How AI outperforms humans at creative idea generation

Noah Castelo
Institution: Alberta School of Business
Title: Assistant Professor
Address: 4-20D Business Building, 11203 Saskatchewan Drive NW, Edmonton, AB, T6G 2R6
Phone number: +1 (587) 991-7326
Email address: ncastelo@ualberta.ca

Zsolt Katona
Institution: Haas School of Business, UC Berkeley
Title: Cheryl and Christian Valentine Professor
Address: Haas School of Business, UC Berkeley, Berkeley, California, 94720-1900
Phone number: +1 (510) 269-7658
Email address: zskatona@berkeley.edu

Peiyao Li
Institution: Haas School of Business, UC Berkeley
Address: Haas School of Business, UC Berkeley, Berkeley, California, 94720-1900
Phone number: +1 (626) 623-2075
Email address: ojhfklsjhl@berkeley.edu

Miklos Sarvary
Institution: Columbia Business School
Title: Carson Family Professor of Business
Address: Columbia Business School, Kravis 747, 665 west 130 street, New York, NY 10027
Phone number: +1 (347) 334-4066
Email address: miklos.sarvary@gsb.columbia.edu

Authors' note: The authors are listed in alphabetical order

How AI outperforms humans at creative idea generation

Feburary 2024

Abstract

This paper investigates the creative potential of artificial intelligence (AI), specifically the GPT4 large language model, in generating new product ideas. We first demonstrate that GPT4 can generate ideas that are rated as more creative than those generated by laypeople and creative professionals working under strong financial incentives. We then propose a novel text analysis methodology that we use to determine what specifically GPT4 is doing to achieve these superior results. We develop measures of two types of creativity: creative form (i.e., the language used to describe an idea being more unusual or unique) and creative substance (i.e., the idea itself is more novel). We find that GPT4 outperforms humans in both types of creativity. Finally, we show that using GPT4 to simply re-write ideas originally written by humans results in those ideas being rated as significantly more creative and that the improvement is explained by creativity in form. These results have immediate implications for managers, researchers, and practitioners in various fields, as integrating AI systems may optimize creative processes, save resources, and accelerate innovation cycles.

Keywords: artificial intelligence, creativity, product design, advertising, marketing research

1 Introduction

Creativity is often defined as the ability to generate novel and useful ideas that can solve problems or create value (Guilford, 1950; Amabile et al., 1988; Burroughs et al., 2011). In today's business environment, creativity is not only a desirable skill but a strategic necessity. Creativity can help businesses innovate, differentiate, and grow and is fundamental to marketing functions including product design, advertising, and consumer research. Many studies have confirmed the importance of creativity for business success. For example, advertisements' creativity can increase their persuasiveness and make consumers more open-minded about a more unfamiliar product (Yang and Smith, 2009). Firms that score higher on a measure of creativity also tend to perform better financially (Brodherson et al., 2017), and a survey of 1,379 chief executive officers in 79 countries found that creativity is seen as the most important skill for organizational success but also the hardest to find in employees (PricewaterhouseCoopers, 2017). Identifying methods for improving organizational creativity is therefore a key management priority. The same report argued that artificial intelligence (AI) "lacks the ... creativity to make sense of data," claiming that "creativity can't be coded." This reflects the common beliefs that AI cannot be creative (Marrone, Taddeo, and Hill, 2022) and that creative jobs are among the least likely to be automated (Frey and Osborne, 2017; Josten and Lordan, 2022; Institute, 2018).

More recently, we polled 200 individuals with management experience recruited from Prolific in May 2023 (average age = 35, 41% female) whether they believed that "today, AI can be as creative as humans who work in creative fields, like product design or advertising." 103 said no, 25 were not sure, and 72 said yes. Thus, most managers still doubt whether AI can be as creative as human professionals. These beliefs may now be outdated in light of recent advances in generative AI models like OpenAI's GPT4 (Eloundou et al., 2023; OpenAI, 2023). In this article, we provide evidence that these models can generate ideas for new products that are judged as equally as creative or more creative than ideas generated by both lay people and human product designers incentivized for creativity. We develop novel text analysis methods to identify the underlying reasons AI is perceived more creative than humans. We demonstrate that GPT4's superior performance can be attributed to both creativity in form (i.e., the specific words it uses to describe its ideas) and creativity in substance (i.e., the substance of the ideas that it generates, independent of the language

used to describe them).

2 Related work

This research provides the first empirical demonstration that AI can now outperform specialized creative professionals at creative tasks, and the first attempt to explain how AI achieves this feat, by disentangling creativity in form and creativity in substance. While prior work has compared AI creativity to that of laypeople or human-machine teams, none has recruited creative professionals, incentivized them to generate highly creative ideas, and directly compared those ideas to AI-generated ideas. It therefore remains unclear how AI's creative performance compares to that of specialized, incentivized creative professionals. At least two published articles have tested AI's abilities to improve creative outcomes in marketing-related outcomes. In the first, GPT-2 was used to draft website content with the goal of improving search engine optimization (SEO); the machine-generated content was edited by humans and this human-machine team outperformed humans working alone in terms of SEO outcomes (Reisenbichler et al., 2022). This study leaves open the question of whether AI alone (without human editing) can outperform human experts. In the second article, an undisclosed AI program was used to assist telemarketers by suggesting answers to customers' questions; this support improved the creativity of the telemarketers' answers, which boosted sales (Jia et al., 2023). This study also focuses on human-machine teams and therefore does not speak to AI's ability to outperform creative professionals without collaboration.

Other recent research has compared AI-generated ideas to ideas generated by laypeople in non-marketing contexts. For example, GPT4 can outperform most laypeople on classic tests of creative thinking such as the Torrance Tests of Creative Thinking (Guzik, Byrge, and Gilde, 2023) and the Alternate Uses Task (Koivisto and Grassini, 2023). Another study found that collaboration between lay people and AI produced the most creative haiku compared to those produced by either AI or humans working alone (Hitsuwari et al., 2023).

Beyond demonstrating for the first time that AI can outperform incentivized creative professionals at creative idea generation, we also shed light on how this feat is accomplished. To do so, we develop a model that separates two mechanisms that contribute to a product idea's perceived creativity – form differences and substantive differences. We design a novel Large Language Model

(LLM)-powered measure of form-level creativity and use it to separately identify the effects of form and substantive differences on perceived creativity. Furthermore, we discuss some insights into the heterogeneous effects of form and substantive differences between genders. Lastly, we provide some useful extensions where practitioners can use LLMs to (i) rewrite human-generated ideas to make them seem more creative and (ii) to rate the creativity of existing product design ideas (Li et al., 2024).

The remainder of this paper proceeds in 4 parts. First, in section 3, we briefly describe two simple experiments, and we use the raw data to show preliminary evidence that GPT4's ideas for new products are rated as more creative than ideas generated by professional product designers and laypeople. Second, in section 4, we introduce our modeling approach in which we empirically separate creativity in form and creativity in substance and apply a full identification strategy of each type of effect, demonstrating that GPT4 outperforms humans on both measures. Thirdly, in section 5, we show that both form and substantive differences between GPT4 and human-generated ideas have significant effects on perceived creativity. Lastly, we conduct practical extensions such as studying heterogeneous effects on rater demographics (males versus females), effects on highly creative product design ideas, and using GPT4 to rewrite ideas originally written by humans to make these ideas appear more creative (while restricting the length of the rewritten ideas to be similar to the original ones).

3 GPT4 outperforms humans at new product ideation

3.1 Originality, innovativeness, and usefulness

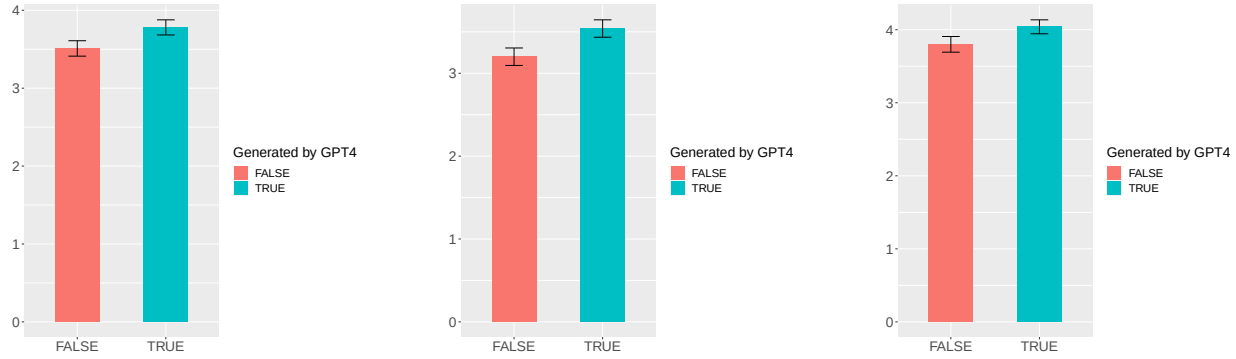
We solicited creative ideas for a new smartphone application (a digital product) that could help users feel less lonely. We used a website called Design Crowd to run a "design contest" in which professional app designers submit their best ideas in response to a prompt, with the best idea receiving a payment of approximately $500. We posted the following prompt: "Please suggest an idea for a new app that could help reduce loneliness. All ideas are welcome and creativity is encouraged! Please describe your idea in 300 words or less and include information about potential features and functionalities the app could have." Twenty app designers submitted an idea, and the creator of the highest-rated idea was paid. We then provided the same prompt to GPT4, twenty

times. Finally, we recruited 100 Prolific users (50% female, mean age = 30) who reported being lonely at least some of the time in a pre-screening survey. These participants are therefore part of the target market for apps that can reduce loneliness. We asked them to rate two ideas each (one human-generated and one GPT4-generated) according to how original, innovative, and useful the ideas seemed, again using 1–5 scales. Participants were not aware of how or by whom the ideas were generated. As shown in figure 1, GPT4-generated ideas were rated as more original ($M_{AI} = 3.54$ vs. $M_{human} = 3.20$, $t = 2.29$, $p = .012$, power = 0.74), more innovative ($M_{AI} = 3.78$ vs. $M_{human} = 3.51$, $t = 1.95$, $p = .026$, power = 0.62), and more useful ($M_{AI} = 4.04$ vs. $M_{human} = 3.80$, $t = 1.67$, $p = .048$, power = 0.51) than human-generated ideas. All of these results are significant at the 5% level after applying the Holm–Bonferroni correction except for the difference in innovativeness which is significant at the 10% level.

In this specific experiment, participants rated one AI-generated idea and one human-generated idea, in random order. We therefore also employed a mixed-effects linear model that incorporated both fixed effects for the source of the idea (AI vs. human) and random effects for participants. The fixed effects within the "human" and "AI" categories allowed for the comparison of average ratings between these two groups, while the random effects model accounted for the non-independence of observations, which were "crossed" by both raters/judges and ideas. This statistical approach ensured that the variability among judges and the potential correlation of ratings among judges were appropriately considered.

The results of the mixed effects model revealed a statistically significant difference in the ratings of ideas generated by AI and humans. We discuss the details in subsection 5.1.

These findings support the conclusion that AI ideas were rated more creatively than human ideas in this study, and this difference remained significant even when accounting for the complex structure of the data and potential confounding effects of judge and idea interdependencies. All remaining experiments described in this article asked participants to rate just a single idea.



(a) Innovativeness (b) Originality (c) Usefulness

Figure 1: Loneliness reduction app design perceived innovativeness, originality, and usefulness

App Idea: "PetMatch"
Description: PetMatch is a heartwarming app that aims to reduce loneliness by connecting users with animals in need of love, attention, and care. The app leverages the power of animal companionship to create lasting bonds, improve mental well-being, and foster a sense of purpose. By matching users with local pet shelters, foster opportunities, or fellow pet enthusiasts, PetMatch encourages meaningful interactions and shared experiences.
Key Features:

1. Shelter Connections: Users can discover nearby pet shelters and learn about the animals available for adoption. PetMatch uses AI to analyze users' preferences and lifestyles, offering personalized pet recommendations to ensure a successful match.
2. Foster Buddies: Users can sign up to become temporary foster parents for pets in need. This feature allows users to experience the joy of pet companionship without long-term commitment and helps shelters by freeing up space for other animals in need.
3. Pet Pals: PetMatch connects users with fellow pet enthusiasts in their area. Users can schedule playdates for their pets, attend pet-friendly events, or simply meet for a walk or coffee, fostering friendships among both humans and pets.
4. Virtual Pet Therapy: Users who may not be able to commit to pet ownership can still benefit from the therapeutic effects of animals. The app offers live video sessions with trained therapy animals, allowing users to interact and engage with them virtually.
5. Pet Care Workshops: The app hosts live and on-demand workshops on various pet care topics, such as training, nutrition, and grooming. These sessions encourage users to learn, connect with experts, and share experiences with fellow pet owners.
6. Volunteering Opportunities: Users can find local volunteering opportunities at animal shelters, wildlife rescues, or pet sanctuaries. This feature fosters a sense of purpose, personal satisfaction, and connections with other volunteers.
7. Pet-Friendly Travel: The app offers a curated list of pet-friendly destinations, accommodations, and activities for users to plan their next adventure with their furry friends.

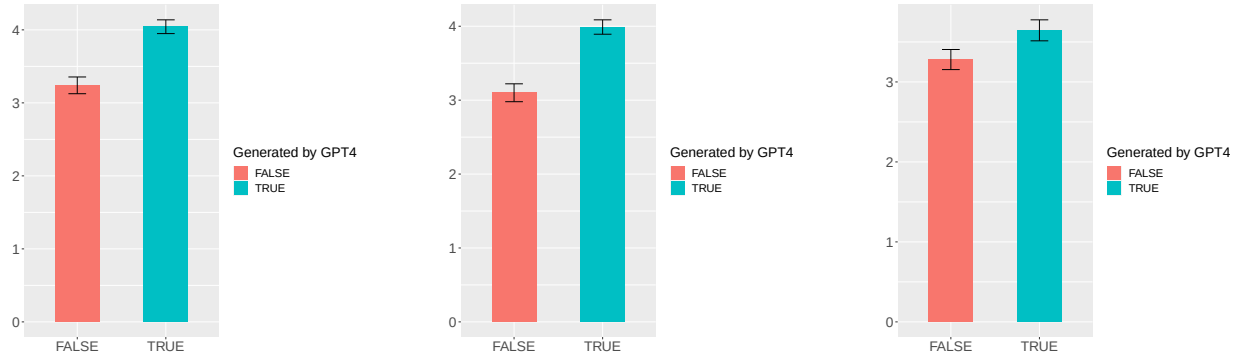
By tapping into the powerful bond between humans and animals, PetMatch creates a platform for meaningful connections and shared experiences, ultimately reducing loneliness and promoting emotional well-being.

Figure 2: Most creative loneliness-reduction app idea

One potential concern with using GPT4 to generate creative ideas is that the model may be simply "regurgitating" the content on which it was trained. The model has indeed been trained on

vast swathes of the internet, such that when asked for creative ways to use a tin can or creative apps to reduce loneliness, it might simply be re-stating ideas that were produced by humans and recorded on the internet, perhaps in a more eloquent manner. It is possible that the GPT4-generated ideas in the first test were variations of ideas contained in its training data. We partly address this concern in the next section by empirically separating creativity in form and creativity in substance. First, however, we address this concern experimentally by asking humans and GPT4 to invent entirely new products. In this way, we can diminish the likelihood that the AI is regurgitating its training data, since the products it invents in this next experiment do not exist and no information about them is available on the internet.

Using Prolific, we recruited ten individuals who work in marketing, sales, or advertising, and gave them the following prompt: "Put yourself in the shoes of a creative director at an award-winning ad agency. Invent a new product that is both creative and useful. Pitch it to potential customers using no more than 200 words." We gave the same prompt to GPT4, ten times. We then recruited 200 individuals from Prolific and asked them to read one idea, blinded to the author's identity, and rate it on the dimensions of innovativeness, originality, and usefulness using 5-point scales. As shown in figure 3, GPT4-generated product ideas were again rated as more innovative ($M_{AI} = 4.00$ vs. $M_{human} = 3.24$, $t = 4.96$, $p < .001$, power > 0.99), more original ($M_{AI} = 3.95$ vs. $M_{human} = 3.10$, $t = 5.30$, $p < .001$, power > 0.99), and more useful ($M_{AI} = 3.55$ vs. $M_{human} = 3.28$, $t = 2.21$, $p = .073$, power $= 0.37$) than human-generated ideas. All of these results are significant at the 5% level after applying the Holm–Bonferroni correction except for the perceived usefulness which is significant at the 10% level. By asking for ideas for entirely new products, we diminish the concern that AI is simply re-stating content from its training data. However, to address this concern in a separate way, we next turn to empirically disentangling creativity in form (i.e. just re-stating existing ideas using more unusual language) and creativity in substance (i.e., generating ideas that are substantively novel).



(a) Innovativeness (b) Originality (c) Usefulness

Figure 3: Perceived innovativeness, originality, and usefulness of new products

3.2 Overall creativity

In subsection 3.1, we analyze each creativity measure separately, and we find that GPT4-generated product design ideas are perceived as more innovative, more original, and more useful. In this subsection, we use Principal Component Analysis (PCA) to combine all three measures into one overall creativity measure and compare GPT4 and human-generated ideas on this holistic measure of creativity.

Table 1: Correlation between types of creativity ratings

	innovative	original
innovative		
original	.76***	
useful	.48***	.37***

Prior literature on creativity in marketing has shown that the creativity of a product is correlated with both novelty and relatedness (Smith et al., 2007; Toubia and Netzer, 2017). We have two measures of novelty – innovativeness and originality, and one measure of relatedness – usefulness. Therefore, to create a measure of overall creativity of a product design idea: we use Principal Component Analysis to find the dimension that explains the largest amount of variation across the three types of creativity ratings. As shown in table 1, we find that all three ratings: innovativeness, originality, and usefulness are significantly positively correlated. In particular, the two novelty measures – innovativeness and originality – are highly correlated. Therefore, we first conduct a PCA on these two dimensions to combine them into one measure. Then, we combine this measure

with usefulness to form the final overall creativity measure. [1]

Table 2: Correlation between overall creativity and individual creativity measures

	Innovativeness	Originality	Usefulness
Experiment 1	0.88	0.86	0.85
Experiment 2	0.77	0.68	0.88
Experiments 1 and 2	0.82	0.77	0.84

As shown in table 2, the constructed overall creativity measure is strongly correlated with innovativeness, originality, and usefulness in each experiment and when both are pooled together. In experiment 1, the average overall creativity of GPT4-generated ideas is significantly larger than the average creativity of human-generated ideas ($t = 2.28$ and $p = 0.011$). In experiment 2, the average overall creativity of GPT4-generated ideas is significantly larger than the average creativity of human-generated ideas ($t = 3.69$ and $p < 0.001$). Therefore, we have suggestive evidence that GPT4-generated ideas are perceived as more creative than human-generated ideas. In addition, in experiment 1, the most highly rated human-generated idea has an average overall creativity of 1.04, and the most highly rated GPT4-generated idea has an average overall creativity of 0.97. The two ratings are not statistically different at the 10% level. In experiment 2, the most highly rated human-generated idea has an average overall creativity of 0.37, and the most highly rated GPT4-generated idea has an average overall creativity of 0.84. The two ratings are not statistically different at the 10% level. This shows that the best GPT4-generated idea is also comparable to the best human-generated product design idea in terms of overall creativity.

4 Model

In section 3, we show suggestive evidence that GPT4 can outperform laypeople and design professionals in some product design tasks. In this section, we examine the mechanisms that cause these differences in the perceived creativity of product design ideas. We model creativity at two levels: creativity in form and creativity in substance. More specifically, in terms of product design, identical products might be assigned varying creativity scores if the language used to describe one

[1] Ratings of innovativeness, originality, and usefulness and standardized before going into the PCA. In addition, PCA is conducted separately for data collected in experiment one and experiment two because the perception of creativity may be different in loneliness reduction apps (experiment one) and pitches for completely new products (experiment two).

product conveys a sense of greater distinctiveness. Alternatively, in some instances, different ratings for two products may arise from differences between the two ideas that are not mediated by word usage. We consider the former word-mediated distinctions as differences in form and distinctions beyond the forms as differences in substance.

4.1 Word-sense divergence

Our goal is the identify the form and substantive differences between GPT4-generated product design ideas and human-generated ones. To do this, we design a measure to control for the difference in form. We measure the form differences among different designs by examining the usage of words. Given a simple sentence such as "Fruit flies like apples.", if we only consider the word "flies" while covering the other 3 words, we would likely think that "flies" as a second word in a short 4-word sentence is commonly used as a verb following a noun. We consider this the *uncontextualized* meaning of "flies" in this sentence. However, if we see the whole text, we would see that in the context of this particular sentence, the word "flies" is a noun that means a type of insect that likes to eat apples. We consider this the *contextualized* meaning of "flies" in this particular sentence.

We consider the description of a product design to be creative if on average it uses words whose contextualized meanings diverge more from its uncontextualized meanings. This measure is composed of two components: the contextualized meaning of a word within the text, and the uncontextualized meaning of a word.

4.1.1 Examples of contextualized and uncontextualized meaning

In the case of product design, one can modify the same idea by replacing some words using another word whose usage in the specific context seems more creative. We illustrate this difference by using examples that embody the same product design idea:

Example 1 (Original idea):

> The all in one toothpaste brush is a toothbrush with toothpaste inside it. Just squeeze the toothbrush handle and the paste will be automatically dispensed. It is ideal for travel. No need to take bulky boxes, just take the toothpaste brush. What's more it is

refillable and therefore more sustainable.

Example 2 (Rewritten more creative idea):

> Experience the convenience of our innovative All-In-One Toothpaste Brush - a uniquely designed toothbrush embedded with toothpaste. With a mere squeeze of the handle, the paste oozes out, ready to refresh your breath. Its compact design is perfect for on-the-go needs, eliminating the need for cumbersome carry-ons. Best of all, its refillable feature guarantees a greener choice for your oral hygiene.

Both examples describe a product that combines toothpaste and toothbrush in one device, and the device is spacing-saving and refillable. However, examples use many words that are not commonly used to describe a tooth-cleaning product. For example, in example 2, the verb "squeeze" has the contextualized meaning of "put pressure on the handle of the toothbrush". However, its more common meaning in the English language (uncontextualized) is to 'squeeze something out" where the "something" that is being squeezed is usually not the handle of the toothbrush. The adjective "greener" in example two has the contextualized meaning of more eco-friendly, but it is more commonly (uncontextualized) used to describe a deeper shade of the color green.

4.2 Quantifying word level creativity using contextualized and uncontextualized meaning

The challenge is to quantify the word-level creativity of each expression of an idea using these abstract word-senses (meaning of a word). The intuitive idea is to first estimate a contextualized (within a given text) representation and an uncontextualized representation (within a robust representation of common English) of the meaning of each word. Then, computing the similarity between these two meanings measures how creatively this word is used relative to the common English language. We use a pre-trained language model to estimate vector embeddings that represent the contextual meaning of each word. A subtlety in our application is that we also a use generative language model, GPT4 to generate the product design ideas; therefore, these ideas are generated based on GPT embeddings (numerical vectors that represent the meaning of texts).

Therefore, using GPT to produce contextualized and uncontextualized embeddings would result in a higher similarity between contextualized and uncontextualized embeddings of product design ideas generated by GPT4 because GPT4 uses a base model that is trained to generate the most likely next-word given the embedding of the previous text. More specifically, GPT4 is likely to use words whose uncontextualized meaning fits the context as defined by GPT embeddings. Thus, to avoid these complications, we compute the contextualized and uncontextualized embeddings using a well-studied language model not in the GPT family: Bidirectional Encoder Representations from Transformers (BERT). BERT is trained on *BookCorpus* and *Wikipedia*. *BookCorpus* is a large collection of free novel books written by unpublished authors. It contains over ten thousand books spanning 16 different sub-genres such as Romance, Historical, Adventure, etc. *Wikipedia* is an open-source library containing descriptions of historical events, biographies, and important scientific discoveries. These two data sets have been widely used and regarded as a reliable representation of word meanings in common English. Thus, if we can extract a BERT embedding of a word without providing specific contextual information, we can get a meaningful representation of the uncontextualized meaning of the word.

Now, we discuss the mathematical details of using BERT to extract contextualized and uncontextualized word embeddings. Given an input sentence made of N subword tokens $s = [t_1, t_2, ..., t_N]$, BERT first assigns an initial static embedding to each token which contains a preliminary raw estimation of the meaning of each token by itself as well as the positional information of this token within the input text. Mathematically, we can represent the token t_i as

$$f(t_i) = f(\text{Pos}_{i,s}, r_i) = E_{0,i} \tag{1}$$

where $\text{Pos}_{i,s}$ is a vector that contains the positional information within s, r_i is a vector that contains the raw representation of the meaning of the token, and $E_{0,i}$ is a vector that represents the 0th layer (input) embedding of the token.

BERT takes the input and refines and contextualizes the embedding of each token through multiple attention layers. Each attention layer takes an N-by-L matrix of N L-dimensional embeddings as input and outputs a new N-by-L matrix of embeddings. More specifically, it uses the

self-attention mechanism which can be expressed as

$$E_t \propto \text{softmax}\left(X_q \cdot X_k\right) X_v \tag{2}$$

where $X_q = E_{t-1}W_q$ (N-by-H), $X_k = E_{t-1}W_k$ (N-by-H), and $X_v = E_{t-1}W_v$ (N-by-L). W_q (L-by-H), W_k (L-by-H), W_v (L-by-L) are matrices of trainable parameters that are defined through pre-training on BookCorpus and Wikipedia.

Therefore, the output for each token of the entire BERT architecture is a representation that contains information about the positional information of this token as context information with respect to other tokens' contextualized embeddings. In short, it represents the meaning of the word within the sentence at the current position of the token. We denote the contextualized embedding of the ith token as $E_{\text{cont}}(i)$.

In addition, we compute an uncontextualized embedding for each token. Intuitively, the uncontextualized embedding of a token is the numerical representation of the meaning of a token irrespective of the context of the product design idea. Each generative language model has a different mapping to generate embeddings. Therefore, to construct the uncontextualized embedding of each token and for it to be comparable to the contextualized embedding, we need to use the same generative language model, respect the token's position in the original text, and omit any meaning carried by other tokens in the same text. To achieve the first goal, we pass the input to the same BERT model. To achieve the second goal, we ensure that the length of the input is the same as the original text, and the focal token is at its original position. To accomplish the third goal, we use the special mask token in BERT to cover the meaning of all other tokens in the edited sequence. Putting everything together, we use the BERT model and the input to the model is a sequence of equal length as the original text used to extract the contextualized embedding of the token, but all tokens except the focal token are replaced with the mask token. Mathematically, to extract the uncontextualized embedding for the ith token in s, the input sequence is

$$s_{\text{uncont}} = [[\text{MASK}], \ldots, [\text{MASK}], t_i, [\text{MASK}], \ldots, [\text{MASK}]] \tag{3}$$

We denote the uncontextualized embedding vector of the ith token as $E_{\text{uncont}}(i)$.

To provide a concrete example of similarity between contextualized and uncontextualized embedding, consider the example we alluded to in the beginning of this section: "Fruit flies like apples." If we mask the identity of all other words except "flies" i.e. "[MASK] flies [MASK] [MASK] [MASK]", we would likely interpret the word "flies" as the third person present tense of the verb "fly" (its uncontextualized meaning). However, given the identity of the other words, we would likely interpret "flies" as the plural form of the noun that describes the animal "fly" (its contextualized meaning).

After constructing the contextualized and uncontextualized embedding of each token, we estimate a measure of the form-level creativity (creativity in using highly distinctive words to describe an idea) we call word-sense divergence using the similarity between contextualized and uncontextualized embeddings. Denote the word-sense divergence of an idea j as a scalar D_j, let N_j denote the number of tokens in idea j

$$D_j = -\frac{\sum_{i=1}^{N_j} \frac{E_{\text{cont}}(i) \cdot E_{\text{uncont}}(i)}{\sqrt{E_{\text{cont}}(i) \cdot E_{\text{cont}}(i)}\sqrt{E_{\text{uncont}}(i) \cdot E_{\text{uncont}}(i)}}}{N_j} \tag{4}$$

In simple terms, the word-sense divergence of an idea is the additive inverse of the mean cosine similarity of each token's contextualized and uncontextualized embedding. If the word-sense divergence of an idea is high, it means on average, the contextualized meaning of words (tokens) in the text of this idea is more different than these words' uncontextualized meaning (the meaning that is not affected by other words in the same text).

4.3 Econometric identification

Using product design ideas and human ratings we have collected shown in section 3. We conduct our identification in two stages.

1. We identify the overall effect of using GPT4 on perceived creativity.

2. We identify the effect of word-sense divergence on perceived creativity.

3. We identify the effect of using GPT4 to generate product design ideas on perceived creativity beyond the creativity of word usage (measured by word-sense divergence).

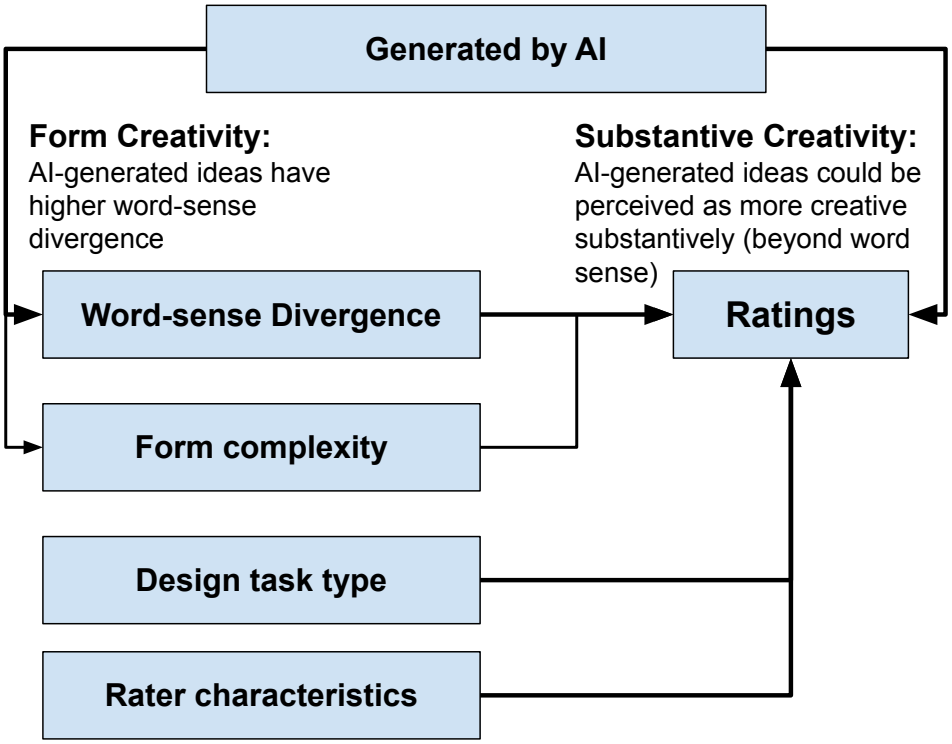


Figure 4: Identification diagram of the effect of form and substantive differences on perceived creativity ratings

In the first stage, we run a mixed-effect regression where the dependent variables are the perceived creativity measures, and the independent variable is whether or not a rating is on a GPT4 or human-generated product design idea. We study four types of creativity-related ratings. First, we ask each rater to rate a product design idea on innovativeness, originality, and usefulness. Then, we use Principal Component Analysis to impute one measure that explains the most amount of variation across the three sets of ratings. We consider this combined measure as a measure of creativity. As shown in figure 4, two other types of variation may bias the estimate: different raters may have different perceptions of creativity, and different tasks (experiment one and experiment) also have different meanings of creativity. Since the participants are recruited from the survey platform *Prolific.com*, we assume the rater effects are random draws from a normal distribution and use a random effect of raters. In addition, we use a fixed effect for each task. In addition,

we consider a form-level complexity that is not accounted for by the task type and word-sense divergence (form-level creativity). We measure this residual form complexity of each idea as the residual when regressing the word count of each idea on the task type (experiment one or two) and word sense divergence. We call this measure the residual word count.

Mathematically, the regression specifications are as follows: for rating k as innovativeness, originality, usefulness, or overall creativity

$$R_{i,k} = \beta_k \mathbf{1}(\text{idea } i \text{ is generated by GPT4}) + \text{FE}_{i,k}^{\text{Design task type}} + \text{RE}_{i,k}^{\text{rater}} + \epsilon_{i,k}. \tag{5}$$

In the second stage, we analyze the effect of word-sense divergence on these dimensions of creativity as well as the overall perceived creativity. To do so, we also run mixed-effect regressions of ratings on word-sense divergence while controlling for the design task fixed effects. GPT4 or human-generated fixed effects, and rater characteristic random effects following figure 4.

For each of the four types of ratings k as innovativeness, originality, usefulness, or overall creativity, we consider the causal structure relating word-sense-divergence, the source of GPT4 (generated by GPT4 or not), each rater's characteristics, and the task of the product design (loneliness app or new product). First, if an idea is generated by GPT4, the word sense divergence may be different because GPT4 and human word usage patterns may not be identical. In addition, An idea generated by GPT4 may receive different ratings because of the differences in word usage, or substantive differences of the idea compared to human-generated ones. Therefore, the "generated by AI" indicator is a confounder for word-sense divergence and ratings. From another perspective, word-sense divergence is a mediator between the "generated by AI" indicator and the ratings. Another confounder for word-sense divergence and ratings is the type of design task. On one hand, the level of word-sense divergence in ideas about loneliness reduction apps may be different from the ideas about new fictional products. On the other hand, product ideas in these two categories may receive different ratings. Therefore, in order to identify the effect of word-sense divergence on ratings, we need to control for the confounders. Once we have the effect of word-sense divergence on ratings, we can control for the effect of using GPT4 to generate product design ideas on the perceived creativity of the ideas through word-sense divergence. Then, we can identify the unmediated effect of using a GPT4 due to the substantive differences between GPT4 and human-generated

ideas. Another set of factors that affects the ratings but not the differences between GPT4 and human-generated ideas are the characteristics of the raters. Different raters may have different rating habits – systematically giving higher or lower ratings. Therefore, to increase the robustness of our specification, we include random effects for raters too.

Mathematically, the regression specifications are as follows: for rating k as innovativeness, originality, usefulness, or overall creativity

$$R_{i,k} = \alpha_k D_i + \lambda_k \text{RWC}_i + \text{FE}_{i,k}^{\text{Design task type}} + \text{FE}_{i,k}^{\mathbf{1}(\text{idea } i \text{ is generated by GPT4})} + \text{RE}_{i,k}^{\text{rater}} + \epsilon_{i,k} \quad (6)$$

In the third stage, we identify the effect of the substantive differences between GPT4 and human-generated product design ideas beyond the creative use of words (word-sense divergence) on human creativity ratings. To do this, we remove the estimated effects of word-sense divergence from regressions 6 from the ratings and run the following regressions to identify the substantive effect of using GPT4 to generate product design ideas on the perceived creativity of the design. Let $R'_{i,k} = R_{i,k} - \alpha_k D_i - \lambda_k \text{RWC}_i$:

$$R'_{i,k} = \tau_k \mathbf{1}(\text{generated by AI})_i + \text{FE}_{i,k}^{\text{Design task type}} + \text{RE}_{i,k}^{\text{rater}} + \epsilon'_{i,k} \quad (7)$$

5 Results

In this section, we show the main results where we identify the effect of using GPT4 to generate product design ideas on the perceived creativity of these ideas in subsection 5.1. Furthermore, we break down the effect into effects due to form differences and substantive differences in subsection 5.2. In the following analysis, we standardize the ratings of innovativeness, originality, usefulness, and creativity in order to make the interpretation of the effect estimates more direct. In particular, even though we show our results for each measure, the most representative one is the overall creativity measure as it provides a comprehensive quantification of creativity.

Table 3: Overall effect of using GPT4 to generate product design ideas of the perceived creativity of the ideas

	Dependent variable:			
	innovativeness	originality	usefulness	creativity
	(1)	(2)	(3)	(4)
Using GPT4	0.421***	0.494***	0.215**	0.397***
	(0.086)	(0.093)	(0.086)	(0.089)
Observations	383	383	383	383
Log Likelihood	-531.558	-536.212	-535.453	-536.782
Note:				*p<0.1; **p<0.05; ***p<0.01

5.1 Identifying the overall effect of using GPT4 to generate product design ideas of the perceived creativity of the ideas

First, we study the overall effect of using GPT4 to generate product design ideas. As shown in table 3, on average, GPT4-generated product design ideas on average have higher overall creativity than human-generated ones by about 0.40 standard deviation. This result is statistically significant at the 1% level.[2] More specifically, using GPT4 to generate product design ideas increases the perceived innovativeness by around 0.42 standard deviation, originality by about 0.49 standard deviation, and usefulness by about 0.22 standard deviation. These results provide evidence that GPT4 can be used to generate product design ideas that are perceived as more creative than human-generated ideas.

5.2 Identifying the effect of form and substantive differences between GPT4 and human-generated product design ideas

In section 3 and subsection 5.1, we have established evidence that GPT4-generated product design ideas are perceived as more creative than human-generated ones. In the following analysis, we aim to disentangle the effect of creativity on rating caused by differences in forms and ideas embodied in the texts. More specifically, we identify the effect of word-sense divergence on ratings and further identify the effect of using a GPT4 to generate product designs on perceived creativity beyond the

[2]When we refer to an estimate being statistically significant for the rest of the paper, we mean the estimate is significantly at the 5% level unless otherwise specified.

use of words (substantive differences).

5.2.1 Identifying the effect of word-sense divergence on perceived creativity – creativity in form

We consider all 57 product design ideas from studies 1 and 2. Using a permutation test on the word-sense divergence of these ideas (27 generated by GPT4 and 30 generated by humans), we find that GPT4-generated ideas have a significantly higher average word-sense divergence than the human-generated ones. We further this finding by studying whether differences in word-sense divergence have any effect on the 4 perceived creativity ratings: innovativeness, originality, usefulness, and overall creativity.

Table 4: Effect of form differences (word-sense divergence) on perceived creativity

	Dependent variable:			
	innovativeness	originality	usefulness	creativity
	(1)	(2)	(3)	(4)
word-sense divergence	0.080*	0.119**	0.016	0.071*
	(0.050)	(0.050)	(0.051)	(0.051)
residual word count	0.003**	0.002**	0.001	0.002**
	(0.001)	(0.001)	(0.001)	(0.001)
Observations	383	383	383	383
Log Likelihood	-532.353	-535.489	-537.472	-537.859
Note:				*p<.1; **p<.05; ***p<.01

As shown in table 4, with the full econometric specification shown in equation 6, we find a significantly positive effect of word-sense divergence on the perceived overall creativity of ideas at the 10% level. In particular, each standard deviation of increase in word-sense divergences leads to 0.07 standard deviation of increase in perceived overall creativity. Breaking it down further, we find that a higher level of word-sense divergence has a significantly positive effect on the perceived originality of the product idea. This means that using words in unusual contexts (relative to common English language patterns defined by the vast training corpus of BERT) makes an idea look more original. In addition, we find a significant positive effect of word-sense divergence on perceived innovativeness

at the 10% level. More specifically, each standard deviation of increase in word-sense divergence corresponds to roughly 0.08 standard deviations of increase in perceived innovativeness, and 0.12 standard deviations of increase in perceived originality. Note that on average GPT4-generated ideas have significantly higher word-sense divergence than human-generated ideas ($t = 3.30$ and $p = 0.001$). Nonetheless, the effect of form level differences (word-sense divergence) makes up a relatively small portion of the overall effect shown in table 3. In addition, we note that the additional form level complexity measured by residual word count also positively contributes to the overall perceived creativity.

5.2.2 Identifying the effect of using GPT4 to generate product design ideas on creativity beyond the form of the text

We calculate the adjusted creativity ratings as $R'_{i,k} = R_{i,k} - \alpha_k D_i$ where α is the average treatment effect of word-sense divergence D_i on perceived creativity rating $R_{i,k}$. This removes the effect of form differences from the perceived creativity ratings. The adjusted ratings reflect only substantive differences between product design ideas (differences beyond the forms).

Table 5: Effect of substantive differences (beyond word-sense divergence) on perceived creativity

	Dependent variable:			
	adj innovative	adj original	adj useful	adj creativity
	(1)	(2)	(3)	(4)
using GPT4	0.356***	0.413***	0.199**	0.343***
	(0.085)	(0.091)	(0.086)	(0.089)
Observations	383	383	383	383
Log Likelihood	-524.260	-527.448	-529.400	-529.810
Note:				*p<0.1; **p<0.05; ***p<0.01

Next, we identify the causal effect of using GPT4 to design product ideas on perceived creativity ratings. As shown in table 5, with the full econometric specification shown in equation 7. We find that after the effect of word sense divergence is removed from the raw creativity ratings, substantive differences between GPT4 and human-generated ideas have significant effects on the perceived overall creativity rating at the 1% level. Using GPT4 to generate product design ideas leads to

an average increase in perceived overall creativity of 0.34 standard deviation. More specifically, GPT4 can generate product design ideas that are substantively significantly higher on perceived innovativeness, originality, usefulness, and overall creativity (all adjusted by the effect of word-sense divergence). More specifically, on average, the perceived adjusted innovativeness is increased by 0.36 standard deviation, the perceived adjusted originality is increased by 0.41 standard deviation, and the perceived adjusted usefulness is increased by 0.20 standard deviation.

The analysis in this subsection shows that using GPT4 to generate product ideas can lead to form and substantive differences that both significantly contribute to the perceived creativity ratings. In particular, once we remove the effect of word usage on perceived creativity, we find that substantive differences between GPT4-generated ideas and human ideas play a more significant role in driving the difference in perceived creativity of ideas.

6 Gender differences in effects on perceived creativity

Product designs are often targeted at specific consumer types, and one of the major differentiations is male versus female. It is well-documented that male and female-centric products have many differences. In addition, it has been shown that males and females have different brain activation functions when engaging in creativity performance assessment (Abraham et al., 2014). In particular, males pay more attention to semantics. In this section, we use our decomposition of form and substantive level creativity to examine whether males and females perceive creativity differently, especially in product design.

In section 5, we find that using GPT4 to generate product design ideas can lead to significantly higher perceived creativity via form and substantive differences. In this section, we decompose the effect of form and substantive differences and identify the heterogeneous effect of using GPT4 to generate product ideas on male versus female raters.

6.1 Heterogeneous effect of form differences

We identify the heterogeneous effect of form differences on perceived creativity using a modified version of equation 6. More specifically, instead of using each rater's gender as a fixed effect, we split the data set into two parts and examine the effect of word-sense divergence on perceived

creativity ratings separately for male and female raters.

Table 6: Effect of form differences (word-sense divergence) on perceived creativity: male

	Dependent variable:			
	innovative	originality	useful	creativity
	(1)	(2)	(3)	(4)
word-sense divergence	0.135**	0.253***	0.042	0.144**
	(0.071)	(0.067)	(0.073)	(0.071)
Observations	192	192	192	192
Log Likelihood	-275.651	-264.336	-280.612	-275.675
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 7: Effect of form differences (word-sense divergence) on perceived creativity: female

	Dependent variable:			
	innovative	original	useful	creativity
	(1)	(2)	(3)	(4)
word-sense divergence	0.024	-0.025	0.004	0.004
	(0.069)	(0.074)	(0.068)	(0.069)
Observations	191	191	191	191
Log Likelihood	-258.961	-271.938	-256.690	-260.700
Note:				*p<0.1; **p<0.05; ***p<0.01

As shown in tables 6 and 7, within male raters a higher word-sense divergence leads to significantly higher perceived originality and overall creativity, and the effect sizes are larger than the effect of word-sense divergence on perceived creativity ratings in the whole sample. In particular, each standard deviation of an increase in word-sense divergence corresponds to a 0.13 standard deviation increase in perceived innovativeness, a 0.25 standard deviation increase in perceived originality, and an over 0.14 standard deviation increase in perceived overall creativity. However, female raters' perceived creativity ratings are not affected by form-level differences (word-sense divergence). This shows that the effect identified in the entire sample is likely driven by male raters. When designing products that target the female subpopulation, the form level difference correlated

with using GPT4 to generate these ideas may not have an edge in the perceived creativity by targeted consumers. On the other hand, the difference in form is likely to affect the perceived creativity if the product is targeted at the male subpopulation.

6.2 Heterogeneous effect of substantive differences

Similar to subsection 6.1, we identify the heterogeneous effect of substantive differences using a modified version of equation 7 where we remove gender as a fixed effect and split the sample based on the reported gender of raters.

Table 8: Effect of substantive differences on perceived creativity: male

	Dependent variable:			
	adj innovative	adj original	adj useful	adj creativity
	(1)	(2)	(3)	(4)
using GPT4	0.449***	0.424***	0.338***	0.465***
	(0.120)	(0.123)	(0.125)	(0.123)
Observations	192	192	192	192
Log Likelihood	-268.275	-256.899	-273.304	-268.316
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 9: Effect of substantive differences on perceived creativity: female

	Dependent variable:			
	adj innovative	adj original	adj useful	adj creativity
	(1)	(2)	(3)	(4)
using GPT4	0.256**	0.403***	0.035	0.200*
	(0.122)	(0.135)	(0.113)	(0.125)
Observations	191	191	191	191
Log Likelihood	-251.498	-264.624	-249.171	-253.261
Note:				*p<0.1; **p<0.05; ***p<0.01

As shown in tables 8 and 9, substantive differences between GPT4 and human-generated product design ideas have a significantly positive effect on adjusted innovativeness, adjusted originality, and

adjusted overall creativity. Particularly, substantive differences have a more significant effect on the male subsample.

Putting together the heterogeneous effects identified in subsections 6.1 and 6.2, we find that within the male subsample, both form and substantive differences achieved by using GPT4 to generate product design ideas have a positive effect on perceived creativity ratings. However, within the female subsample of raters, the effects likely come from substantive differences between GPT4 and human-generated ideas and the effect is less significant ($p = 0.052$).

7 Analysis of ideas with high word-sense divergence

When selecting product design ideas for further development, practitioners may focus on only ideas that have a higher probability of receiving high creativity ratings. Therefore, we focus our analysis on human and GPT4-generated ideas that have above-median creativity ratings. We use ideas collected in both experiments (loneliness reduction app and new product advertisement) and look at GPT4 and human-generated ideas and only keep the ones that have average overall creativity ratings above the median of each idea generator (GPT4 or human) and each experiment. We follow the same procedure described in section 4.3 to identify the effects of form and substantive differences between GPT4 and human-generated highly word-sense divergent ideas.

Table 10: Effect of form differences on perceived creativity among ideas with high overall creativity

	Dependent variable:			
	adj innonvativeness	adj originality	adj usefulness	adj creativity
	(1)	(2)	(3)	(4)
word-sense divergence	-0.011	-0.066	0.032	0.066
	(0.168)	(0.175)	(0.163)	(0.165)
Observations	176	176	176	176
Log Likelihood	-244.320	-251.553	-239.421	-241.218
Note:				*p<0.1; **p<0.05; ***p<0.01

As shown in table 10, when restricting our analysis to product design ideas that are perceived as highly creative (above median average overall creativity rating), the effect of form level differences on perceived creativity measures is insignificant. This means that creative use of words (higher

word-sense divergence) does not lead to a higher perceived creativity if the idea is already highly creative.

Table 11: Effect of substantive differences on perceived creativity among ideas with high average overall creativity ratings

	Dependent variable:			
	adj innonvativeness	adj originality	adj usefulness	adj creativity
	(1)	(2)	(3)	(4)
using GPT4	0.424***	0.628***	-0.046	0.235**
	(0.124)	(0.131)	(0.120)	(0.124)
Observations	176	176	176	176
Log Likelihood	-238.422	-245.914	-233.211	-234.755
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 5 shows the effect of substantive differences between GPT4 and human-generated ideas on perceived creativity ratings, and table 11 shows this effect only among ideas with high average overall creativity ratings. In particular, we find that when conditioning on an idea's average overall creativity rating is above the median among ideas from the same generator and for the same task, the substantive difference induced by using GPT4 to generate the ideas corresponds to about 0.42 standard deviations of increase in the adjusted innovativeness rating, about 0.63 in the adjusted originality rating, and about 0.24 in the adjusted overall creativity rating. Comparing these values with the ones shown in table 5, we see that the size of the effect on the overall creativity measure is smaller when we focus on highly creative ideas. This suggests that even though GPT4 can be used to generate product design ideas that are, on average, perceived as more creative due to substantive differences between GPT4 and human-generated ideas, this improvement is more moderate when comparing high creative human and GPT4-generated ideas.

8 Using GPT4 to rewrite human-generated ideas

As shown in our main results in section 5, we have documented some use cases where GPT4 can be used to generate more creative ideas both in form and substance. However, many practitioners

may not have full faith in using an LLM like GPT4 to conduct the entire product design ideation. In addition, in some other cases, firms may have some ready-to-go product design ideas. Therefore, in application, we may want to use an LLM like GPT4 to improve existing product design ideas instead of generating new ones. In this section, we examine whether we can use GPT4 to rewrite human-generated product ideas more creatively.

We test GPT4's design-improving ability using human-generated ideas in study 2, where each professional is asked to write a pitch for a new product. In total, we have 10 product design ideas that range from an all-in-one toothpaste brush to a solar tower. We first ask GPT4 to rewrite each idea 5 times using the prompt: "You are a professional product designer. Your job is to rewrite the given idea to make it more creative. You cannot change the idea. Please rewrite the following idea in about N words:" where N is the word count of the original idea. Indeed, we find that the average word count for human original ideas is 71.5, and for GPT4-rewritten ideas, the average length is 72.7. Therefore, we do not include the extra text complexity measure in the regression. The two numbers are about 10% of the standard error of the mean word count in the two sets. Then, we rank the rewrites of each idea by word-sense divergence. We have shown in subsubsection 5.2.1 that a higher word-sense divergence can lead to a higher perceived innovativeness, originality, and overall creativity of the idea. Therefore, we choose the rewrite with the highest word-sense-divergence for each of the 10 human-generated ideas.[3] Using a matched t-test, we find that the best GPT4 rewrites have a significantly higher word-sense divergence than human-generated ideas.

We recruit 300 participants on *Prolific.com* to rate each of the 20 ideas (10 human-generated ideas and 10 best GPT4 rewrites) on innovativeness, originality, and usefulness. We randomly assign one idea to each participant.

The identification strategy in this experiment is straightforward. Given that we use a random rater assignment to ideas, the effect we aim to identify is the average treatment effect of using GPT4 on the perceived creativity measures of each idea. Therefore, we use the idea ID as a fixed effect, the calculate the effect of using GPT4 on innovativeness, originality, and usefulness.

[3]We manually checked each of the 10 rewrites to ensure that they are not changing the original human-generated ideas at a substantive level.

For rater i rating on creativity measure k of text j ($j \in \{1, 2, ..., 20\}$) and idea ID m ($\{1, 2, ..., 10\}$)

$$R_{i,j,k,m} = \tau 1(\text{text j is rewritten by GPT4}) + \text{FE}^m + \epsilon_{i,j,k} \tag{8}$$

where τ is the effect of using GPT4 to rewrite human-generated ideas on perceived creativity and FE^m is the fixed effect of the mth idea.[4]

Table 12: Effect of using GPT4 to rewrite human-generated ideas

	Dependent variable:			
	innovative	original	useful	creativity
	(1)	(2)	(3)	(4)
GPT4 rewrite	0.174*	0.234**	-0.001	0.181**
	(0.108)	(0.105)	(0.107)	(0.106)
Observations	298	298	298	298
R^2	0.207	0.249	0.222	0.238
Adjusted R^2	0.179	0.223	0.195	0.211
Residual Std. Error (df = 287)	0.906	0.882	0.897	0.888
Note:				*p<0.1; **p<0.05; ***p<0.01

As shown in table 12, using GPT4 to rewrite human-generated ideas significantly increases the perceived originality and overall creativity of these ideas. To put these numbers in perspective: using GPT4 to rewrite product design ideas can improve their perceived innovativeness by about 0.17 standard deviation, originality by 0.23 standard deviation, and overall creativity by 0.18 standard deviation. This result aligns with our result in subsubsection 5.2.1 where innovativeness, originality, and overall creativity are significantly positively affected by word-sense divergence. This is unsurprising because, in this study, we fix the substantive ideas, so the difference is driven by the difference in form level differences between GPT4 and human writing.[5] Given the results in sections 6 and 7, our estimates suggest that this form-level improvement would be more effective for product categories that have a nontrivial proportion of males in their target consumer group, and this approach is likely to work better at improving lower quality (creativity) ideas than ones

[4]Note that here we do not use a rater random effect because each rater is only asked to rate one idea (either original or rewritten), and ideas are randomly assigned to raters.

[5]We also asked GPT4 to generate "less creative ideas". The resulting average word-sense divergence of these ideas is lower than the "more creative" ones but still higher than human original ideas.

that are already highly creative.

9 Conclusion

In this paper, we demonstrated GPT4's ability to generate creative ideas that are rated as more creative than those generated by human creative professionals incentivized for performance. Furthermore, we showed that the model's superior performance can be attributed not only to creativity in form (i.e., using more unusual language), but also to creativity in substance (i.e., the ideas themselves are more creative). Nonetheless, the most powerful applications of these tools are likely to be in conjunction with humans, insofar as the most creative humans can still outperform current LLMs and their ideas can be further enhanced by LLMs.

These findings contribute to prior work in two primary ways. First, existing research has compared AI creativity only to laypeople or to human-AI teams; ours is the first, to our knowledge, to compare AI creativity to that of specialized and incentivized creative professionals. Using this more demanding benchmark allows us to provide stronger evidence of AI's creative capabilities. Second, prior research has been unable to disentangle creativity in form and creativity in substance; our novel modeling strategy provides both a methodological and theoretical contribution by enabling this disentanglement and showing that current LLMs can outperform humans on both types of creativity.

The most significant limitation facing this work is that LLMs continue to evolve and the exact pattern of results we observe may therefore change as future versions of LLMs become available. Nevertheless, these models are expected to become more capable and more human-like in the future, such that any changes to our results are likely to favor AI over humans to an even greater degree than we currently observe. Another limitation is that we were unable to sample ideas from the very highest echelons of human creativity. While we were able to collect ideas from specialized product designers incentivized for strong creative performance, even more creative humans – such as best-selling novelists, prolific inventors, or award-winning scientists – are likely to still retain an advantage over current LLMs. Relatedly, our research was limited to the context of new product ideation, such that we cannot conclude anything regarding AI's creative performance in other domains such as artistic production, creating advertisements, or generating research ideas.

Future research on AI creativity is therefore warranted. Testing AI's capabilities relative to the most creative humans in multiple domains remains an important task. Another important direction for future research is to measure the economic impacts of AI creativity: do AI-generated creative ideas outperform those generated by humans on measures such as click-through rates (i.e., for digital advertising) or even sales (i.e., for new product designs)? Such incentive-compatible behavioral outcomes would provide stronger evidence of AI's creative potential to meaningfully improve marketing outcomes.

References

Abraham, Anna, Kristin Thybusch, Karoline Pieritz, and Christiane Hermann (2014), "Gender differences in creative thinking: behavioral and fMRI findings," *Brain imaging and behavior*, 8, 39–51.

Amabile, Teresa M et al. (1988), "A model of creativity and innovation in organizations," *Research in organizational behavior*, 10 (1), 123–167.

Brodherson, Marc, Jason Heller, Jesko Perrey, and David Remley (2017), "Creativity's bottom line: How winning companies turn creativity into business value and growth," *Digital McKinsey. Available online: https://www. mckinsey. com/capabilities/mckinseydigital/our-insights/creativitys-bottom-line-how-winning-companies-turn-creativity-into-business-value-and-growth (accessed on 12 June 2022)*.

Burroughs, James E, Darren W Dahl, C Page Moreau, Amitava Chattopadhyay, and Gerald J Gorn (2011), "Facilitating and rewarding creativity during new product development," *Journal of Marketing*, 75 (4), 53–67.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock (2023), "Gpts are gpts: An early look at the labor market impact potential of large language models," *arXiv preprint arXiv:2303.10130*.

Frey, Carl Benedikt and Michael A Osborne (2017), "The future of employment: How susceptible are jobs to computerisation?," *Technological forecasting and social change*, 114, 254–280.

Guilford, Joy Paul (1950), "Fundamental statistics in psychology and education,".

Guzik, Erik E, Christian Byrge, and Christian Gilde (2023), "The originality of machines: AI takes the Torrance Test," *Journal of Creativity*, 33 (3), 100065.

Hitsuwari, Jimpei, Yoshiyuki Ueda, Woojin Yun, and Michio Nomura (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.

Institute, Mckinsey Global (2018), "Skill shift: Automation and the future of the workforce,".

Jia, Nan, Xueming Luo, Zheng Fang, and Chengcheng Liao (2023), "When and how artificial intelligence augments employee creativity," *Academy of Management Journal*, (ja).

Josten, Cecily and Grace Lordan (2022), "Automation and the changing nature of work," *Plos one*, 17 (5), e0266326.

Koivisto, Mika and Simone Grassini (2023), "Best humans still outperform artificial intelligence in a creative divergent thinking task," *Scientific reports*, 13 (1), 13601.

Li, Peiyao, Noah Castelo, Zsolt Katona, and Miklos Sarvary (2024), "Frontiers: Determining the Validity of Large Language Models for Automated Perceptual Analysis," *Marketing Science*.

Marrone, Rebecca, Victoria Taddeo, and Gillian Hill (2022), "Creativity and artificial intelligence—A student perspective," *Journal of Intelligence*, 10 (3), 65.

OpenAI (2023), "GPT4 Technical Report,".

PricewaterhouseCoopers (2017), "20th CEO survey,".

Reisenbichler, Martin, Thomas Reutterer, David A Schweidel, and Daniel Dan (2022), "Frontiers: Supporting content marketing with natural language generation," *Marketing Science*, 41 (3), 441–452.

Smith, Robert E, Scott B MacKenzie, Xiaojing Yang, Laura M Buchholz, and William K Darley (2007), "Modeling the determinants and effects of creativity in advertising," *Marketing science*, 26 (6), 819–833.

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

Yang, Xiaojing and Robert E Smith (2009), "Beyond attention effects: Modeling the persuasive and emotional effects of advertising creativity," *Marketing Science*, 28 (5), 935–949.

Appendix

A Using GPT4 to rate product design ideas

We have shown in section 8 that we can use GPT4 and word-sense divergence to generate more creative rewrites of human-generated ideas. In this section, we test whether we can use GPT4 to generate creativity ratings that are similar to human-generated ratings. We use the 10 human-generated ideas and 10 GPT4-rewritten ideas in section 8. In addition, we use two types of agreement measures to compare GPT4 and human-generated ratings: the (continuous) Pearson correlation test and the (discrete) triplet method developed in **Add citation**. Before conducting these tests, we first compute $\bar{r}_{i,j,k}$ the average rating of each product design idea $i \in \{1, 2, ..., 10\}$, each perceived creativity factor $j \in \{\text{innovativeness}, \text{originality}, \text{usefulness}, \text{creativity}\}$, and each source $k \in \{\text{human}, \text{GPT4}\}$. We compute the Pearson correlation between human and GPT4-generated ratings as

$$\rho = \frac{\text{cov}_j(\bar{r}_{\text{human}}, \bar{r}_{\text{GPT4}})}{\sigma_{j,\text{human}}, \sigma_{j,\text{GPT4}}} \tag{9}$$

where cov_j is the covariance over average ratings of factor j from human versus GPT4, $\sigma_{j,\text{human}}$ is the standard deviation of average ratings of factor j from human, and $\sigma_{j,\text{GPT4}}$ is the standard deviation of average ratings of factor j from GPT4.

In addition, the adjusted triplet agreement rate between human and GPT4-generated ratings can computed as

$$\text{AR}_{\text{adj}}(\text{human}, \text{GPT4}) = \frac{\text{AR}(\text{human}, \text{GPT4})}{\text{SAR}_{\text{adj}}(\text{human}, \text{GPT4})} \tag{10}$$

where $\text{AR}(\text{human}, \text{GPT4})$ is the raw agreement rate between human and GPT4 generated ratings and $\text{SAR}_{\text{adj}}(\text{human}, \text{GPT4})$ is the average of the self-consistency rates of human and GPT4 generated ratings. where the agreement rate $\text{AR}(\text{human}, \text{GPT4})$ between human and GPT4 ratings is the probability that given the ratings for any two product design ideas a and b on any creativity factor j, both human and GPT4 rates a higher than b or b higher than a. The average self-consistency rate $\text{SAR}_{\text{adj}}(\text{human}, \text{GPT4})$ is the average of the self-consistency rates of the human ratings and GPT4

ratings. The self-consistency rate of a set of ratings can be computed by randomly drawing N pairs of equal-sized samples from the ratings with replacement and computing the average agreement rate between all pairs of samples[6].

Table 13: Consistency between human and GPT4-generated creativity ratings

	Pearson correlation	Adjusted triplet agreement rate
Innovativeness	0.616***	0.761***
Originality	0.703***	0.691***
Usefulness	0.780***	0.833***
Creativity	0.497***	0.608***
Overall	0.814***	0.721***

Correlation significance: *p<0.1; **p<0.05; ***p<0.01

As shown in table 13, both the Pearson correlations and adjusted triplet agreement rates are highly significant relative to a null hypothesis that human and GPT4-generated ratings are uncorrelated. In the case of Pearson correlation, two uncorrelated sets would have a correlation of 0. In the case of adjusted triplet agreement rates for two highly self-consistent data sets (SAR = 0.9) would have an agreement rate of 0.56. The significance of the adjusted triplet agreement rates is computed by bootstrapping the GPT and human-generated ratings each 500 times. These results show that GPT4 perceived creativity ratings of product design ideas are highly correlated with human-generated ones; therefore, given a set of product design ideas, practitioners can use GPT4 as a pre-filter: they can generate creativity ratings using GPT4 and eliminate ideas that are likely to be perceived as creative.

B Using GPT4 to rate form-level creativity (word differences)

As a sanity check, we explore whether GPT4's interpretation of creativity is correlated with its understanding of word-level creativity. To do this, we use GPT4 to generate ratings for word-level creativity of product design ideas. More specifically, we use the prompt "You are a participant in a consumer survey. Your job is to rate the following pitches of product design ideas on the creativity of their use of words. Do not consider the creativity of the ideas, only focus on how the idea is described in words. Please rate each idea with just an integer from 1 (least creative) to 5 (most

[6]We choose N=500 in our study

creative), do not use any words: {*TEXT OF IDEA*}". We collect 30 ratings for each GPT4 and human-generated idea studied in Appendix A, and compute the Kendall-Tau ranked correlation between the AI ratings for creativity and word differences. As shown in table 14, we find that GPT4's ratings of innovativeness, originality, and overall creativity are moderately correlated with its ratings of word-level creativity.

Table 14: Correlation between GPT4-generated word-level creativity rating and overall creativity rating.

	Pearson correlation
Innovativeness	0.267*
Originality	0.140*
Usefulness	−0.102
Creativity	0.239*

Correlation significance: *p<0.1; **p<0.05; ***p<0.01