Working Paper

Mack Institute for Technological Innovation – The Wharton School, University of Pennsylvania
Cornell Tech and Johnson College of Business – Cornell University

Ideas are Dimes a Dozen: Large Language Models for Idea Generation in Innovation

Karan Girotra, Lennart Meincke, Christian Terwiesch, and Karl T. Ulrich[1]

July 10, 2023

Abstract

Large language models (LLMs) such as OpenAI's GPT series have shown remarkable capabilities in generating fluent and coherent text in various domains. We compare the ideation capabilities of ChatGPT-4, a chatbot based on a state-of-the-art LLM, with those of students at an elite university. ChatGPT-4 can generate ideas much faster and cheaper than students, the ideas are on average of higher quality (as measured by purchase-intent surveys) and exhibit higher variance in quality. More important, the vast majority of the best ideas in the pooled sample are generated by ChatGPT and not by the students. Providing ChatGPT with a few examples of highly-rated ideas further increases its performance. We discuss the implications of these findings for the management of innovation.

Keywords: innovation, idea generation, creativity, creative problem solving, LLM, large-scale language models, AI, artificial intelligence, ChatGPT

[1] Girotra: Cornell Tech, 2 West Loop Rd, New York, NY, 10044, girotra@cornell.edu | Meincke, Terwiesch, Ulrich: The Wharton School, 500 Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104, lennart@sas.upenn.edu, terwiesch@wharton.upenn.edu, ulrich@wharton.upenn.edu

Electronic copy available at: https://ssrn.com/abstract=4526071

1 Introduction

Generative artificial intelligence has made remarkable advances in creating life-like images and coherent, fluent text. Open AI's ChatGPT chatbot, based on the GPT series of large language models (LLM) can equal or surpass human performance in academic examinations and tests for professional certifications (OpenAI, 2023). Github Co-Pilot based on the same LLMs can help with writing, commenting, and debugging code. Other models can provide valuable professional advice in fields like 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 (i.e., *hallucinations*). The models are optimized to generate the most statistically likely sequences of words with an injection of randomness. They are not designed to exercise any judgment on the veracity or feasibility of the output. Further, the underlying optimization algorithms provide no performance guarantees and their output can thus be of inconsistent quality. Hallucinations and inconsistency are critical flaws that limit the use of LLM-based solutions to low-stakes settings or in conjunction with expensive human supervision.

In what applications can we leverage artificial intelligence that is brilliant in many ways yet cannot be trusted to produce reliably accurate results? One possibility is to turn their weaknesses – hallucinations and inconsistent quality – into a strength (Terwiesch, 2023).

In most management settings, we expect to make use of each unit of work produced. As such, consistency is prized and is, therefore, the focus of contemporary performance management. (See, for example, the *Six Sigma* methodology.) Erratic and inconsistent behavior is to be eliminated. For example, 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. In creative tasks, given that only one or a few ideas will be pursued, only a few extremely positive outcomes matter. Similarly, an ideator that generates 30 ideas is likelier to have one brilliant idea than an ideator that generates just 10. Overall, in creative problem-solving, variability in quality, and productivity, as reflected in the number of ideas generated, are more valuable than consistency (Girotra et al., 2010).

To achieve high variability in quality and high productivity, most research on ideation and brainstorming recommends enhancing performance by generating many ideas while postponing evaluation or judgment of ideas (Girotra et al., 2010). This is hard for human ideators to do, but LLMs are designed to do exactly this— quickly generate many somewhat plausible solutions without exercising much judgment. Further, the hallucinations and inconsistent behavior of LLMs increase the variability in quality, which, on average, improves the quality of the best ideas. For ideation, an LLM's lack of judgment and inconsistency could be prized features, not bugs.

Thus, we hypothesize that LLMs will be excellent ideators. The purpose of this paper is to test this hypothesis by evaluating the performance of LLMs in generating new ideas.

Electronic copy available at: https://ssrn.com/abstract=4526071

Specifically, we compare three pools of ideas for new consumer products. The first pool was created by students at an elite university enrolled in a course on product design prior to the availability of LLMs. The second pool of ideas was generated by OpenAI's ChatGPT-4 with the same prompt as that given to the students. The third pool of ideas was generated by prompting ChatGPT-4 with the task as well as with a sample of highly rated ideas to enable some in-context learning (i.e., *few-shot prompting*).

We address three questions. First, how productive is ChatGPT-4? That is, how much time and effort is required to generate ideas and how many can reasonably be generated compared to human efforts?

Second, what is the quality distribution of the ideas generated? We are particularly interested in the extreme values – the quality of the best ideas in the three pools. We measure the quality of the ideas using the standard market research technique of eliciting consumer purchase intent in a survey. Given an estimate of the quality of each idea, we can then compare the distributional characteristics of the quality of the three pools of ideas.

Third, given the performance of ChatGPT-4 in generating new product ideas, how can LLMs be used effectively in practice and what are the implications for the management of innovation?

2 Approach

We have over 20 years of experience teaching product design and innovation courses at Wharton, Cornell Tech, and INSEAD. We have used similar innovation challenges dozens of times with thousands of students. Most of our courses embody the 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 available to enhance ideation.

We randomly selected 200 ideas from the pool of ideas generated in our class in 2021 (i.e., at a time prior to the widespread availability of ChatGPT and other LLMs). These ideas comprise a descriptive title and a paragraph of text. They were all generated in response to the challenge of creating a new physical product for the college student market that would be likely to retail for less than USD 50. (This price cap is imposed to limit the complexity of the projects in a one-semester course.) Here is an example of a submitted idea:

> **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?*

The set of 200 ideas forms the baseline for comparison with the ideas generated using LLMs. The average description is 63 words long, with a standard deviation of 34.

We use Open AI's GPT-4 API access to prompt ChatGPT-4 with essentially the same prompt we gave the students. No LLM yet acts fully autonomously. Rather they are tools used by humans to complete tasks. Still, for the purpose of this study, we aim for minimal prompt engineering, thus representing a novice user scenario.

Electronic copy available at: https://ssrn.com/abstract=4526071

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 are 40-80 words, similar to 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 induce randomness, and thus greater creativity.

An obstacle to using ChatGPT-4 for generating 100s of ideas is its finite memory, typically limited to the number of tokens (i.e., semantic chunks used for representational efficiency) 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 ChatGPT-4 that we had access to is about 8000, which is roughly 7000 words or approximately 80 ideas (some tokens are used for the system and user prompt and for idea titles).

To generate more than about 80 ideas while wrestling with the context limit, we asked GPT-4 to “compress” the previously generated ideas into shorter summaries. These summaries were then provided to the model prior to generating the next batch of ideas, ensuring that the model knows the previously generated ideas while remaining within the context limits. To generate ideas beyond the token limit, 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.

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.”

Electronic copy available at: https://ssrn.com/abstract=4526071

General-purpose LLMs may be used as is or may be fine-tuned with examples. We generated a second batch of ideas after providing the LLM with examples of high-quality ideas generated by students. In particular, we appended our prompts to provide the LLM with seven highly-rated ideas from a separate student set that did the same exercise and informed ChatGPT-4 that these ideas had been well-received. We used seven examples to keep the overall contribution to the context window moderate as well as drawing on previous experience from in-context few-shot learning.

Good Ideas Prompt

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

Overall, we generated 100 ideas without providing examples of good ideas and another 100 after providing access to examples of good ideas.

Prior work in other domains suggests that the text generated by LLMs is not distinguishable from that generated by humans (Brown et al., 2020). While we do not test this question in this study, our impression is that any particular idea generated by ChatGPT can not easily be distinguished from those generated by our students.

3 Do LLMs Enhance Productivity in Generating Ideas?

The answer to this question is straightforward. ChatGPT-4 is very efficient at generating ideas. This question does not require much precision to answer. Two hundred ideas can be generated by one human interacting with ChatGPT-4 in about 15 minutes. A human working alone can generate about five ideas in 15 minutes (Girotra et al., 2010). Humans working in groups do even worse. In short, the productivity race between humans and ChatGPT is not even close.

Still, the old saying that *ideas are a dime a dozen* is perhaps a tad optimistic. A professional working with ChatGPT-4 can generate ideas at a rate of about 800 ideas per hour. At a cost of USD 500 per hour of human effort, a figure representing an estimate of the fully loaded cost of a skilled professional, ideas are generated at a cost of about USD 0.63 each, or USD 7.50 (75 dimes) per dozen. At the time we used ChatGPT-4, the API fee for 800 ideas was about USD 20. For that same USD 500 per hour, a human working alone, without assistance from an LLM, only generates 20 ideas at a cost of roughly USD 25 each, hardly a dime a dozen. For the focused idea generation task itself, a human using ChatGPT-4 is thus about 40 times more productive than a human working alone.

In prior work, (Kornish and Ulrich, 2011) found that a typical new-product innovation domain contains thousands of unique ideas, ranging from about 1300 ideas for narrow challenges (e.g., use of technology in the classroom) to 3000 for more open-ended challenges (e.g., new consumer products). These numbers are large enough that a human working alone or in a small group is unlikely to identify most of them. However, LLMs are so productive that a human working with an LLM might reasonably fully articulate nearly every idea in an opportunity space. That is, it may now be possible to identify essentially every idea that a very large group of individuals working in parallel might identify after working for a long time, say, days or weeks. Prior work (Kornish and Ulrich 2014, Girotra et al. 2010) showed that the idea generation process in humans is essentially stationary, so ideas 2901 - 3000 exhibit the same quality distribution as ideas 1-100.

Electronic copy available at: https://ssrn.com/abstract=4526071

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.

4 What is the Quality Distribution of the Ideas Generated using LLMs?

A “stochastic parrot” can generate ideas, and LLMs do so shockingly productively. But we don’t care about quantity alone. More typically, the objective of idea generation is to generate at least a few truly exceptionally good ideas. In most innovation settings, we’d rather have 10 great ideas and 90 terrible ideas than 100 ideas of average quality.

We, therefore, care about the quality distribution of the ideas, and in particular, the quality of the best few ideas in a sample. Of course, we might as well also measure the mean and standard deviation of the three sets of ideas, and we do so. Two useful measures of the extreme values are: What is the average quality of the ideas in the top decile of each of the three samples? Which sources provided the ideas comprising the top 10 percent of the ideas in the pooled sample?

Measuring Idea Quality

Of course, what we want to know in most innovation settings is which idea has the highest expected future economic value given the uncertainty in how the ideas are developed and in the exogenous factors. This rationale is explored thoroughly in (Kornish and Ulrich, 2014) in the development of the *VIDE model*. Value (V) is a function of the idea itself (I), the development of that idea (D), and the exogenous factors (E). This value is not directly observable. To measure it we would need to develop and launch all ideas under all future states of the world. In very limited settings, we can estimate financial value, as done in (Kornish and Ulrich, 2014). That study showed that the best single indicator of future value creation is the *average purchase intent expressed by a sample of consumers in the target market*. Furthermore, (Kornish and Ulrich, 2014) showed that no single individual, expert or novice, is particularly good at estimating value. Rather, a sample of expressed purchase intent from about 15 individuals in the target market is a reliable measure of idea quality.

After obtaining the required IRB approvals, we used mTurk to evaluate all 400 ideas (200 created by humans, 100 created by ChatGPT without examples and 100 with training examples). The panel comprised college-age individuals in the United States. Ideas were presented in random order. Each respondent evaluated an average of 40 ideas. On average, each idea was evaluated 20 times[2].

Respondents were asked 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. Of course, many other weightings are possible. We report results using the Jameson and Bass weights, but the results are robust to other convex weighting schemes.

[2] In Summer 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. Thus, we believe that we were indeed surveying humans.

Electronic copy available at: https://ssrn.com/abstract=4526071

Results

The full quality distribution of ideas generated by the three pools is shown in Figure 1.

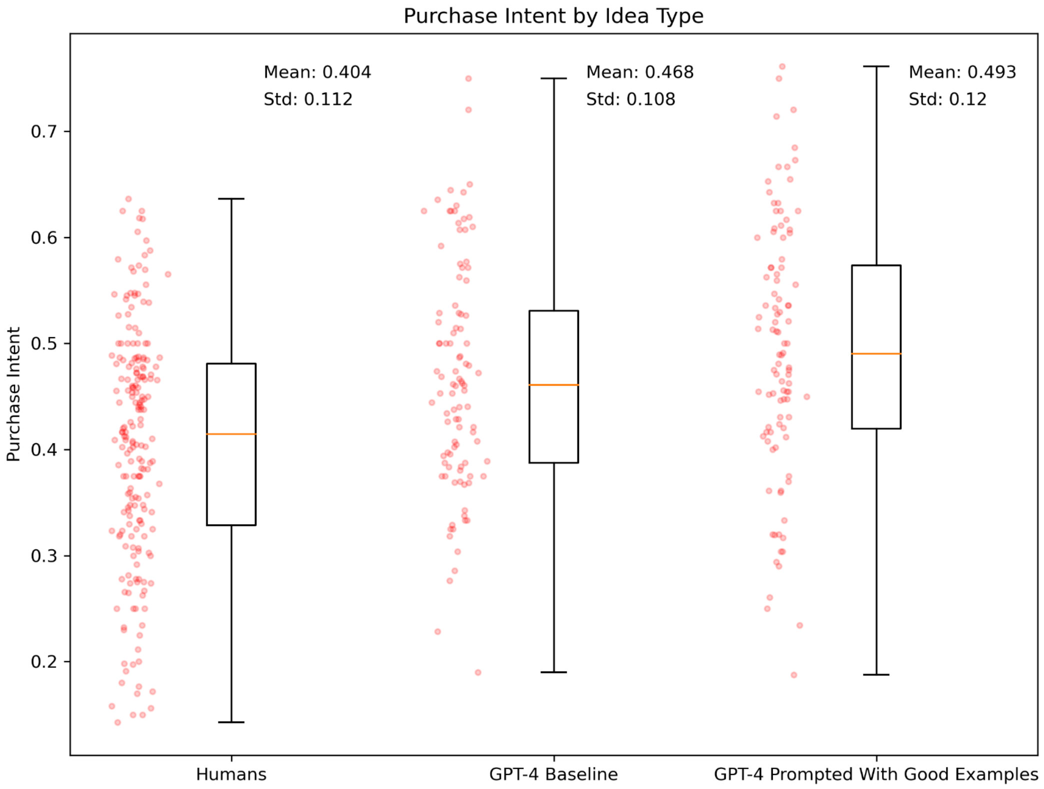

Figure 1 - Distribution of idea quality for three sets of ideas. Purchase intent is the weighted average of the five-box response scale per Jameson and Bass (1989).

The average quality of ideas generated by ChatGPT is higher than the average quality of ideas generated by humans, as measured by purchase intent. The average purchase probability of a human-generated idea is 40.4%, that of vanilla GPT-4 is 46.8%, and that of GPT-4 seeded with good ideas is 49.3%. The difference in average quality between humans and ChatGPT is statistically significant (p<0.001), but the difference between the two GPT models is not statistically significant (p=0.11). See Table 1.

The standard deviation of the quality of ideas is comparable, with ChatGPT trained with examples having the highest standard deviation.

Electronic copy available at: https://ssrn.com/abstract=4526071

Table 1 - Summary Statistics

	Human Generated Ideas	ChatGPT-4	ChatGPT-4 trained with examples
N Ideas	200	100	100
Average Length of Description	63 words	69 words	71 words
Average Quality	0.404	0.468	0.493
Standard Deviation of Quality	0.112	0.108	0.120
Best Idea	0.64	0.70	0.75
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
P-value (Is the average quality different?)		vs. humans <0.001	vs. humans <0.001 vs. baseline LLM 0.11

Most interesting are the differences in the quality of the best ideas. Chat-GPT generated the best-rated idea in our sample, with an 11% higher purchase probability than the best human idea. The average quality of the top decile in each of the three pools also follows the same pattern as average quality— seeded Chat-GPT ≻ ChatGPT ≻ Humans. Finally, most striking are the differences in each treatment's contribution to the top decile of all ideas we generated. Overall, we have 400 ideas, with an equal number generated by Chat-Gpt and humans. *In the top 40 ideas (top decile) a full 35 (87.5%) are those generated by Chat-Gpt. In other words, in a head-to-head match most of the winners come from ChatGPT.*

Titles of the top 40 ideas in our pool are reported in Table A1.

Novelty

Given that LLMs are designed to generate approximately the statistically most plausible sequence of text based on their training data, perhaps they generate less-novel ideas. Novelty is not a goal expressed in the prompt for either humans or Chat-GPT. It is typically not a primary objective in commercial product development efforts, nor does it have commercial value in itself. Still, we are

Electronic copy available at: https://ssrn.com/abstract=4526071

curious about how the novelty of ideas varies between LLM-generated ideas and those generated by humans.

We adopt the survey instrument of Shibayama, Yin, and Matsumoto (2021) to assess the novelty of the ideas. mTurk respondents answered this question:

Relative to other products you have seen, how novel do you consider the idea for this new product?

1. Not at all novel
2. Slightly novel
3. Moderately novel
4. Very novel
5. Extremely novel

We weigh these responses 0, 0.25, 0.50, 0.75, and 1.00 to produce a novelty score for each idea. By this measure, the mean novelty of the human-generated ideas is higher than that of the LLM-generated ideas (p<0.001). The mean novelty of the two different pools of LLM-generated ideas is not statistically significantly different from each other. (Figure 2)

Novelty does not appear to be significantly correlated with purchase intent. The correlation coefficient is slightly negative at -0.08 (p=0.12).

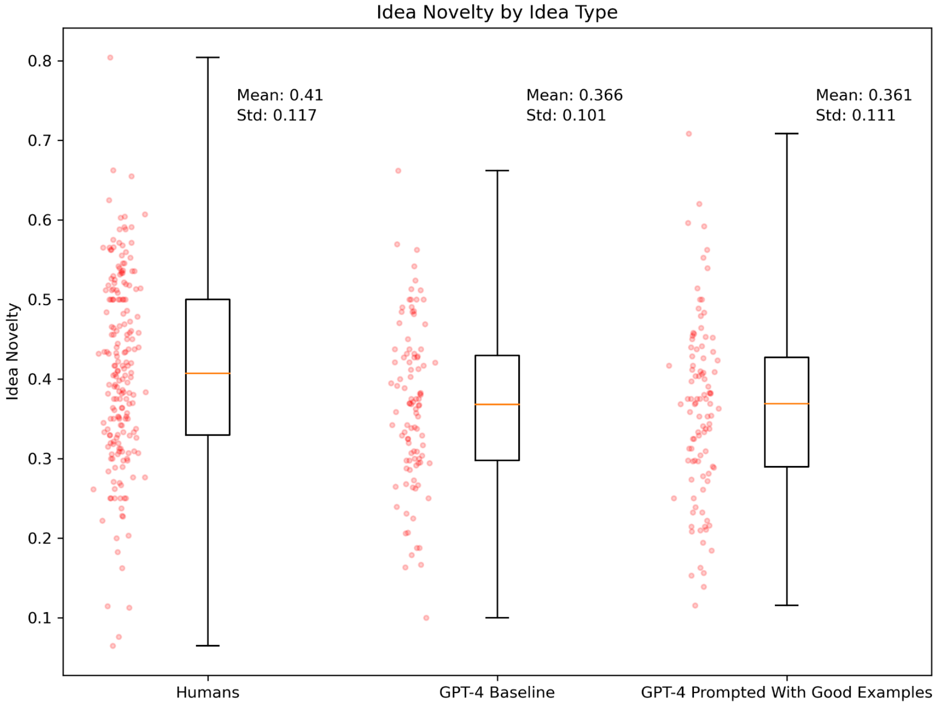

Figure 2 - Distribution of novelty ratings for three samples of ideas. Novelty based on mTurk assessment per Kwon, Kim, and Lee (2009).

Electronic copy available at: https://ssrn.com/abstract=4526071

We note here that the average novelty of all ideas, irrespective of source, lies between slightly and moderately novel. While human ideas are a bit more novel, there is little reason to believe that novelty – being the first to think of an idea – leads to a significant financial advantage in domains associated with off-the-shelf technology, low entry barriers, and limited intellectual property protection. As such, from a commercial point of view, we don't believe novelty provides sufficient advantage, if any, to overcome the productivity and quality benefits of the LLMs. Further, recall that novelty was not an explicit objective for any of our ideation schemes. In settings where novelty is the goal, it should be part of the prompts.

5 Limitations

Student Subjects

It is possible that professional product innovators would generate better ideas than our students. However, that is not our intuition having worked in many product development settings. Many students in this course have gone on to be product innovators, sometimes based on ideas from the course tournament. We have not produced evidence that ChatGPT is better than the best human product innovators working today. However, we believe that we can claim conservatively that ChatGPT 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.

Domain

Our results are set in a common widely understood domain, for consumer products likely selling at a price less than USD 50. Presumably, there is a lot of commentary and data around these domains in the training data used by the GPT class of language models. As such, It is possible that in more specialized domains, say surgical instruments, our results will no longer hold with the current class of models. That said, to us, if this is true, this is likely driven simply by the paucity of training data. An organization looking for opportunities in these specialized domains should presumably be able to fine-tune language models with their own proprietary data and achieve comparable or better performance.

Misbehavior

Most language models do not provide any performance guarantees and it is possible they can generate offensive, illegal, or inappropriate ideas. Ideators using models for ideation should exercise caution. Of course, the same caution is warranted with human idea generators.

Similarity

For most innovation settings, the goal is to thoroughly explore the landscape of possibilities. Doing so enhances confidence that the most reasonable opportunities have been unearthed and considered. To this extent, we prefer a process that generates 200 diverse ideas to one that generates 200 highly similar ideas. Our analysis does not speak to the similarity or variability in the content of ideas. This remains an open question for further study.

Electronic copy available at: https://ssrn.com/abstract=4526071

6 Concluding Remarks

In this study, we showed that the LLM technology in the form of ChatGPT4, a technology available for just a few months at the time of our experiments, is already significantly better at generating new product ideas than motivated, trained engineering and business students at a highly selective university.

Our results examined ideation productivity and quality separately. In each match-up, ChatGPT came out ahead. Combined, the effects of much higher productivity and the higher quality of the best ideas will likely completely trounce human ideators. The order of magnitude advantage in productivity itself is nearly insurmountable, and the higher quality of the best ideas further adds to the advantage of the LLM.

We can now put these tools in the hands of any innovator at extremely low cost. This suggests that the critical task in innovation practice may shift from idea generation to idea evaluation and selection, a task for which LLMs do not yet appear to be particularly well suited.

It is striking that conventional wisdom prior to 2022 was that AI tools would likely be most useful in rote tasks and that creative work would likely remain the domain of humans. In some ways, the opposite is true of LLMs. The tools are not perfectly reliable oracles providing information, but their lack of judgment leads to extreme productivity and high variance in idea quality resulting, at least in one setting, to creativity greater than that of the average human.

Acknowledgments and Funding Sources

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

References

OpenAI. 2023. GPT-4 Technical Report. https://cdn.openai.com/papers/gpt-4.pdf

Brown TB, et al. 2020. Language models are few-shot learners. arXiv:2005.14165.

Girotra K, Terwiesch C, Ulrich KT. 2010. Idea generation and the quality of the best idea. Manage Sci 56:591–605.

Jamieson LF, Bass FM. 1989. Adjusting stated intention measures to predict trial purchase of new products: A comparison of models and methods. J Mark Res 26:336–345.

Kornish LJ, Ulrich KT. 2014. The importance of the raw idea in innovation: Testing the sow's ear hypothesis. J Mark Res 51:14–26.

Kornish LJ, Ulrich KT. 2011. Opportunity spaces in innovation: Empirical analysis of large samples of ideas. Manage Sci 57:107–128.

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

Electronic copy available at: https://ssrn.com/abstract=4526071

Terwiesch C, Ulrich K. 2023. The Innovation Tournament Handbook: A Step-by-Step Guide to Finding Exceptional Solutions to Any Challenge (University of Pennsylvania Press).

Terwiesch C. 2023. Let’s cast a critical eye over business ideas from ChatGPT. Financ Times March 12.

Shibayama S, Yin D, Matsumoto K. 2021. Measuring novelty in science with word embedding. PLOS ONE July.

Electronic copy available at: https://ssrn.com/abstract=4526071

Appendix

Table A1 Top 10% Ideas (By Purchase Intent)

Title	Source	Purchase Intent	Novelty
Compact Printer	GPT-4 (Examples)	0.76	0.55
Solar-Powered Gadget Charger	GPT-4 (Examples)	0.75	0.44
QuickClean Mini Vacuum	GPT-4 (Base)	0.75	0.30
Noise-Canceling Headphones	GPT-4 (Examples)	0.72	0.18
StudyErgo Seat Cushion	GPT-4 (Base)	0.72	0.39
Multifunctional Desk Organizer	GPT-4 (Examples)	0.71	0.21
Reusable Silicone Food Storage Bags	GPT-4 (Examples)	0.68	0.34
Portable Closet Organizer	GPT-4 (Examples)	0.67	0.23
Dorm Room Chef [oven, microwave and toaster]*	GPT-4 (Examples)	0.67	0.71
Collegiate Cookware	GPT-4 (Examples)	0.67	0.45
Collapsible Laundry Basket	GPT-4 (Examples)	0.65	0.21
On-the-Go Charging Pouch	GPT-4 (Examples)	0.65	0.33
GreenEats Reusable Containers	GPT-4 (Base)	0.65	0.21
HydrationStation [bottle with filter]*	GPT-4 (Base)	0.64	0.19
Reusable Shopping Bag Set	GPT-4 (Examples)	0.64	0.19
CollegeLife Collapsible Laundry Hamper	GPT-4 (Base)	0.64	0.26
Adaptiflex [cord extension to fit big adapters]*	Student	0.64	0.44
SpaceSaver Hangers	GPT-4 (Base)	0.64	0.33
Dorm Room Air Purifier	GPT-4 (Examples)	0.63	0.29
Smart Power Strip	GPT-4 (Examples)	0.63	0.22
CampusCharger Pro	GPT-4 (Base)	0.63	0.31
Kitchen Safe Gloves	Student	0.62	0.31
Nightstand Nook [charging, cup holder]*	GPT-4 (Examples)	0.62	0.43
Mini Steamer	GPT-4 (Examples)	0.62	0.41
CollegeCare First Aid Kit	GPT-4 (Base)	0.62	0.26
StudySoundProof [soundproofing panels]*	GPT-4 (Base)	0.62	0.57
FreshAir Fan	GPT-4 (Base)	0.62	0.29
StudyBuddy Lamp [portable, usb charging]*	GPT-4 (Base)	0.62	0.43
Bluetooth Signal Merger [share music]*	Student	0.62	0.41
Adjustable Laptop Riser	GPT-4 (Examples)	0.62	0.21
EcoCharge [solar powered charger]*	GPT-4 (Base)	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 (Base)	0.62	0.42
Multipurpose Fitness Equipment	GPT-4 (Examples)	0.62	0.37
CollegeCooker	GPT-4 (Base)	0.61	0.50
Multifunctional Wall Organizer	GPT-4 (Examples)	0.61	0.31
DormDoc Portable Scanner	GPT-4 (Base)	0.61	0.49
Mobile Charging Station Organizer	GPT-4 (Examples)	0.61	0.26
StudyMate Planner	GPT-4 (Examples)	0.61	0.22
DormChef Kitchen Set	GPT-4 (Base)	0.61	0.33
LaundryBuddy [laundry basket]*	GPT-4 (Base)	0.61	0.30

* Text in square brackets [] is not part of the original title and was added to clarify the idea.

Electronic copy available at: https://ssrn.com/abstract=4526071