our students. However, this has not been the experience of the paper's authors, who have taught many academic courses and worked in many product development settings. Many students who participated in the innovation contests have gone on to be product innovators, sometimes based on ideas from the course tournament.

Nevertheless, we have not produced evidence that GPT-4 is better than the best product innovators working today. However, we believe that we can claim conservatively that GPT-4 is better than many human product innovators working today and probably better than average. Thus, at a very minimum, an LLM could elevate the least capable humans to a better-than-average level of performance.

Third, GPT might be a great salesperson. As such, it is possible that the writing style ("pitch") convinces the customers rather than the idea itself. Prior work in other domains suggests that the text generated by LLMs is not distinguishable from that generated by humans (Brown et al. 2020), though recent work has developed sophisticated measures to detect LLM-generated text (Mitchell et al. 2023, Kobak et al. 2024, Venkatraman et al. 2024). For example, Kobak et al. (2024) provide intuitions that could be used to identify LLM-generated text, such as words that are not commonly used by the majority of English speakers like "delve." However, it is unlikely that these characteristics were known to our survey participants at the time of our experiment in May and June 2023, and that any particular idea generated by GPT-4 could easily be distinguished from those generated by our students. Future research could use LLMs to present human-generated ideas in a way that more closely mimics the presentation style of LLM-generated ideas, ensuring that the quality of the idea is not confounded by its presentation style.

Fourth, our study is set in the widely understood domain of consumer products for the college students market that cost less than $50. Presumably, there exists a lot of commentary and data about such products in the training data used by the GPT class of language models. As such, it is unclear whether our results would generalize to more specialized domains, such as surgical instruments. Organizations looking for opportunities in these specialized domains should fine-tune language models with their own proprietary data to achieve comparable or better performance.

Fifth, innovation often benefits from collaboration and is not solely focused on one ideator generating many ideas. Liu et al. (2018) show that collaborating with other innovators improves the creative process by enabling the transfer of critical skills and knowledge, particularly when those collaborations involve highly skilled innovators. Future work should investigate whether this can be applied to human and LLM interaction, and whether an LLM could help a novice human innovator become better.

8.2 Robustness There are different ways to analyze the data. Here, we provide additional robustness checks that investigate the validity of our results under various specifications.

8.2.1 Study 1 To measure purchase intent, it is possible to use other convex weighting schemes. Ulrich and Eppinger (2007) weigh 'definitely would purchase' as 0.4 and 'probably would purchase' as 0.2 with all other responses rated as 0. When using this alternative set of weights, we find the same significant differences between pools.

As a robustness test for our primary purchase-intent analysis using a linear mixed-effects model, we also

conduct a simpler linear regression focusing on the average perceived quality of product ideas across different sources. This model aggregates individual ratings at the idea level, removing the random effects to capture the overall influence of the source on rating averages. The results confirm our previous findings and show that ideas from GPT-4 (zero-shot) are rated higher than human ones by an average of 0.256 points (95% CI [0.15, 0.37]; $t = 4.602$, $p < 0.001$), and ideas from GPT-4 (few-shot) are rated higher by an average of 0.358 points (95% CI [0.25, 0.47]; $t = 6.435$, $p < 0.001$).

In addition, we estimate a cumulative link mixed model (CLMM) to treat the ratings outcome as a factor. We find significant differences in the perceived quality, measured as purchase intent of product ideas, between sources. Ideas generated by GPT-4 (zero shot) receive a significantly greater average rating ($\beta = 0.395$; 95% CI [0.215, 0.575]; $z = 4.31$, $p < 0.001$). Similarly, ideas generated by GPT-4 (few-shot) receive even higher ratings ($\beta = 0.581$; 95% CI [0.400, 0.762]; $z = 6.30$, $p < 0.001$) compared to human-generated ideas. These findings suggest that LLM-generated ideas are perceived as more likely to be purchased than human-generated ideas, with the highest perceived quality attributed to few-shot GPT-4-generated ideas.

8.2.2 Study 2 Our chosen threshold of $\theta = 0.8$ has been established through experimentation by comparing ideas as pairs of two and their respective similarity scores. However, our findings are robust to other values such as 0.7 (25 and 37 overlapping ideas for zero-shot and few-shot GPT-4 respectively) and 0.75 (16 and 23 overlapping ideas). At $\theta = 0.85$, the zero-shot GPT-4 pool only features two overlapping ideas, whereas the few-shot pool features one. Because these are extreme values that approach zero, we used 0.8 as our main threshold. We compute the pairwise similarity for an idea compared to all other ideas in the pool and calculate the average. Mean pairwise similarity is a common measure in ideation (Siangliulue et al. 2016, Cox et al. 2021) and similar text-mining tasks (Doshi and Hauser 2024) but it is not without issues, as it lacks sensitivity to highly clustered ideas. As an additional specification, we consider the per-pool collective diversity of all ideas by following the work in Cox et al. (2021) and construct a minimum spanning tree (MST) which spans all points (ideas) in space with the smallest total distance along the edges. In 2D space, an MST would be the tree that contains all points with the shortest overall length of edges. We compute the mean of all edge distances as a measure of how distributed ideas are in the high-dimensional space. The spanning tree is constructed in high-dimensional space (512 dimensions), its edge weights summed up and divided by the number of edges, resulting in a range from 0 (not diverse at all) to 1 (very diverse). Based on this measure, the student idea pool is the most diverse (0.53), GPT-4 zero-shot is the second most diverse (0.33) and GPT-4 few-shot is the least diverse (0.3) pool.

Similar to purchase intent, we also conduct a simpler linear regression focusing on the average perceived novelty of product ideas across different sources. This model aggregates individual ratings at the idea level, removing the random effects to capture the overall influence of the source on rating averages. We find that

ideas from GPT-4 (zero-shot) are significantly less novel than human ones (β = -0.177; 95% CI [-0.286, -0.069]; t = -3.22, p < 0.0014). Ideas from GPT-4 (few-shot) are rated as significantly less novel than human ones (β = -0.197; 95% CI [-0.305, -0.089]; t = -3.58, p < 0.001). This simpler analysis reinforces that human ideas are more novel than AI-generated ones, even when using zero-shot prompting.

In addition, we estimate a cumulative link mixed model (CLMM) to treat the ratings outcome as a factor. We find significant differences in the perceived novelty. Ideas generated by GPT-4 (zero shot) receive a significantly lower average rating (β = -0.306; 95% CI [-0.514, -0.1]; z = -2.89, p < 0.01). Similarly, ideas generated by GPT-4 (few-shot) receive even lower ratings (β = -0.39; 95% CI [-0.6, -0.18]; z = -3.66, p < 0.001) compared to human-generated ideas. These findings suggest that LLM-generated ideas are perceived as less novel than human-generated ideas, with the lowest perceived novelty attributed to few-shot GPT-4-generated ideas.

8.2.3 Study 3 In this study, we present our results for the 90th percentile of all aggregated ideas. Table 4 shows that using other percentiles yields similar results.

Table 4 Top 5 and 15 Percent of Ideas Pool Distributions

	Student Ideas	GPT-4 zero-shot	GPT-4 few-shot
Average Quality of Top 5%	0.64	0.67	0.68
Fraction of the top 5% of pooled ideas from this source	1/20	6/20	14/20
Average Quality of Top 15%	0.60	0.62	0.64
Fraction of the top 15% of pooled ideas from this source	11/60	22/60	27/60

9. Summary

GenAI has demonstrated remarkable advancements in creating coherent and fluent text, equaling or surpassing human performance in various academic and professional domains. In this study, we explored the ideation capabilities of OpenAI's GPT-4, a state-of-the-art large language model, in comparison to the ideation abilities of university students when generating ideas for new products targeted toward college

students at a price point of $50 or less. Specifically, we make three main contributions to the literature of innovation and the role of AI.

First, GPT-4 produces high-quality ideas that are perceived as more likely to be purchased than human-generated ideas. Second, consumers perceive AI-generated ideas as less novel. Third, when considering the quality of the best ideas, AI outperforms human creativity significantly. To put these findings in context, innovation favors a few great ideas over a large number of solid ideas and our results show that AI-generated ideas are seven times more likely to be among the top 10% of ideas considered for our experiment compared to human ideas. Despite the reduction in novelty, the overall AI advantage thus remains substantial.

The fact that GPT-4 is very efficient at generating ideas does not require a formal research study. Two hundred ideas can be generated by one human interacting with GPT-4 in about 15 minutes. A human working alone can generate about five ideas in 15 minutes and humans working in groups do even worse (Girotra et al., 2010). In short, the productivity race between humans and GPT-4 is not even close. However, as we show in this article, the enormous potential of LLMs in ideation does not result only from their ability to quickly and inexpensively generate ideas, but in the remarkable quality of their output.

Importantly, these ideas can be produced at a fraction of the cost it would take humans, generating hundreds of high-quality ideas. This previously unimaginable productivity in generating ideas may substantially reduce the importance of the idea-generation phase of innovation and shift managerial focus to the idea-evaluation phase. Can an LLM also take on the task of idea evaluation? From our viewpoint, this is a fascinating question for future research.

References

Bellaiche L, Shahi R, Turpin MH, Ragnhildstveit A, Sprockett S, Barr N, Christensen A, Seli P (2023) Humans versus AI: whether and why we prefer human-created compared to AI-created artwork. *Cogn. Research* 8(1):42.

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, et al. (2020) Language Models are Few-Shot Learners. (July 22) http://arxiv.org/abs/2005.14165.

Chao RO, Kavadias S (2008) A Theoretical Framework for Managing the New Product Development Portfolio: When and How to Use Strategic Buckets. *Management Science* 54(5):907–921.

Cochran WG (1978) Laplace's Ratio Estimator. David HA, ed. *Contributions to Survey Sampling and Applied Statistics*. (Academic Press), 3–10.

Connolly T, Jessup LM, Valacich JS (1990) Effects of Anonymity and Evaluative Tone on Idea Generation in Computer-Mediated Groups. *Management Science* 36(6):689–703.

Cox SR, Wang Y, Abdul A, Von Der Weth C, Y. Lim B (2021) Directed Diversity: Leveraging Language Embedding Distances for Collective Creativity in Crowd Ideation. *Proceedings of the 2021 CHI*

Conference on Human Factors in Computing Systems. (ACM, Yokohama Japan), 1–35.
Dahan E, Mendelson H (2001) An Extreme-Value Model of Concept Testing. *Management Science* 47(1):102–116.
Dell'Acqua F, McFowland III 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. (September 15) https://papers.ssrn.com/abstract=4573321.
Doshi AR, Hauser OP (2024) Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Sci. Adv.* 10(28):eadn5290.
Girotra K, Terwiesch C, Ulrich KT (2010) Idea Generation and the Quality of the Best Idea. *Management Science* 56(4):591–605.
Goldenberg J, Mazursky D, Solomon S (1999) Creative Sparks. *Science* 285(5433):1495–1496.
Guilford JP (1967) Creativity: Yesterday, Today and Tomorrow. *Journal of Creative Behavior* 1(1):3–14.
Haase J, Hanel PHP (2023) Artificial muses: Generative artificial intelligence chatbots have risen to human-level creativity. *Journal of Creativity* 33(3):100066.
Hitsuwari J, Ueda Y, Yun W, Nomura M (2023) Does human–AI collaboration lead to more creative art? Aesthetic evaluation of human-made and AI-generated haiku poetry. *Computers in Human Behavior* 139:107502.
Hubert KF, Awa KN, Zabelina DL (2024) The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Sci Rep* 14(1):3440.
Huchzermeier A, Loch CH (2001) Project Management Under Risk: Using the Real Options Approach to Evaluate Flexibility in R…D. *Management Science* 47(1):85–101.
Jamieson LF, Bass FM (1989) Adjusting Stated Intention Measures to Predict Trial Purchase of New Products: A Comparison of Models and Methods. *Journal of Marketing Research* 26(3):336–345.
Jia N, Luo X, Fang Z, Liao C (2024) When and How Artificial Intelligence Augments Employee Creativity. *AMJ* 67(1):5–32.
Kobak D, González-Márquez R, Horvát EÁ, Lause J (2024) Delving into ChatGPT usage in academic writing through excess vocabulary. (July 3) http://arxiv.org/abs/2406.07016.
Koenker R, Hallock KF (2001) Quantile Regression. *Journal of Economic Perspectives* 15(4):143–156.
Koivisto M, Grassini S (2023) Best humans still outperform artificial intelligence in a creative divergent thinking task. *Sci Rep* 13(1):13601.
Kornish LJ, Ulrich KT (2011) Opportunity Spaces in Innovation: Empirical Analysis of Large Samples of Ideas. *Management Science* 57(1):107–128.
Kornish LJ, Ulrich KT (2014) The Importance of the Raw Idea in Innovation: Testing the Sow's Ear

Hypothesis. *Journal of Marketing Research* 51(1):14–26.
Lincoln FC (1930) *Calculating waterfowl abundance on the basis of banding returns* (U.S. Dept. of Agriculture, Washington, D.C.).
Liu H, Mihm J, Sosa ME (2018) Where Do Stars Come From? The Role of Star vs. Nonstar Collaborators in Creative Settings. *Organization Science* 29(6):1149–1169.
Loch CH, Terwiesch C, Thomke S (2001) Parallel and Sequential Testing of Design Alternatives. *Management Science* 47(5):663–678.
March JG (1991) Exploration and Exploitation in Organizational Learning. *Organization Science* 2(1):71–87.
Mihm J, Schlapp J (2019) Sourcing Innovation: On Feedback in Contests. *Management Science* 65(2):559–576.
Mitchell E, Lee Y, Khazatsky A, Manning CD, Finn C (2023) DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. (July 23) http://arxiv.org/abs/2301.11305.
Meincke L, Carton A (2024) Beyond Multiple Choice: The Role of Large Language Models in Educational Simulations. (May 26) https://papers.ssrn.com/abstract=4873537.
OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. (2024) GPT-4 Technical Report. (March 4) http://arxiv.org/abs/2303.08774.
Osborn AF (1953) *Applied imagination* (Scribner'S, Oxford, England).
Rashidi HH, Fennell BD, Albahra S, Hu B, Gorbett T (2023) The ChatGPT conundrum: Human-generated scientific manuscripts misidentified as AI creations by AI text detection tool. *Journal of Pathology Informatics* 14:100342.
Shank DB, Stefanik C, Stuhlsatz C, Kacirek K, Belfi AM (2023) AI composer bias: Listeners like music less when they think it was composed by an AI. *J Exp Psychol Appl* 29(3):676–692.
Shibayama S, Yin D, Matsumoto K (2021) Measuring novelty in science with word embedding Muscio A, ed. *PLoS ONE* 16(7):e0254034.
Si H, Kavadias S, Loch CH (2022) Managing Innovation Portfolios: From Project Selection to Portfolio Design. (March 6) https://papers.ssrn.com/abstract=4050940.
Siangliulue P, Chan J, Dow SP, Gajos KZ (2016) IdeaHound: Improving Large-scale Collaborative Ideation with Crowd-Powered Real-time Semantic Modeling. *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. (ACM, Tokyo Japan), 609–624.
Sommer SC, Loch CH (2004) Selectionism and Learning in Projects with Complexity and Unforeseeable Uncertainty. *Management Science* 50(10):1334–1347.
Sutton RI, Hargadon A (1996) Brainstorming Groups in Context: Effectiveness in a Product Design Firm. *Administrative Science Quarterly* 41(4):685.

Terwiesch C (2023) Let's cast a critical eye over business ideas from ChatGPT. *Financial Times* (March 12) https://www.ft.com/content/591ad272-6419-4f2c-9935-caff1d670f08.
Terwiesch C, Ulrich K (2023) The innovation tournament handbook: a step-by-step guide to finding exceptional solutions to any challenge (Wharton School Press, Philadelphia, PA).
Terwiesch C, Ulrich KT (2009) Innovation tournaments: creating and selecting exceptional opportunities (Harvard Business Press, Boston, Mass).
Terwiesch C, Xu Y (2008) Innovation Contests, Open Innovation, and Multiagent Problem Solving. *Management Science* 54(9):1529–1543.
Torrance EP (1968) A Longitudinal Examination of the Fourth Grade Slump in Creativity. *Gifted Child Quarterly* 12(4):195–199.
Ulrich K, Eppinger S (2007) *Product Design and Development* (McGraw-Hill Education)
Venkatraman S, Uchendu A, Lee D (2024) GPT-who: An Information Density-based Machine-Generated Text Detector. (April 3) http://arxiv.org/abs/2310.06202.
Wang H, Zou J, Mozer M, Goyal A, Lamb A, Zhang L, Su WJ, et al. (2024) Can AI Be as Creative as Humans? (January 25) http://arxiv.org/abs/2401.01623.
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. (January 10) http://arxiv.org/abs/2201.11903.
Weitzman ML (1979) Optimal Search for the Best Alternative. *Econometrica* 47(3):641.
Zhou E, Lee D (2024) Generative artificial intelligence, human creativity, and art Harding M, ed. *PNAS Nexus* 3(3):pgae052.
Zlatkov D, Ens J, Pasquier P (2023) Searching for Human Bias Against AI-Composed Music. Artificial Intelligence in Music, Sound, Art and Design: 12th International Conference, EvoMUSART 2023, Held as Part of EvoStar 2023, Brno, Czech Republic, April 12–14, 2023, Proceedings. (Springer-Verlag, Berlin, Heidelberg), 308–323.

Appendix A. Quantile Regression Results

The regression model considered quantiles 0.1 to 0.9 with a 0.1 step. For each quantile, it estimated *MeanRating ~ SourceAI*. *SourceAI* is a dummy variable that indicates whether the idea source was a student (*SourceAI*=0) or GPT-4 (*SourceAI*=1). A positive value for *SourceAI* indicates that ideas by GPT-4 performed better than human ideas. Negative values indicate the opposite.

Table A.1. Quantile Regression Results for quantiles 0.1 to 0.9

Quantile	Intercept	Source AI	Conf. Int. Low	Conf. Int. High
0.1	1	0.3*	0.128399	0.471601
0.2	1.230768	0.290958*	0.152504	0.429411
0.3	1.388859	0.277678*	0.149304	0.406051
0.4	1.549946	0.262418*	0.139183	0.385653
0.5	1.666666	0.227943*	0.10915	0.346736
0.6	1.789528	0.210472*	0.095888	0.325057
0.7	1.882355	0.260502*	0.137852	0.383152
0.8	1.954555	0.445445*	0.328038	0.562851
0.9	2.181822	0.318235*	0.190531	0.445939

Notes. *($p < 0.1\%$).

Appendix B. Supplementary Regression Tables

	Purchase Intent		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	0.40	0.38 – 0.43	<0.001
Source [Zero-Shot]	0.06	0.03 – 0.09	<0.001
Source [Few-Shot]	0.09	0.06 – 0.12	<0.001
Random Effects			
σ^2	0.07		
$\tau_{00\ IdeaID}$	0.01		
$\tau_{00\ RespondentID}$	0.02		
$\tau_{11\ IdeaID.SourceZero\text{-}Shot}$	0.01		

$\tau_{11\ IdeaID.SourceFew\text{-}Shot}$	0.03
$\tau_{11\ RespondentID.SourceZero\text{-}Shot}$	0.01
$\tau_{11\ RespondentID.SourceFew\text{-}Shot}$	0.01
ρ_{01}	-0.64
	-0.97
	-0.06
	-0.28
ICC	0.28
$N_{\ RespondentID}$	200
$N_{\ IdeaID}$	400
Observations	7982
Marginal R^2 / Conditional R^2	0.014 / 0.290

Purchase Intent Alternative Weights

Predictors	*Estimates*	*CI*	*p*
(Intercept)	0.08	0.07 – 0.08	<0.001
Source [Zero-Shot]	0.02	0.01 – 0.03	0.001
Source [Few-Shot]	0.03	0.02 – 0.04	<0.001
Random Effects			
σ^2	0.01		
$\tau_{00\ IdeaID}$	0.00		
$\tau_{00\ RespondentID}$	0.00		
$\tau_{11\ IdeaID.SourceZero\text{-}Shot}$	0.00		
$\tau_{11\ IdeaID.SourceFew\text{-}Shot}$	0.00		
$\tau_{11\ RespondentID.SourceZero\text{-}Shot}$	0.00		
$\tau_{11\ RespondentID.SourceFew\text{-}Shot}$	0.00		
ρ_{01}	0.04		
	0.48		
	-0.00		
	-0.16		
ICC	0.21		

$N_{RespondentID}$	200
N_{IdeaID}	400
Observations	7982
Marginal R^2 / Conditional R^2	0.009 / 0.215

Purchase Intent (Simple)

Predictors	*Estimates*	*CI*	*p*
(Intercept)	1.62	1.55 – 1.68	<0.001
Source [Zero-Shot]	0.26	0.15 – 0.37	<0.001
Source [Few-Shot]	0.36	0.25 – 0.47	<0.001
Observations	400		
R^2 / R^2 adjusted	0.108 / 0.104		

Purchase Intent (no weights, zero-shot baseline)

Predictors	*Estimates*	*CI*	*p*
(Intercept)	1.85	1.73 – 1.98	<0.001
Source [Student]	-0.24	-0.35 – -0.12	<0.001
Source [Few-Shot]	0.12	-0.00 – 0.24	0.058
Random Effects			
σ^2	1.18		
$\tau_{00\ IdeaID}$	0.12		
$\tau_{00\ RespondentID}$	0.42		
$\tau_{11\ IdeaID.SourceStudent}$	0.33		
$\tau_{11\ IdeaID.SourceFew\text{-}Shot}$	0.12		
$\tau_{11\ RespondentID.SourceStudent}$	0.13		
$\tau_{11\ RespondentID.SourceFew\text{-}Shot}$	0.01		
ρ_{01}	-0.77		
	-0.36		
	-0.50		
	-0.99		

ICC	0.31
$N_{RespondentID}$	200
N_{IdeaID}	400
Observations	7982
Marginal R^2 / Conditional R^2	0.014 / 0.322

Purchase Intent (ordered logistic regression)

Predictors	*Odds Ratios*	*CI*	*p*
0|1	0.25	0.21 – 0.29	<0.001
1|2	1.06	0.90 – 1.26	0.484
2|3	3.01	2.53 – 3.57	<0.001
3|4	19.07	15.84 – 22.97	<0.001
Source [Zero-Shot]	1.48	1.24 – 1.78	<0.001
Source [Few-Shot]	1.79	1.49 – 2.14	<0.001
Random Effects			
σ^2	3.29		
$\tau_{00\ IdeaID}$	0.39		
$\tau_{00\ RespondentID}$	0.92		
ICC	0.28		
$N_{RespondentID}$	200		
N_{IdeaID}	400		
Observations	7982		
Marginal R^2 / Conditional R^2	0.014 / 0.294		

Novelty

Predictors	*Estimates*	*CI*	*p*
(Intercept)	0.41	0.39 – 0.43	<0.001
Source [Zero-Shot]	-0.04	-0.07 – -0.01	0.008
Source [Few-Shot]	-0.05	-0.08 – -0.02	0.001
Random Effects			

σ^2	0.05
$\tau_{00\ IdeaID}$	0.01
$\tau_{00\ RespondentID}$	0.01
$\tau_{11\ IdeaID.SourceZero\text{-}Shot}$	0.02
$\tau_{11\ IdeaID.SourceFew\text{-}Shot}$	0.03
$\tau_{11\ RespondentID.SourceZero\text{-}Shot}$	0.01
$\tau_{11\ RespondentID.SourceFew\text{-}Shot}$	0.01
ρ_{01}	-0.87
	-0.99
	0.14
	0.06
$N_{\ RespondentID}$	201
$N_{\ IdeaID}$	400
Observations	8023
Marginal R^2 / Conditional R^2	0.009 / NA

Novelty (Simple)

Predictors	*Estimates*	*CI*	*p*
(Intercept)	1.64	1.58 – 1.70	<0.001
Source [Zero-Shot]	-0.18	-0.29 – -0.07	0.001
Source [Few-Shot]	-0.20	-0.31 – -0.09	<0.001
Observations	400		
R^2 / R^2 adjusted	0.042 / 0.037		

Novelty (no weights, zero-shot baseline)

Predictors	*Estimates*	*CI*	*p*
(Intercept)	1.48	1.37 – 1.59	<0.001
Source [Student]	0.15	0.05 – 0.26	0.004
Source [Few-Shot]	-0.04	-0.16 – 0.07	0.493
Random Effects			

σ^2	0.90
$\tau_{00\ IdeaID}$	0.11
$\tau_{00\ RespondentID}$	0.35
$\tau_{11\ IdeaID.SourceStudent}$	0.33
$\tau_{11\ IdeaID.SourceFew\text{-}Shot}$	0.44
$\tau_{11\ RespondentID.SourceStudent}$	0.03
$\tau_{11\ RespondentID.SourceFew\text{-}Shot}$	0.04
ρ_{01}	-0.71
	-0.98
	-1.00
	-0.23
$N_{\ RespondentID}$	201
$N_{\ IdeaID}$	400
Observations	8023
Marginal R^2 / Conditional R^2	0.009 / NA

Novelty (zero-shot baseline)

Predictors	*Estimates*	*CI*	*p*
(Intercept)	0.37	0.34 – 0.40	<0.001
Source [Student]	0.04	0.01 – 0.07	0.008
Source [Few-Shot]	-0.01	-0.04 – 0.02	0.449
Random Effects			
σ^2	0.05		
$\tau_{00\ IdeaID}$	0.01		
$\tau_{00\ RespondentID}$	0.02		
$\tau_{11\ IdeaID.SourceStudent}$	0.02		
$\tau_{11\ IdeaID.SourceFew\text{-}Shot}$	0.03		
$\tau_{11\ RespondentID.SourceStudent}$	0.01		
$\tau_{11\ RespondentID.SourceFew\text{-}Shot}$	0.00		
ρ_{01}	-0.74		

	-1.00
	-0.69
	-0.95
ICC	0.35
$N_{RespondentID}$	201
N_{IdeaID}	400
Observations	8023
Marginal R^2 / Conditional R^2	0.006 / 0.356

Novelty (ordered logistic regression)

Predictors	*Odds Ratios*	*CI*	*p*
0|1	0.16	0.13 – 0.19	<0.001
1|2	0.87	0.72 – 1.04	0.118
2|3	4.51	3.76 – 5.43	<0.001
3|4	29.60	24.09 – 36.37	<0.001
Source [Zero-Shot]	0.74	0.60 – 0.91	0.004
Source [Few-Shot]	0.68	0.55 – 0.84	<0.001
Random Effects			
σ^2	3.29		
$\tau_{00\ IdeaID}$	0.57		
$\tau_{00\ RespondentID}$	0.91		
ICC	0.31		
$N_{RespondentID}$	201		
N_{IdeaID}	400		
Observations	8023		
Marginal R^2 / Conditional R^2	0.006 / 0.315		