

Expertise elevates AI usage: experimental evidence comparing laypeople and professional artists

Thomas F. Eisenmann^{*1} Andres Karjus^{*2,3} Mar Canet Sola^{1,2,4} Levin Brinkmann¹

Bramantyo Ibrahim Supriyatno¹ Iyad Rahwan¹

¹Center for Humans and Machines, Max Planck Institute for Human Development

²Tallinn University ³Estonian Business School ⁴Academy of Media Arts Cologne

*shared first authorship

Abstract. Novel capacities of generative AI to analyze and generate cultural artifacts raise inevitable questions about the nature and value of artistic education and human expertise. Has AI already leveled the playing field between professional artists and laypeople, or do trained artistic expressive capacity, curation skills and experience instead enhance the ability to use these new tools? In this pre-registered study, we conduct experimental comparisons between 50 active artists and a demographically matched sample of laypeople. We designed two tasks to approximate artistic practice for testing their capabilities in both faithful and creative image creation: replicating a reference image, and moving as far away as possible from it. We developed a bespoke platform where participants used a modern text-to-image model to complete both tasks. We also collected and compared participants' sentiments towards AI. On average, artists produced more faithful and creative outputs than their lay counterparts, although only by a small margin. While AI may ease content creation, professional expertise is still valuable — even within the confined space of generative AI itself. Finally, we also explored how well an exemplary vision-capable large language model (GPT-4o) would complete the same tasks, if given the role of an image generation agent, and found it performed on par in copying but outperformed even artists in the creative task. The very best results were still produced by humans in both tasks. These outcomes highlight the importance of integrating artistic skills with AI training to prepare artists and other visual professionals for a technologically evolving landscape. We see a potential in collaborative synergy with generative AI, which could reshape creative industries and education in the arts.

1 Introduction

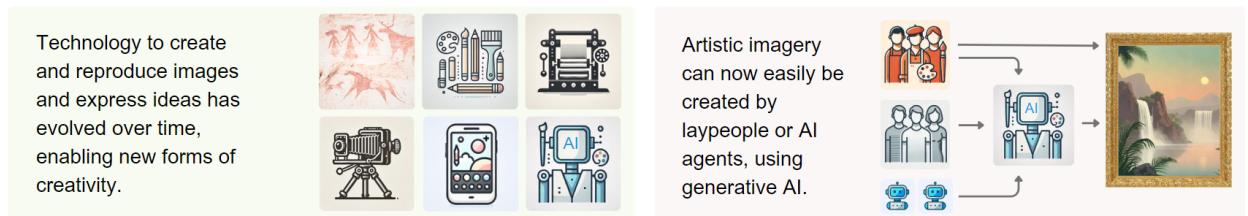
Generative machine learning models' increasing prevalence and capacities are transforming creative processes. Large language model-driven text generators and chatbots like ChatGPT or Copilot enable native speaker-like text production. Generative image models and services such as Stable Diffusion, Dall-E or Midjourney enable the creation of artistic, photo-realistic, and illustrative visual materials without necessarily having professional training in these fields. Such tools are also rapidly being integrated into word processors and graphic design software like Canva or Adobe Photoshop. By far the most common interface to these generative artificial intelligence models is natural language instructions or "prompting". The entry threshold is thus very low: only a minimal command of the input language(s), typically English, is required to get a given app or model to start generating artistic materials or grammatically coherent text. Generative AI usage typically combines both creation and curation, as it is easy—and often the default workflow in many applications—to produce concurrent variants based on a single input, and choose the most suitable output.

It has been shown that current generation LLMs are not only theoretically capable of creativity (Wang et al. 2024), but already out-performing humans in various creative writing and divergent thinking tasks (Bellemare-Pepin et al. 2024; Hubert et al. 2024; Mirowski et al. 2024). This is in addition to more formal tasks such as various forms of textual annotation and analysis, where similar results have emerged (Ziems et al. 2023; Törnberg 2023; Gilardi et al. 2023; Karjus 2023). Naturally, it depends on which humans exactly the machines are being compared to, the average or the top performers, experts or laypeople (Koivisto and Grassini 2023; Haase and Hanel 2023; Porter and Machery 2024). Other work has reported humans as being more imaginative than LLMs (Beguš 2024), or LLMs being more creative but less diverse (Doshi and Hauser 2024). Furthermore, general creativity as such is notoriously difficult to define and measure (Miravete and Tricot 2024). Instructing generative text and image models has been argued to be a skill in itself (to the point of being called "prompt engineering"), as informative inputs that are well aligned with the pre-trained "expectations" of the model tend to produce superior results (Wei et al. 2022; Liu and Chilton 2022; Oppenlaender et

al. 2024). Additionally, it has been shown that prompt content continues to matter even as generative models improve (Jahani et al. 2024). Image generators have been assessed in several studies highlighting their usefulness as supporting creativity (Sáez-Velasco et al. 2024; Gu et al. 2024; Braguez 2023), and such usage likely benefits from relevant skills.

Here we focus on text-to-image models and the domain of visual art, broadly construed. The aesthetics, value, and quality of the outputs of these tools may be criticized, but their usage is widespread and likely increasing (von Garrel and Mayer 2023; Shen et al. 2023; Walkowiak and Potts 2024). While people may prefer (at least the idea of) naturally produced art (Bellaiche et al. 2023), artificially generated content is becoming increasingly difficult, if not impossible, to detect, especially for the untrained eye and ear (Lu et al. 2024; Frank et al. 2023; Cooke et al. 2024; Porter and Machery 2024). This adoption of such easy content creation technologies across various professions is blurring the lines between professional artists and hobbyists. It also inevitably raises questions about the nature of and need for expertise and the role of AI therein. Professional artists and amateurs may also use AI tools simply for different ends, the former to improve art quality and try new styles, the latter for e.g. entertainment and exploration (Elfa et al. 2023; Braguez 2023; Shen et al. 2023). Most recent comparisons in this domain cited above have focused on laypeople; we aim to build on this with explicit comparisons with experts.

(A) Challenges and opportunities for artistic professions



(B) The experimental design

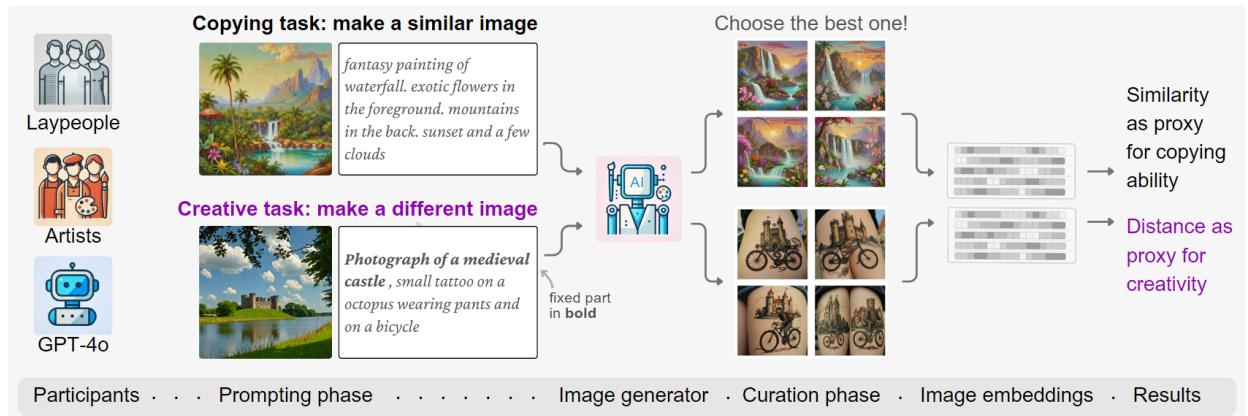


Figure 1. Motivation for the study (A) and pipeline overview (B). The participants were asked to view a reference image and write a prompt for a generative model to either create a similar image (copying task) or a maximally different image (creative task). In the curation phase, they were shown four generated variants and asked to select the most suitable one given the goal of the task. We later compared the similarity of the reference and generated images using an image embedding model, and used cosine vector similarity to operationalize the results. The prompts and variants are from one of the artist participants; the rest of the images created by this participant can be found in the Supplementary Information.

Naturally, these developments have raised various concerns from legal and moral issues around training data (Goetze 2024) to ethical practices of AI usage and its impact on the creative professions and labor markets (Lovato et al. 2024; Miyazaki et al. 2024; Walkowiak and Potts 2024). A recent study found a significant drop in job postings for writing and image creation jobs on online freelance platforms after the introduction of chatbots like ChatGPT and various image generation tools like Midjourney (Demirci et al. 2023). The rapid uptake of these technologies has caused various reactions in societies and the media, ranging from claims that "art is dead" (Roose 2022) and that "this changes everything" (Klein 2023), to reports of "AI anxiety" of workers fearing for their jobs (Cox 2023), and questions about "creators becoming redundant" (Dege 2023). On the other hand, AI adoption has been associated with gains in both artistic productivity and output novelty (Zhou and Lee 2024). Such tools clearly have the potential to change education and career paths in the arts, yet there is uncertainty about how the focus should be distributed between traditional art education, "AI prompting", and the synergy of the two.

Here we explore the hypothesis that expertise in a given domain should also lead to superior results in using AI, on the example of text-to-image prompting and the curation of its results (Figure 1). Professionals can be expected to be more creative, aware of artistic principles, better at commanding domain-relevant vocabulary to more accurately describe the expected outputs of a creative process, and skilled at curating among multiple generated alternatives. We aim to measure how much of an advantage this provides, if any, compared to layperson-users of the same AI tool, representing the general population. We devised two controlled tasks, emulating the artistic practices of replication or copying of existing art, and the creative production of novel or divergent art (see Figure 1 and Methods for details). Specifically, we test the following preregistered hypotheses:

H1: Artists' images in the copying task will overall be closer to the original images than laypeople's.

H2: Artists' images in the creative task will overall be more distant from the original images than laypeople's.

H3: Artists' curated images in the copying task will be closer to the original images than laypeople's.

H4: Artists' curated images in the creative task will be more distant from the original images than laypeople's.

H5: Curation independent from prompting: Artists will consistently choose more suitable images from the selection of four images in the pooled data of both tasks, compared to laypeople.

Additionally, we explore a comparison of the human groups to the outputs of a similarly instructed AI language model in an approach comparable to H1-2.

2 Methods and materials

The current experimental design draws from several experimental traditions. These include interactive behavioral experiments often used in cognitive science and adjacent disciplines (Kirby et al. 2008; Okada and Ishibashi 2017; Nölle et al. 2018; Müller et al. 2019; Karjus et al. 2021; Kim et al. 2024), research comparing the behavior or outcomes of domain professionals or experts with some control group of laypeople (Kozbelt 2001; Bhattacharya and Petsche 2005; Bezruczko and Schroeder 1994; Torngren and Montgomery 2004), and studies on individual aesthetic perception and preferences (Porter and Machery 2024; Cela-Conde et al. 2009; Lakhal et al. 2020). The online experiment was carried out with two participant groups, asked to complete two distinct tasks, which we call "copying" and "creative" for short. The tasks were divided into "prompting" and "curation" sub-tasks or phases, and each task contained four such image generation and curation trials. Figure 1.B illustrates the pipeline from experimental tasks to results. The sections below describe the details of the design, sample, procedure, and data analysis.

2.1 Participants

Our final sample consists of 99 participants, 50 professional visual artists, and a matched control group of 49 laypeople recruited via the crowdsourcing platform Prolific. The artist sample was recruited first, via invitation emails sent directly through the coauthors' (mostly MS) personal and professional networks. This way, we ensured artists fulfilled our pre-registered recruitment criteria (see below) and avoided excessive over-recruiting and exclusion, although it was rather labor-intensive. The final artist sample resulted from a total of 213 invitations, distributed in rounds of 10 to 15 invitations per day over several weeks to control the load put on the Stable Diffusion model used in the experiment. After the full artist sample had been completed, we recruited the laypeople sample to match its demographics.

All participants, regardless of the subsample, were paid €3.75 after completing the full experiment, in compliance with the ethical approval received for the study beforehand (by the IRB of the Max Planck Institute for Human Development, approval number A2024-15). Participants not using € currency received the equivalent amount in their local currency instead. Artists were paid via direct transfers using Paypal, while laypeople received the money in their Prolific account. Some artists (22%) were more motivated by participating in a scientific experiment than by the remuneration itself and waived their payment.

All participants were required to be fluent in English (self-reporting native or near-native competence) but were allowed to have a different first language. Most importantly for the purpose of the study, the artist sample had to actually represent professionally working artists, whereas the laypeople sample must not, to ensure a clean separation of the two conditions. We made sure of this by recruiting the artists from our networks of professional artists, and double-checking that the Prolific sample did not include any. There was an item in the post-questionnaire to screen for participants' correct assignment to their condition ("Do you have work experience in visual art?"). If a participant's answer did not match their condition, they were excluded from the sample. This happened in 5 cases for artists (e.g. having art education but

not working professionally as an artist) and in 10 cases for laypeople (due to chance, because we could not initially screen them against that). We also inquired about the number of years the artists had been active in visual arts, which was high in our final sample ($M = 22.4$, $\sigma = 10.7$). We chose to focus on artists in professional careers, defined as those actively engaged in creating art and participating in the distribution system, such as exhibiting or selling their work. This has been called the most salient marker distinguishing serious artists from amateurs (Becker 1982). Thus, art students and academics not actively working in the field were excluded, to ensure a clean comparison between professionals who create art and laypeople who do not. We further narrowed our focus to *visual* artists, as this was the medium of the task.

While recruiting the artist sample, we recorded participants' highest level of education and their first language, identically to two screening items on Prolific. We then used the screening function on Prolific to match the participants faithfully, according to their specific combinations of education and language (binary English/not English). As expected, artists' highest level of education was very high, with 24 participants of the 50 stating they had a Master's degree and 23 more reporting a PhD. 14 artists reported being native speakers of English.

Apart from group assignment, we specified two more exclusion criteria in the preregistration. Participants who did not enter a prompt in two or more trials were replaced due to too much missing data. Images that were blanked out for participants due to our implemented filter (see below) were counted toward this limit. These criteria led to the exclusion of two more participants (one per subsample), who were immediately replaced. After data collection of all 100 participants and exclusion via the preregistered criteria had concluded, it became apparent that the prompting data of a single participant from the laypeople group had not been passed on due to a technical error (while the participant provided all information otherwise), and was thus unusable; hence we arrived at our final sample of 99 participants.

2.2 Experimental design

This monadic design consists of instructing each participant, working independently but within a time limit, to complete two multimodal tasks where they are shown a reference image and asked to write short textual inputs ("prompts") to produce new images and select their preferred output. We created a bespoke online platform for the experiment, aimed to broadly mimic currently common generative AI apps and platforms. The experiment started with an instructions and consent page, followed by two training trials where the participants were free to generate images from their own prompts to familiarize themselves with our generator, the interface, and the prompting and curation phases. They were then instructed in the first task, completed it, got instructed in the second task, and finally were asked to complete a short questionnaire. In total, the average participant took about 18 minutes to complete the experiment.

Participants completed the copying task first and creative task second. There were 4 trials within each task which displayed a unique reference image. The order of images was randomized within each task. Each of the 8 trials consisted of a "prompting phase" and a "curation phase". The interfaces were the same for both tasks. In the prompting phase, participants were tasked with writing prompts to create images from. Here, the interface showed the unique reference image, a text entry box, and a "submit" button. Prompt length per trial was limited to 30 (space-separated) words, visible as a counter below the text box. Furthermore, there was a time limit of 2 minutes per attempt, which was also displayed. If time ran out before participants had submitted their prompt, their current input was submitted and they moved to the curation phase automatically.

After prompting, our generative model created 4 different images for the participant to choose from in the curation phase. The reasoning behind this was threefold. First, this mimics real-world generative AI usage, where tools often provide multiple variants by default. It also balances the randomness inherent in different seeds (see below). Finally, this enables testing whether curating would also allow for the expression of artistic expertise, as per our second set of hypotheses. The interface showed the 4 generated images arranged horizontally, with the reference image displayed on top for comparison. Here, the time limit was 30 seconds, as participants only needed to click their preferred image to submit. If time ran out, the images were stored without curation data.

In the copying task, participants were asked to prompt the model to generate an image "as close as possible" to the reference image. The goal of the task is to compare the ability of different groups to produce faithful copies of an example, demonstrating their familiarity with a given genre (see Stimuli below), ability to envision and describe a subject, and an eye for relevant details. This is meant to partially approximate traditional art education practices that also emphasize honing skills by copying masterpieces to internalize techniques and styles.

In the creative task, participants were instead asked to make a new image that would be "as different from the original

as possible". Their instructions further clarified that 1) ideally, they should aim for "an image with different **content** as well as different **visuals**", and that 2) negation does not work well for this and would actually increase the likelihood of adding something to the image. To make sure results remain comparable and the task reasonably challenging, the start of the prompt was (visibly) fixed in the text box to the first words used to create the reference image. For example, if the prompt we used to create the image was "A photo of a classic gray Mini Cooper in a parking lot of a shopping mall", the text box would contain the locked starting phrase "A photo of a gray classic Mini Cooper". This task was designed to measure creativity in addition to the descriptive skills of the first task, as it requires the addition of something novel and divergent, not just reproduction. As such, professional artists would be expected to excel. As discussed in the Introduction, measuring creativity is challenging, but we nevertheless propose ways to compare the groups as described below. An alternative for this task we considered would have been to instruct them to prompt for an image that would be just "more creative". We opted for the more indirect task of creating a different image, as it arguably still requires expressing creativity, while being much more straightforward to operationalize than measuring what is "more" and "less" creative (see below).

2.3 The generative model, stimuli, and data

The text to image model in the backend was Stable Diffusion XL Turbo (henceforth SD), chosen for its ability to generate images within seconds (Sauer et al. 2023). As with all deep neural network-based models of that nature, SD has a random seed hyperparameter, which affects the outputs. To account for this inherent randomness, we used four fixed seeds to generate the variants that participants curated from. These seeds were distinct from the seeds used to create the reference images, so participants could not reproduce the reference images even if they found the exact corresponding prompt. This also makes the outputs comparable: as all parameters are fixed for all users, if two participants entered the exact same prompt, they would get the exact same set of four images. As for other relevant hyperparameter, the inference steps was set to 4 to ensure sufficiently fast image generation. We implemented a keyword-based filter to prevent users from creating sensitive or obscene images. The filter worked by comparing the CLIP embedding of the image to the embeddings of the filter words (Rando et al. 2022). If such an image was detected, it was not shown to the participant in the following curation step. This was for ethical concerns and because creating such images was not instrumental to the task. Participants were informed about the filter and asked to avoid creating inappropriate content. They were also instructed not to include any personal details in their prompts.

The eight reference images were created using the same SD model by our professional artist coauthor (MS) to ensure sufficient quality and task difficulty. The number of images and thus trials was small enough to keep the experiment manageable and within budget constraints. We aimed to cover a diverse range of artistic imagery, subjects, and styles: landscapes, architecture, vehicles, food, humans, and cartoon characters; and photographic, animation, painting, drawing, and video game 3D styles. Additionally, we kept the original prompts in the copying task relatively complex and the ones in the creative task relatively plain; this was to avoid ceiling or floor effects.

We have published all the image data resulting from the experiments (see Data Availability) and also produced an interactive dashboard based on the Collection Space Navigator (Ohm et al. 2023) that enables easy exploration of the dataset, available here: <https://artistlaypeopleaiexperiment.github.io>.

2.4 Measuring the results

After data collection and manual inspection, we identified and removed four trials that we deemed ineligible. These included errors like a single-letter prompt, two cases where participants seemed to have initially misunderstood the task (but carried on as expected afterward), and one case where the fixed lead of a creative trial merged with the input due to a bug in the interface (the data is still available in the open code and database accompanying this paper). We came across 11 occasions where participants had apparently circumvented the whitespace-operationalized word limit by concatenating a few words at the end of the prompt, but did not deem this an exclusion criterion. While clever, omitting spaces carries the risk of confusing the underlying text tokenizer of the model and therefore unexpected visual outputs.

The resulting final dataset consists of 3148 images. As preregistered, we used the CLIP (clip-vit-base-patch16) embeddings (Radford et al. 2021) to quantify the similarity between reference images and participant creations, via cosine similarity of the embedded image vectors. SD models are known to use CLIP internally for encoding prompts, making this a natural choice. In the copying task, the goal was to create a close replication of the reference image — therefore,

higher cosine similarity indicates a better match. In the creative task, where the goal was to create a different image, lower cosine indicates a better outcome. For comparing groups in H1 and H2, the values were averaged for each set of four variant images (which on average have high inter-similarity), as we are more interested in prompt outcomes rather than individual images.

An image embedding is by no means an absolute or objective measure of visual similarity (if such a thing exists), but rather a practical solution for comparing thousands of images using a consistent metric. Based on a limited manual evaluation of piloting data, CLIP appears to capture similarity and dissimilarity well enough for our purposes, both in terms of the objects and style of an image. Importantly, we do not use embedding vector similarity as a measure of creativity or artistic skill, but rather as a way to measure behavioral outcomes in a task that is adjacent to or emulates artistic practice (although not entirely ecologically valid, as is typical for most, if not all, such experiments).

While not the main object of study, we also measured several summary statistics to gain some insight into the data: image colorfulness, complexity, and prompt length (in characters). For color, we used the "M3" measure from Hasler and Suesstrunk (2003), and for complexity the file size of the PNG compressed image. Compression is a well-known estimate of image complexity that aligns with human perceptions (Machado et al. 2015; Chamorro-Posada 2016; Karjus et al. 2023). We find no notable differences between groups along these axes. We use self-reported AI usage experience levels as a control variable in the Results section; artist participants were more experienced (mean 1.68) than laypeople (1.04, on a scale of 0 to 3). A graph visualizing all the aforementioned metadata and summary variables is found in the Supplementary Information.

We also wanted to make sure the creative task could not be solved successfully by simply filling the prompt with random words or gibberish, which might conceivably lead to images different from the reference. We ran a small simulation to test that, by producing a set of creative task prompts of maximal length (including the fixed prefixes) out of random words sampled from an English word list, and another set consisting of pseudo-words of randomly sampled letters, and generating new images based on those. We found that while it is possible to occasionally "get lucky" with this strategy, adding random things to a prefixed prompt generally leads to images still close to the reference image, and the simulated results were on average far worse than the real data (see SI for details).

2.5 Bringing in an AI

Inspired by the growing literature on comparing humans and various AI agents discussed in the Introduction, we also carried out a small comparison with an "AI", the vision-capable large language model GPT-4o by OpenAI (specifically gpt-4o-2024-08-06, via its API service). The LLM was given roughly the same instructions as the human participants in the experiments, with added context that was otherwise implicit in the task interface (such as the word limit), and general guidance to act in the role of a creative professional. It was instructed to write a prompt of up to 30 words just like human participants. Any words above the limit would be clipped to ensure comparability with the rest of the experiment (there were only 3 such occasions, however). In short, we created something akin to a simple image evaluation and generation agent, consisting of three models: the LLM interpreting the input and producing the prompt, the image generating SD, and the CLIP model yielding embeddings for our goal of image comparison. Unlike participants, the LLM here was not granted a "memory" of past completed attempts. While in principle achievable by feeding previous inputs and outputs into the context window of the model, it is unclear to what extent this would be comparable to human memory and is not explored here.

The procedure was as follows: for each trial, the input was the reference image and the instructions to carry out the given task, e.g., to write a prompt that would generate an image like the reference forest image. The model was set to generate 10 output variations on each trial (by setting the relevant parameter in the API). The prompts were subsequently entered into the same image generator used in the experiment, with the same parameters, to produce images that could then each be compared to the reference, as described above. Here the goal was neither model comparison nor parameter space exploration; the relevant parameter to set was temperature, which governs token sampling in LLMs. 0 temperature means the model always picks the most likely next word to generate (precluding generating variations from the same input), while higher values lead to more stochastic and unexpected outputs, and maximal values can end up generating nonsense. We set it at 0.7 (on a scale of 0 to 2) for the copying task, to sample variation while still being relatively precise, and 1.5 for the creative task, simply following Bellemare-Pépin et al. (2024) who observed that higher temperatures indeed help LLMs in creative association and semantics tasks.

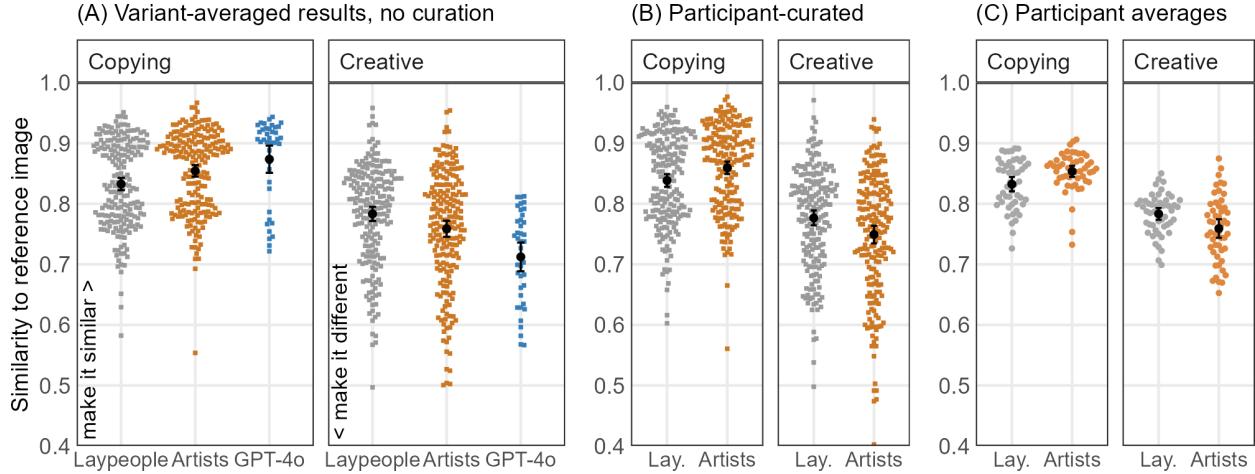


Figure 2. Experiment results: artists are (slightly) better at using generative AI than laypeople. The vertical axis is cosine similarity; for the copying task, the goal was to produce similar images (high values), and for creative, dissimilar. The panels display three views of the data. Each square in (A) represents the distance of a generated image (averaged across its four variants) from the reference. The bimodal distributions in the copying task arise from the apparent variable difficulty in copying the reference images (this is controlled for in the statistical modeling; see also extended graph in the SI). Panel (B) shows only image variants selected by participants in the curation phase. Additionally, panel (C) shows participants (round points), averaged across all their (non-curated) outputs. Black bars are means with 95% confidence intervals for reference.

3 Results

We use mixed effects generalized linear regression to analyze the differences between the outputs of our two groups, in the following form: cosine ~ group + experience + (1|subject) + (1|item). In both tasks, the dependent variable is the cosine similarity between the original and generated image. The main variable of interest is the group: the coefficient value of this shows how much the groups differ in terms of outcomes. We also control for participant background by fitting a fixed effect for (self-reported, numeric) prior AI usage experience, and fit random intercepts for stimuli items and participants to account for individual-level variation. As discussed above, we use the averaged distance of the four generated variants as the unit of data here, yielding 787 cases.

We find support for both hypotheses 1 and 2. In the copying task, the estimated cosine similarity of laypeople is -0.03 lower than that of artists (the 95% confidence interval is between [-0.04, -0.01]). The effect is significant (at $\alpha = 0.05$), as determined via a likelihood ratio test comparing the full model to a partial one without the group variable ($\chi^2 = 11, p < 0.001$). In the creative task, laypeople are 0.02 [0.002, 0.04] closer to the reference than artists — here the goal is to be distant, so the artists' result is superior ($\chi^2 = 4.56, p = 0.03$). The coefficient confidence intervals were estimated using bootstrapping with 1000 model replicates. In the latter task, the lower bound only narrowly excludes zero. In summary, while narrowly statistically significant, these are not very large differences, as also visible in Figure 2.A. The reference images of both tasks and the top results are visualized in Figure 3.

3.1 Curation effects

The experiment combined both artistic image creation and curation, as participants were presented with four variant images to choose from after every generation trial. We can therefore assess the effect of curation as well. We constructed two sets of models to do so. Here we excluded the few trials that produced NSFW content, as the participants were not given a chance to curate those (blank images were displayed), leaving 758 cases. In short, we found no differential effect of curation between groups. First, we ran the same mixed effects linear models as above on this dataset reflecting curation choices. For both tasks, it resulted in similarly small differences: laypeople were more distant than artists in copying (-0.03[-0.04, -0.01], $p = 0.001$) and more similar in the creative task (0.026[0.003, 0.05], $p = 0.02$), narrowly supporting H3 and H4.

To test H5, we also ran cumulative link mixed models (with logit link function), again also controlling for prior experience and using the same random effect structure. The dependent is a 4-level ordinal variable, reflecting the rank of the

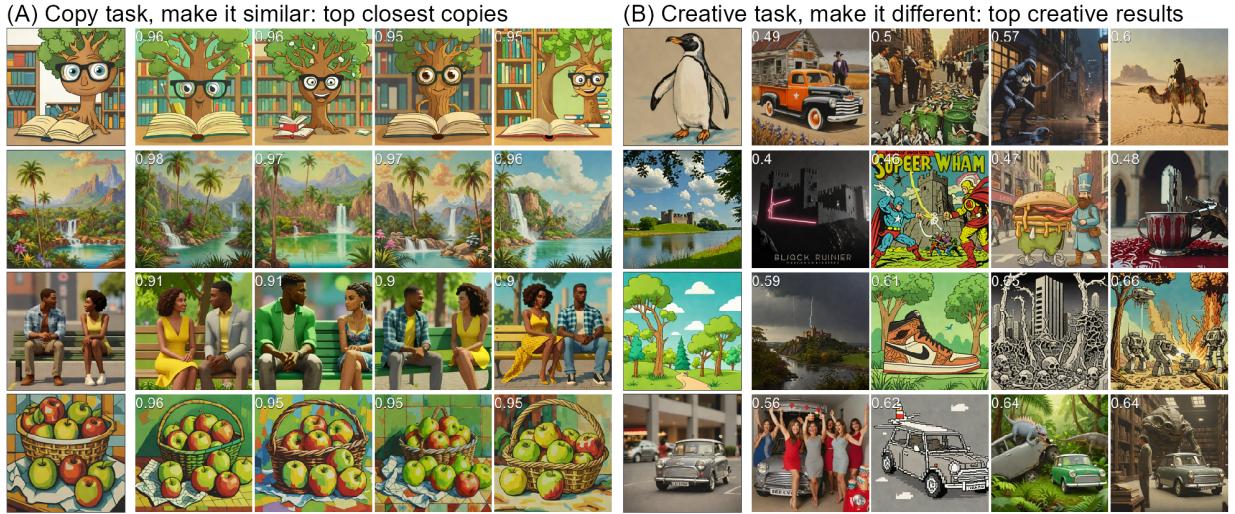


Figure 3. Examples of most successful trials across the experiment; cosine similarity in the corners. Panel (A): closest copies for each reference image (left, black border). (B): the most successful creative results, diverging furthest from the reference, either by managing to shift the style, hide the prefixed subject, or transform it. A larger version of this graph, comparing the best to the worst results, can be found in the appended Supplementary Information.

participant's choice in terms of the best option among the four, as measured by CLIP similarity (where 1 is the furthest, 4 is the closest). However, the p-values for the group variable were well above 0.05 in both the copying and creative tasks (0.7 and 0.4; and bootstrapped coefficient confidence intervals span 0), indicating no discernible difference between the abilities of artists and laypeople when it comes to choosing between image alternatives. This is of course a very limited emulation of curation, and in retrospect, most variant images were already highly similar, leaving little room for potential differences in curation skills to be expressed. A dedicated curation experiment with more dissimilar choices would have likely provided better insights and could be pursued in the future.

3.2 Comparing to AI performance

We also explore an additional comparison to GPT-4o, one of the frontier multimodal LLMs currently also powering the popular ChatGPT chatbot service. The statistical modeling solution here is less complex, as there is only one "participant" in the GPT group, and there is no comparable experience variable for a machine. We therefore fit a simple fixed-effects linear regression model, where the group variable now contains the GPT, set as the reference level (we fitted a random effects model as well which yielded very similar results). As in the first comparison, we average the variant image cosine similarities for each trial. In the copying task, there is no significant difference between GPT and artists ($p = 0.11$). Laypeople are an estimated -0.04 further from the reference (CI = $[-0.06, -0.02]$, $p < 0.001$; model adjusted $R^2 = 0.03$, $F(2, 429) = 8.03$, $p < 0.001$), meaning GPT did better here. In creative, where the goal is to have a lower value, both laypeople ($\beta = 0.07[0.04, 0.1]$, $p < 0.001$) and artists ($\beta = 0.05[0.02, 0.08]$, $p = 0.002$) are on average higher than GPT ($R^2 = 0.05$, $F(2, 432) = 11.94$, $p < 0.001$). This means GPT-4o was able to write prompts that led to on average more creative visual results than both human groups, within the narrow definition of the task. The variance described is very low in both models, however, less than 5%, corresponding to the rather small absolute differences, and the top results in each task are still achieved by the most successful human participants, mirroring results by Koivisto and Grassini (2023). Even if the advantage for GPT is small, the perhaps interesting takeaway here is that many professional artists with years of training and experience, not to mention laypeople, were *not* able to score much better in this (admittedly narrow) task than a simply next-word-predicting large language model.

Artists collectively generated slightly more diversity or variability in the creative task than both laypeople and GPT-4o. We calculated this by first measuring variation in the images and then fitting a linear mixed-effects model. Variation is implemented as the cosine similarity of each image embedding vector to its group centroid (average) vector, computed for each trial image and group (as images differ, and we mean to compare groups). This is analogous to mean absolute deviation or MAD. The regression then measures the effect of the group on the dependent variable of the variation metric (with random effects for group and trial reference image). The unit is still cosine, so higher values here indicate closer proxim-

ity to the center, i.e. lower variability. Laypeople ($\beta = 0.02$, CI = [0.003, 0.03]) and GPT-4o ($\beta = 0.05$, [0.001, 0.11]) have both higher cosine, i.e. lower collective diversity than the reference level of artists (model $p = 0.01$ compared to reduced model). The coefficients are again quite small (for reference, see the vertical axis of Figure 2 which is on the same cosine similarity scale).

Another possible approach to measure diversity would be within participants: how much the four images (and their variants) differ from each other on average. In the creative task, some participants (and GPT) used similar strategies for all four trials, while others varied (see the example image sets in the SI for intuition). We measure this as the average of all pairwise embedding similarities within a given participant. For GPT, we simulate subjects by replicating the dataset 10 times, randomly recombining the aforementioned 10 image prompt-variant sets (keeping the output images of a prompt intact). Here the statistical model is just simple regression, as trials and participants are already averaged. The artists have an estimated average intra-similarity of 0.63 (still on the same cosine scale as above), and laypeople do not significantly differ ($p = 0.91$), while GPT yields higher estimated similarity, i.e. less diversity than artists ($\beta = 0.07$, $p < 0.001$, model $F(2, 196) = 155.8$, $p < 0.001$).

It should be noted that the LLM "participant" was at a disadvantage here in terms of diverse outputs. We used only a single instruction prompt per task (see Supplementary), so the inputs differed only in terms of the input image, and in the case of the creative task, the explanation of the fixed prefix. This being a generative model highly sensitive to the output, using a larger array of different instruction prompts may well have increased output variance as well (e.g. by prompting for various types of artists, different styles, and approaches to undertaking such tasks).

3.3 Exploring the results: how do people (and machines) use an image generator?

Figure 3 displays the most successful generations in the two tasks. For the copying task, precise and relevant wording naturally worked the best, e.g. the top result for the jungle image was prompted as "a painting of an idyllic landscape, lake, mountains, waterfalls, palm trees, matte painting, detailed, tropical, jungle, inspired by mark keathley". Some humorous prompts but still worked fairly well, for example, "a bad Cezanne style painting of green and red apples in a baskek with ceramic tiles behind it" resulted in an average 0.93 similarity (the typo may have affected the outcome). While the copying task had participants mostly just trying to describe the scenes as best they could, the creative task elicited various interesting strategies.

Some participants either knew or realized that text to image generators can be confused by compounds. Both of these prompts made the output diverge from the reference: "*Photograph of a medieval castle* caricature walking down a busy new york street eating a hotdog. *pastel colours*" (third best castle result), and "*Painting of a penguin* colored pickup truck in rural America. The truck is black, white, and orange, but the driver is dressed up as the Joker. *Photorealistic*" (the best penguin result; the locked prefixes in italics). This did not always work, however, e.g. in "*A cartoon landscape with trees* bark, texture picture". Making use of the word allowance to describe a very different style or objects often worked: "*Photograph of a medieval castle* in black and white, neon lights, blade runner, 4k unreal engine, octane. Night-time blurred motion lights" (best castle result) or "*A photo of a gray classic Mini Cooper* is in the closet of Maria. She wears a red dress and her 20 friends are with her celebrating her birthday." (top car result). An attempt to hide the penguin in a box also worked quite well, as did an idea to transform the castle into a tattoo (for these further examples, see the SI).

In terms of image embedding similarity, there was little to no difference between human diversity and that of the LLM-driven outputs. Based on close viewing of subsets of the results, humans did seem to have utilized more variable underlying strategies in the creative task (properly quantifying this intuition would require deeper study, however). Given the currently utilized single instruction input (see SI), GPT-4o mostly produced prompts that mostly led to overwhelming the output with other objects or different scenery, e.g. "*A cartoon landscape with trees* transforms into futuristic cityscape, towering skyscrapers with neon lights, bustling drones flying, abstract robotic figures gliding past, shimmering holographic displays projecting vibrant surreal content." A common (but successful) theme was placing the objects underwater, e.g. "*A photo of a gray classic Mini Cooper* transformed into a steampunk airship floating in a vibrant underwater fantasy world, spires of colossal coral retro-futuristically outfitted with gears gleam luminously". It was instructed to be expressive and detailed, which paid off in the copying task as well, where precise prompts led to faithful copies of the reference image, e.g. "Tropical landscape with lush greenery, vibrant flowers, tall palm trees, cascading waterfall, serene turquoise river, and majestic mountain backdrop under a warm, colorful sky. Bright, vivid colors, serene atmosphere." or "Tree character with glasses reading a book, surrounded by bookshelves, colorful books, cartoon style, whimsical and educational setting, vibrant colors, playful and imaginative atmosphere." Given these reference images and instruction prompts, GPT-4o did

use the word "vibrant" a lot: this was the top non-function word it used (in 48 outputs of the 80). The LLM was instructed to make use of the word allowance, and as a result its outputs were quite long, on average 201 characters across all trials, compared to both laypeople (103) and artists (116).

3.4 What makes a good prompt?

We also carried out an explorative analysis of prompt components, using a multi-factorial feature analysis or quantitizing design. As abundantly demonstrated in recent literature, modern LLMs can be utilized as convenient on-demand classifiers and information retrieval engines in lieu of traditional NLP pipelines or human annotators (Karjus 2023; Ziems et al. 2023; Rathje et al. 2024). We use GPT-4o in a zero-shot manner with instructions to retrieve the following components, if present, from each of the 787 prompts: artistic style or style period, genre (e.g. landscape, still life), medium (photo, painting, etc.), mentions of any color, presence of describing adjectives, mentions of subjects and tangible objects, and in contrast, just setting or background descriptions (jungle, library, night-time). We selectively reviewed the LLM outputs and found they generally matched the data very well.

Here the two task datasets from human participants are concatenated for a joint analysis. They are made comparable by inverting the cosine similarities of the creative task and z-scoring (centering and dividing by standard deviation) these values separately for the two tasks. Prompt length, in characters, is similarly normalized by z-scoring separately, as creative task prompts had a fixed prefix the participants could not change and is not counted towards length. The prefix was removed from the feature analysis, and the creative prompts automatically coded as having an object and medium, as both were present in all prefixes. Figure 4.A illustrates the results of a linear regression model predicting simple additive contributions of the (binomially-coded) presence of the components and prompt length to task success. Longer and more informative prompts do slightly better as expected. Specifying stylistic aspects and setting a scene helps, but concrete descriptions of subjects (along with adjectives and colors) are a double edged sword, as mis-specifying them can easily lead to suboptimal results. Figure 4.B illustrates the relative usage of these components between groups, which are surprisingly not that different, but artists do appear more likely to define the style, which their training would be expected to support.

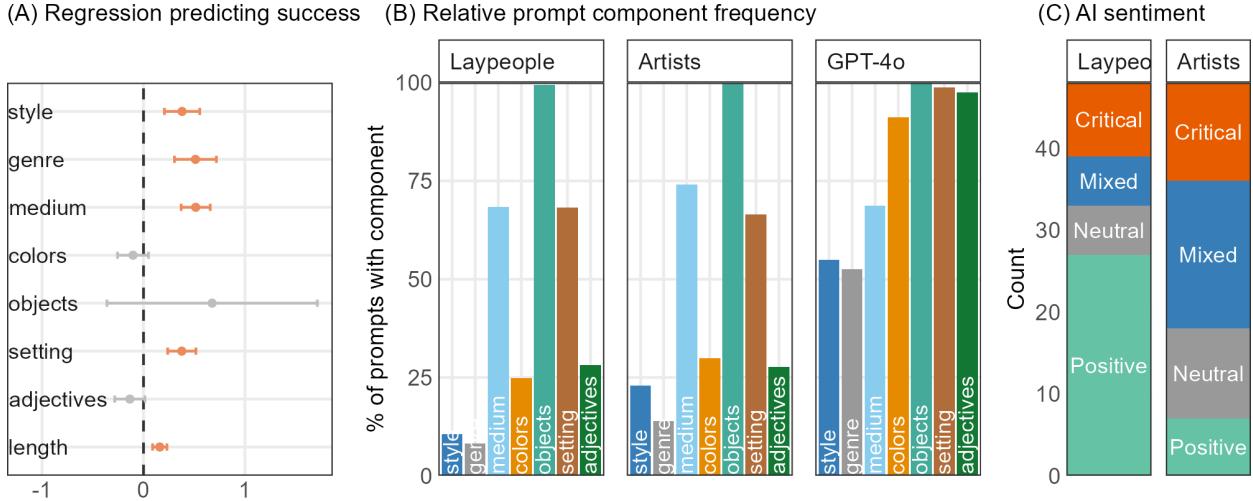


Figure 4. How do people write prompts? (A) Coefficients with 95% confidence intervals, from a linear regression model predicting the (z-scored cosine) success score: higher values indicate better result. Orange color indicates significant effects ($p < 0.05$). All variables except length are categorical, with no presence as the reference level (presence predicts the x-axis value of increase in standard deviations in score). Prompt length in characters is scaled separately for the tasks; a 1σ increase in length predicts a 0.18σ increase in score. Panel (B) shows the relative usage frequency of the same components by the two groups.

3.5 How do laypeople and artists feel about generative AI?

We also analyzed the participant sentiment or stance towards AI, elicited in the post-experiment questionnaire, asking how they feel about AI. Most left some comments, either on generative AI in general or on the generative tool used in the experiment (96 comments could be used, as a few were empty or illegible). As above, we used GPT-4o as a zero-

shot stance detection tool, requesting the following categories: positive (including excited or enthusiastic), critical or concerned, mixed (multiple or ambivalent feelings expressed), and neutral or indifferent (including discussing pragmatic technical aspects). Figure 4.C shows here differences do emerge: 27 of the laypeople are positively minded about AI compared to only 9 artists, who were more likely to have mixed or critical feelings. Naturally, as a survey, this is a small sample, and such attitudes have been surveyed more extensively elsewhere (Novozhilova et al. 2024; Goetze 2024; Lovato et al. 2024), but it serves to provide some insight into our sample.

4 Discussion

The results indicate that expertise in visual art does provide an advantage in using generative AI for image generation. Artist participants created more faithful copies and more creative imagery, even under the severe time and technological constraints of the experiment, limited to a single generative model in a restricted manner. While generative AI has made content creation accessible and fast, the deeper understanding and skills of professionals remain relevant in effectively using these tools as well. We see some qualitative evidence for this also within the content of artists' prompts, which were longer and contained more references to the style, genre, or medium of the image; all of which also contributed to better-performing results. At the same time, we can exclude that our main results were merely due to differences between the two samples in prompting experience, education level, or fluency in English, as these were controlled and matched for.

As AI becomes more integrated into the creative sectors, tools, and workflows, the skills defining professional artists and illustrators are changing. Traditional skills remain important, but integrating them with technological tools can enhance artistic training and professional work, while also unlocking new possibilities for artistic expression. This includes empowering individuals to realize visions that were previously unattainable due to physical or skill limitations. Contrasting with artificial (only) intelligence, this potential has recently been referred to as collaborative intelligence among other things (Mollick 2024).

4.1 Limitations and challenges

Unexpectedly, we did not find differences in image curation between the two samples that went beyond the fact that the average image created by artists was already performing better due to better prompting. As outlined in the results, we believe that this is due to the variance between the 4 created images being very low, leaving little opportunity for the artist group to demonstrate their skills. However, this is a useful insight by itself, suggesting that seed randomness is not a threat to our results.

As is common in such experiments, there are limitations to the interpretation and extrapolation of our results. Our participant groups were not entirely random samples. The Prolific participants are crowd workers who may not represent the average general population — more so as we intentionally matched them to the education level of the artists. The artists were carefully selected in hopes to be representative of the artistic profession. This is essentially an attempt to balance control, feasibility, and ecological validity, and ability to recruit a sample of professionals of sufficient size. It could also be argued that a potential difference in compensation relative to income could introduce a differential in economic motivations in completing the task. It could be equally argued that it may be impossible to create perfectly equivalent motivations for groups of potentially different income levels and relationships to the topic of the task. Here, it was closely related to the core activity and interests of one group, yet just another arbitrary task on a gig platform for the other.

The sample of artists may also seem quite small in absolute terms, but in practice convincing 50 busy professionals to lend their time was far from trivial. Our stimuli range was also quite small. While we attempted to cover a range of artistic expressions, obviously it represents only a fraction of what is considered art across different cultures, and inevitably biased by being produced by a single artist. Future research could attempt this type of experimental paradigm on larger or more diverse stimuli. The same goes for our two tasks, which do not cover the entire spectrum of artistic skills and practice. Again, the results and discussion should be interpreted with this in mind.

The effect sizes were also quite small, which is not surprising in a noisy artificial task with moderate ecological validity, especially for artists who in real practice have the choice of tools, media and processes at their disposal. Yet even under these constraints, they outperformed laypeople, which could be seen as a strong result in that sense. Moreover, it has

been argued that it is namely constraints that give rise to creativity and novelty (Feiten et al. 2023). This might also explain why the participants found the creative task particularly exciting and motivating.

4.2 Future research

While we focused on a few aspects of artistic practice and pitted professionals directly against laypeople, it would be interesting to see how these contrasts play out in a larger variety of tasks and setups. Such controlled experimental approaches can be used in synergy with more naturalistic data collection methods such as direct observation or interviews. These different data points can be used to inform decisions about AI integration in professional education. Besides the visual domain, it would be informative to carry out similar experiments with artistic professions in performance, film or music, other creative and content-producing professions, or people now identifying as "prompt engineers", using various generative models relevant to their disciplines. It could be compared how much of an edge professional background provides in concrete tasks, compared to an untrained person using AI assistance, or a person completing an equivalent task without any AI tools.

Here we also carried out a limited, non-preregistered comparison with an LLM, prompted to complete the same task. Surprisingly, it performed on par if not better than the average artist. Recent research has carried out similar comparisons of humans and machines, although often using crowdsourced laypeople to represent the former (Koivisto and Grassini 2023; Haase and Hanel 2023; Porter and Machery 2024; Beguš 2024). These types of studies provide information on how the uninitiated, amateurs or hobbyists fare against AI. However, a more pressing question with economic and social implications concerns professionals, as various fields may need to reevaluate aspects of training and daily practices. Systematic studies including experiments can be used to inform these processes. GPT-4o performance in the creative task does not necessarily mean that "AI is more creative than humans now". The top results were still all human. The LLM-generated prompts were longer on average and quite descriptive, leading to images on average diverging more from the reference in terms of CLIP embedding distance (we did check that simply adding a large number of any random words does not improve outcomes on this task; see SI). Its creations were also less diverse than the images produced by artists, indicating potential issues with its "creativity" at this different level relying on a few limited but high-performing "tricks". The AI system was good at optimizing on our key measure of distance, but part of the reason for this might be the constrained nature of the task. Future research could explore how diversity may be affected by initial instructions, and how images produced by independent AI agent pipelines (inferring the task, writing instructions for other generative models) would compare to human-prompted AI images in terms of creativity, novelty, or aesthetics, as perceived by humans (but also machines).

4.3 Future of art education

While various ethical, moral, and pedagogical issues continue to be discussed in arts and education communities, the spread of AI usage inevitably raises questions about the future importance of traditional artistic skills and education. It has been suggested that modern AI should be incorporated into art education, having been shown to aid the development of both technical and creative skills (Pavlik and Pavlik 2024; Sáez-Velasco et al. 2024; Fathoni 2023). Gu et al. (2024) show that AI tools enhance students' confidence in the creative process and suggest providing prompt engineering training to aid creativity and cognition. Generative AI enables rapid prototyping, content creation and curation at a groundbreaking pace and scale. This does not only affect text and visual art anymore, as video and music generation tools are beginning to mature. Disruptions in the arts are not unprecedented, however. Photography and film redefined artistic expression over a century ago, but oil painting remained relevant, itself having revolutionized art half a millennium earlier. Education should be critically engaging with the possibilities and limitations of such tools, as well as their underlying technical principles and ethical considerations, just as with other tools and materials.

5 Conclusions

We carried out a controlled behavioral experiment to compare visual generative AI tool usage by artists and untrained laypeople. Unlike preceding similar research, we recruited a sample of active professional artists as the domain experts. In our two tasks, we found that expertise does provide an edge and leads to better results in both faithful copying and creative novel production. Then again, the laypeople's results were only a small step behind. These findings suggest

several directions for future research. We also experimented with letting an example AI language model, GPT-4o, complete analogous tasks. We found it performs on par, if not better than humans in some cases, but does not surpass the best human results. This study serves as a preliminary exploration into how experts and non-experts compare in their ability to utilize novel and transformative AI tools, and how AI agents themselves can solve tasks that until only recently were thought the domain of human-only proficiency.

Author contributions and funding

T.F. Eisenmann designed the study, implemented and carried out the human experiments, wrote the preregistration, preprocessed the data, wrote parts of the paper (mostly Method), and provided edits and comments. A. Karjus designed the study, analyzed the data, carried out the LLM experiments, designed the figures, and wrote the paper. M. Canet Sola contributed to designing the study, created the stimuli, gathered the expert participants, wrote parts of the paper (mostly Method), produced the online dashboard, and co-designed the conceptual figure. L. Brinkmann designed the experimental infrastructure and co-designed the conceptual figure. B.I. Supriyatno implemented and monitored the text-to-image inference backend and set up the database. I. Rahwan commented on the framing of the manuscript and on the figures.

A.K. and M.C.S were supported during the initial research process by the CUDAN ERA Chair project for Cultural Data Analytics, funded through the European Union Horizon 2020 research and innovation program (Project No. 810961).

Acknowledgments

We would like to thank Yi-Tong Chen, Samira Fakhri, Diana Paola Americano Guerrero, Valerii Chirkov, and Omar Sherif for their assistance with implementing the experimental design. We further thank Ivan Soraperra for discussion on the statistical analyses and Yvonne Bialek for her help with reaching out to professional artists.

Data availability

The code and metadata to reproduce the analyses is available at <https://github.com/andreskarjus/genAExperiment>. The preregistration document is available at <https://aspredicted.org/t3mz-yjpz.pdf>. The image data is available at <https://doi.org/10.5281/zenodo.14710706>. See also <https://artistlaypeopleaiexperiment.github.io> for an interactive dashboard to explore the data.

References

- Becker, Howard Saul (1982). *Art Worlds*. University of California Press. 410 pp.
- Beguš, Nina (2024). "Experimental Narratives: A Comparison of Human Crowdsourced Storytelling and AI Storytelling". In: *Humanities and Social Sciences Communications* 11.1, pp. 1–22. doi: 10.1057/s41599-024-03868-8.
- Bellaiche, Lucas, Rohin Shahi, Martin Harry Turpin, Anya Ragnhildstveit, Shawn Sprockett, Nathaniel Barr, Alexander Christensen, and Paul Seli (2023). "Humans versus AI: Whether and Why We Prefer Human-Created Compared to AI-created Artwork". In: *Cognitive Research: Principles and Implications* 8.1, p. 42. doi: 10.1186/s41235-023-00499-6.
- Bellemare-Pépin, Antoine, François Lespinasse, Philipp Thölke, Yann Harel, Kory Mathewson, Jay A. Olson, Yoshua Bengio, and Karim Jerbi (2024). *Divergent Creativity in Humans and Large Language Models*. doi: 10.48550/arXiv.2405.13012. url: <http://arxiv.org/abs/2405.13012> (visited on 11/14/2024). Pre-published.
- Bezruczko, Nikolaus and David H. Schroeder (1994). "Differences in Visual Preferences and Cognitive Aptitudes of Professional Artists and Nonartists". In: *Empirical Studies of the Arts* 12.1, pp. 19–39. doi: 10.2190/92C4-L35B-FC60-89UH.
- Bhattacharya, Joydeep and Hellmuth Petsche (2005). "Drawing on Mind's Canvas: Differences in Cortical Integration Patterns between Artists and Non-Artists". In: *Human Brain Mapping* 26.1, pp. 1–14. doi: 10.1002/hbm.20104.
- Braguez, Joana (2023). "AI as a Creative Partner: Enhancing Artistic Creation and Acceptance". In: *ISSN: 2435-9475 – The Barcelona Conference on Arts, Media & Culture 2023: Official Conference Proceedings*, pp. 121–131.

- Cela-Conde, Camilo J., Francisco J. Ayala, Enric Munar, Fernando Maestú, Marcos Nadal, Miguel A. Capó, David del Río, Juan J. López-Ibor, Tomás Ortiz, Claudio Mirasso, and Gisèle Marty (2009). "Sex-Related Similarities and Differences in the Neural Correlates of Beauty". In: *Proceedings of the National Academy of Sciences* 106.10, pp. 3847–3852. doi: 10.1073/pnas.0900304106.
- Chamorro-Posada, Pedro (2016). "A Simple Method for Estimating the Fractal Dimension from Digital Images: The Compression Dimension". In: *Chaos, Solitons & Fractals* 91, pp. 562–572. doi: 10.1016/j.chaos.2016.08.002.
- Cooke, Di, Abigail Edwards, Sophia Barkoff, and Kathryn Kelly (2024). *As Good As A Coin Toss: Human Detection of AI-generated Images, Videos, Audio, and Audiovisual Stimuli*. doi: 10.48550/arXiv.2403.16760. URL: <http://arxiv.org/abs/2403.16760> (visited on 11/14/2024). Pre-published.
- Cox, Josie (2023). *AI Anxiety: The Workers Who Fear Losing Their Jobs to Artificial Intelligence*. BBC. URL: <https://www.bbc.com/worklife/article/20230418-ai-anxiety-artificial-intelligence-replace-jobs> (visited on 01/08/2025).
- Dege, Stefan (2023). *AI and Art — Are Creators about to Become Redundant?* dw.com. URL: <https://www.dw.com/en/ai-and-art-are-creators-about-to-become-redundant/a-64537364> (visited on 01/08/2025).
- Demirci, Ozge, Jonas Hannane, and Xinrong Zhu (2023). *Who Is AI Replacing? The Impact of Generative AI on Online Freelancing Platforms*. doi: 10.2139/ssrn.4602944. URL: <https://papers.ssrn.com/abstract=4602944> (visited on 11/14/2024). Pre-published.
- Doshi, Anil R. and Oliver P. Hauser (2024). "Generative AI Enhances Individual Creativity but Reduces the Collective Diversity of Novel Content". In: *Science Advances* 10.28, eadn5290. doi: 10.1126/sciadv.adn5290.
- Elfa, Ali, Mayssa Ahmad, and Mina Eshaq Tawfils Dawood (2023). "Using Artificial Intelligence for Enhancing Human Creativity". In: *Journal of Art, Design and Music* 2.2, Article 3. doi: 10.55554/2785-9649.1017.
- Fathoni, Ahmad Faisal Choiril Anam (2023). "Leveraging Generative AI Solutions in Art and Design Education: Bridging Sustainable Creativity and Fostering Academic Integrity for Innovative Society". In: *E3S Web of Conferences* 426, p. 01102. doi: 10.1051/e3sconf/202342601102.
- Feiten, Tim Elmo, Zachary Peck, Kristopher Holland, and Anthony Chemero (2023). "Constructive Constraints: On the Role of Chance and Complexity in Artistic Creativity". In: *Possibility Studies & Society* 1.3, pp. 311–323. doi: 10.1177/27538699231193539.
- Frank, Joel, Franziska Herbert, Jonas Ricker, Lea Schönherr, Thorsten Eisenhofer, Asja Fischer, Markus Dürmuth, and Thorsten Holz (2023). *A Representative Study on Human Detection of Artificially Generated Media Across Countries*. doi: 10.48550/arXiv.2312.05976. URL: <http://arxiv.org/abs/2312.05976> (visited on 11/14/2024). Pre-published.
- Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli (2023). "ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks". In: *Proceedings of the National Academy of Sciences* 120.30, e2305016120. doi: 10.1073/pnas.2305016120.
- Goetze, Trystan S. (2024). "AI Art Is Theft: Labour, Extraction, and Exploitation: Or, On the Dangers of Stochastic Pollocks". In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '24. New York, NY, USA: Association for Computing Machinery, pp. 186–196. doi: 10.1145/3630106.3658898. URL: <https://doi.org/10.1145/3630106.3658898> (visited on 11/14/2024).
- Gu, Quan, Yiduo Wang, Xiaoxiao Hu, and Orit Shaer (2024). "Exploring the Impact of Human-AI Collaboration on College Students' Tangible Creation: Building Poetic Scenes with LEGO Bricks". In: *Joint Proceedings of the ACM IUI Workshops 2024, March 18-21, 2024, Greenville, South Carolina, USA*.
- Haase, Jennifer and Paul H. P. Hanel (2023). "Artificial Muses: Generative Artificial Intelligence Chatbots Have Risen to Human-Level Creativity". In: *Journal of Creativity* 33.3, p. 100066. doi: 10.1016/j.yjoc.2023.100066.
- Hasler, David and Sabine E. Suesstrunk (2003). "Measuring Colorfulness in Natural Images". In: *Human Vision and Electronic Imaging VIII*. Human Vision and Electronic Imaging VIII. Vol. 5007. SPIE, pp. 87–95. doi: 10.1117/12.477378. URL: <https://www.spiedigitallibrary.org/conference-proceedings-of-spie/5007/0000/Measuring-colorfulness-in-natural-images/10.1117/12.477378.full> (visited on 04/02/2022).
- Hubert, Kent F., Kim N. Awa, and Darya L. Zabelina (2024). "The Current State of Artificial Intelligence Generative Language Models Is More Creative than Humans on Divergent Thinking Tasks". In: *Scientific Reports* 14.1, p. 3440. doi: 10.1038/s41598-024-53303-w.
- Jahani, Eaman, Benjamin Manning, Joe Zhang, Hong-Yi TuYe, Mohammed Abdulrahman M. Alsobay, Christos Nicolaides, Siddharth Suri, and David Holtz (2024). *As Generative Models Improve, We Must Adapt Our Prompts*. doi: 10.31219/osf.io/9rhku. URL: <https://osf.io/9rhku> (visited on 12/12/2024). Pre-published.
- Karjus, Andres (2023). "Machine-Assisted Quantitizing Designs: Augmenting Humanities and Social Sciences with Artificial Intelligence". In: *ArXiv e-prints*.

- Karjus, Andres, Richard A. Blythe, Simon Kirby, Tianyu Wang, and Kenny Smith (2021). "Conceptual Similarity and Communicative Need Shape Colexification: An Experimental Study". In: *Cognitive Science* 45.9, e13035. doi: 10.1111/cogs.13035.
- Karjus, Andres, Mar Canet Solà, Tillmann Ohm, Sebastian E. Ahnert, and Maximilian Schich (2023). "Compression Ensembles Quantify Aesthetic Complexity and the Evolution of Visual Art". In: *EPJ Data Science* 12.1 (1), pp. 1–23. doi: 10.1140/epjds/s13688-023-00397-3.
- Kim, Yoolim, Vita V Kogan, and Cong Zhang (2024). "Collecting Big Data Through Citizen Science: Gamification and Game-based Approaches to Data Collection in Applied Linguistics". In: *Applied Linguistics* 45.1, pp. 198–205. doi: 10.1093/applin/amad039.
- Kirby, Simon, Hannah Cornish, and Kenny Smith (2008). "Cumulative Cultural Evolution in the Laboratory: An Experimental Approach to the Origins of Structure in Human Language". In: *Proceedings of the National Academy of Sciences* 105.31, pp. 10681–10686. doi: 10.1073/pnas.0707835105.
- Klein, Ezra (2023). *Opinion | This Changes Everything*. The New York Times. URL: <https://www.nytimes.com/2023/03/12/opinion/chatbots-artificial-intelligence-future-weirdness.html> (visited on 03/16/2023).
- Koivisto, Mika and Simone Grassini (2023). "Best Humans Still Outperform Artificial Intelligence in a Creative Divergent Thinking Task". In: *Scientific Reports* 13.1, p. 13601. doi: 10.1038/s41598-023-40858-3.
- Kozbelt, Aaron (2001). "Artists as Experts in Visual Cognition". In: *Visual Cognition* 8.6, pp. 705–723. doi: 10.1080/13506280042000090.
- Lakhal, Samy, Alexandre Darmon, Jean-Philippe Bouchaud, and Michael Benzaquen (2020). "Beauty and Structural Complexity". In: *Physical Review Research* 2.2, p. 022058. doi: 10.1103/PhysRevResearch.2.022058.
- Liu, Vivian and Lydia B Chilton (2022). "Design Guidelines for Prompt Engineering Text-to-Image Generative Models". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New York, NY, USA: Association for Computing Machinery, pp. 1–23. doi: 10.1145/3491102.3501825. URL: <https://dl.acm.org/doi/10.1145/3491102.3501825> (visited on 11/14/2024).
- Lovato, Juniper, Julia Zimmerman, Isabelle Smith, Peter Dodds, and Jennifer Karson (2024). *Foregrounding Artist Opinions: A Survey Study on Transparency, Ownership, and Fairness in AI Generative Art*. doi: 10.48550/arXiv.2401.15497. URL: <http://arxiv.org/abs/2401.15497> (visited on 11/14/2024). Pre-published.
- Lu, Zeyu, Di Huang, Lei Bai, Jingjing Qu, Chengyue Wu, Xihui Liu, and Wanli Ouyang (2024). "Seeing Is Not Always Believing: Benchmarking Human and Model Perception of AI-generated Images". In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., pp. 25435–25447.
- Machado, Penousal, Juan Romero, Marcos Nadal, Antonino Santos, João Correia, and Adrián Carballal (2015). "Computerized Measures of Visual Complexity". In: *Acta Psychologica* 160, pp. 43–57. doi: 10.1016/j.actpsy.2015.06.005.
- Miravete, Sébastien and André Tricot (2024). "Are Some People Generally More Creative Than Others? A Systematic Review of Fifty Years' Research". In: *Educational Psychology Review* 36.4, p. 99. doi: 10.1007/s10648-024-09926-6.
- Mirowski, Piotr, Juliette Love, Kory Mathewson, and Shakir Mohamed (2024). "A Robot Walks into a Bar: Can Language Models Serve as Creativity Support Tools for Comedy? An Evaluation of LLMs' Humour Alignment with Comedians". In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '24. New York, NY, USA: Association for Computing Machinery, pp. 1622–1636. doi: 10.1145/3630106.3658993. URL: <https://dl.acm.org/doi/10.1145/3630106.3658993> (visited on 11/14/2024).
- Miyazaki, Kunihiro, Taichi Murayama, Takayuki Uchiba, Jisun An, and Haewoon Kwak (2024). "Public Perception of Generative AI on Twitter: An Empirical Study Based on Occupation and Usage". In: *EPJ Data Science* 13.1 (1), pp. 1–20. doi: 10.1140/epjds/s13688-023-00445-y.
- Mollick, Ethan (2024). *Co-Intelligence*. Penguin Random House.
- Müller, Thomas F., James Winters, and Olivier Morin (2019). "The Influence of Shared Visual Context on the Successful Emergence of Conventions in a Referential Communication Task". In: *Cognitive Science* 43.9, e12783. doi: 10.1111/cogs.12783.
- Nölle, Jonas, Marlène Staib, Riccardo Fusaroli, and Kristian Tylén (2018). "The Emergence of Systematicity: How Environmental and Communicative Factors Shape a Novel Communication System". In: *Cognition* 181, pp. 93–104.
- Novozhilova, Ekaterina, Kate Mays, and James E. Katz (2024). "Looking towards an Automated Future: U.S. Attitudes towards Future Artificial Intelligence Instantiations and Their Effect". In: *Humanities and Social Sciences Communications* 11.1, pp. 1–11. doi: 10.1057/s41599-024-02625-1.
- Ohm, Tillmann, Mar Canet Solà, Andres Karjus, and Maximilian Schich (2023). "Collection Space Navigator: An Interactive Visualization Interface for Multidimensional Datasets". In: *arXiv preprint*. doi: 10.48550/arXiv.2305.06809.

- Okada, Takeshi and Kentaro Ishibashi (2017). "Imitation, Inspiration, and Creation: Cognitive Process of Creative Drawing by Copying Others' Artworks". In: *Cognitive Science* 41.7, pp. 1804–1837. doi: 10.1111/cogs.12442.
- Oppenlaender, Jonas, Rhema Linder, and Johanna Silvennoinen (2024). *Prompting AI Art: An Investigation into the Creative Skill of Prompt Engineering*. doi: 10.48550/arXiv.2303.13534. URL: <http://arxiv.org/abs/2303.13534> (visited on 11/14/2024). Pre-published.
- Pavlik, J. and O. Pavlik (2024). "Art Education and Generative AI: An Exploratory Study in Constructivist Learning and Visualization Automation for the Classroom". In: *Creative Education* 15, pp. 601–616. doi: 10.4236/ce.2024.154037.
- Porter, Brian and Edouard Machery (2024). "AI-generated Poetry Is Indistinguishable from Human-Written Poetry and Is Rated More Favorably". In: *Scientific Reports* 14.1, p. 26133. doi: 10.1038/s41598-024-76900-1.
- Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever (2021). *Learning Transferable Visual Models From Natural Language Supervision*. doi: 10.48550/arXiv.2103.00020. URL: <http://arxiv.org/abs/2103.00020> (visited on 11/14/2024). Pre-published.
- Rando, Javier, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr (2022). *Red-Teaming the Stable Diffusion Safety Filter*. doi: 10.48550/arXiv.2210.04610. URL: <http://arxiv.org/abs/2210.04610> (visited on 01/21/2025). Pre-published.
- Rathje, Steve, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjeh, Claire E. Robertson, and Jay J. Van Bavel (2024). "GPT Is an Effective Tool for Multilingual Psychological Text Analysis". In: *Proceedings of the National Academy of Sciences* 121.34, e2308950121. doi: 10.1073/pnas.2308950121.
- Roose, Kevin (2022). *An A.I.-Generated Picture Won an Art Prize. Artists Aren't Happy*. The New York Times. URL: <https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html> (visited on 01/08/2025).
- Sáez-Velasco, Sara, Mario Alaguero-Rodríguez, Vanesa Delgado-Benito, and Sonia Rodríguez-Cano (2024). "Analysing the Impact of Generative AI in Arts Education: A Cross-Disciplinary Perspective of Educators and Students in Higher Education". In: *Informatics* 11.37 (2). doi: 10.3390/informatics11020037.
- Sauer, Axel, Dominik Lorenz, Andreas Blattmann, and Robin Rombach (2023). *Adversarial Diffusion Distillation*. doi: 10.48550/arXiv.2311.17042. URL: <http://arxiv.org/abs/2311.17042> (visited on 01/06/2025). Pre-published.
- Shen, Shuhan, Yuetong Chen, Min Hua, and Mai Ye (2023). "Measuring Designers 'use of Midjourney on the Technology Acceptance Model". In: *IASDR 2023: Life-Changing Design, 9–13 October, Milan, Italy*. doi: 10.21606/iasdr.2023.794.
- Törnberg, Petter (2023). *ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning*. doi: 10.48550/arXiv.2304.06588. URL: <http://arxiv.org/abs/2304.06588> (visited on 09/12/2023). Pre-published.
- Torngren, Gustaf and Henry Montgomery (2004). "Worse Than Chance? Performance and Confidence Among Professionals and Laypeople in the Stock Market". In: *The Journal of Behavioral Finance*. doi: 10.1207/s15427579jpfm0503_3.
- Von Garrel, Jörg and Jana Mayer (2023). "Artificial Intelligence in Studies—Use of ChatGPT and AI-based Tools among Students in Germany". In: *Humanities and Social Sciences Communications* 10.1, pp. 1–9. doi: 10.1057/s41599-023-02304-7.
- Walkowiak, Emmanuelle and Jason Potts (2024). *Generative AI, Work and Risks in Cultural and Creative Industries*. doi: 10.2139/ssrn.4830265. URL: <https://papers.ssrn.com/abstract=4830265> (visited on 11/14/2024). Pre-published.
- Wang, Haonan, James Zou, Michael Mozer, Anirudh Goyal, Alex Lamb, Linjun Zhang, Weijie J. Su, Zhun Deng, Michael Qizhe Xie, Hannah Brown, and Kenji Kawaguchi (2024). *Can AI Be as Creative as Humans?* doi: 10.48550/arXiv.2401.01623. URL: <http://arxiv.org/abs/2401.01623> (visited on 11/14/2024). Pre-published.
- Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou (2022). "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., pp. 24824–24837. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.
- Zhou, Eric and Dokyun Lee (2024). "Generative Artificial Intelligence, Human Creativity, and Art". In: *PNAS Nexus* 3.3, pgae052. doi: 10.1093/pnasnexus/pgae052.
- Ziems, Caleb, Omar Shaikh, Zhehao Zhang, William Held, Jiaao Chen, and Diyi Yang (2023). "Can Large Language Models Transform Computational Social Science?" In: *Computational Linguistics*, pp. 1–53. doi: 10.1162/coli_a_00502.

Supplementary Information

S 1 Additional insights into the experimental data

Figure S1, as mentioned in the Methods section of the main text, provides further insight into the properties of the generated images and background of the participants. In general, we find no notable differences in terms of image colorfulness or complexity between groups. The only slight difference in average prompt length between groups was not significant (simple linear regression, $F(1, 3146) = 3.56, p = 0.06$).

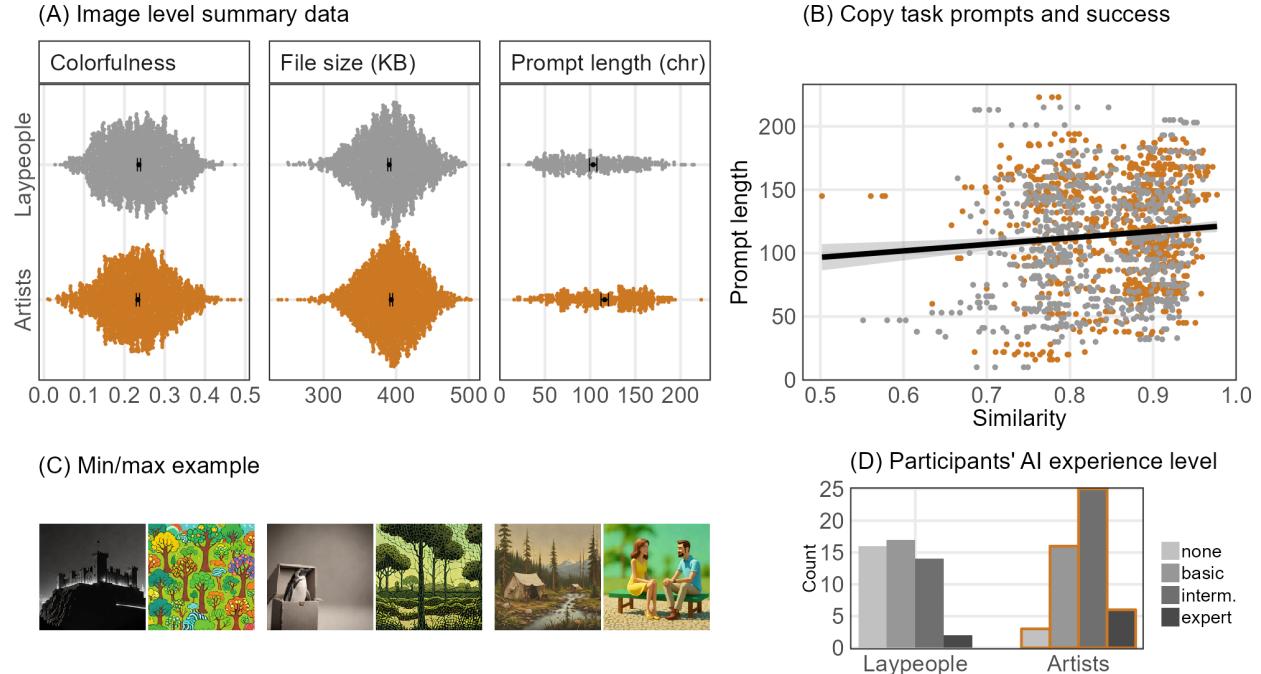


Figure S1. Overview of the experiment data, including all variant images. Panel (A): image colorfulness, estimated complexity (compressed size) and prompt length (characters) are fairly similar for both groups (each point is an image). There are fewer data points in the latter panel, as the prompt is the same for all four variants. Panel (C) displays the image with the lowest and highest value in each of the metrics, e.g. for colorfulness (left) the gray-scale castle is the lowest while the bright cartoony forest is the most colorful. The complexity (middle) of the monochrome penguin with a plain background is lowest while the intricately textured green forest is highest. Longer prompts can help specify an output, but shorter prompts can still create complex images, e.g. the cabin (right) is generated by a one-word "wilderness". Longer prompts can help but do not always lead to more accurate depictions when copying (B). Panel (D) illustrates prior experience with AI tools on a scale of 0 (none) to 3 (expert).

S 2 Adding random words to the prompt does not help much on average in the creative task

As discussed in Methods, we also experimented with two random baselines to make sure the creative task could not be completed successfully by just adding random words or gibberish to the prefixed part. Neither leads to better results than the human (or GPT) written prompts, on average, but a long list of random words can indeed do better than a badly written or too short of a prompt, when the task is to write a prompt that would diverge from the fixed prefix as far as possible.

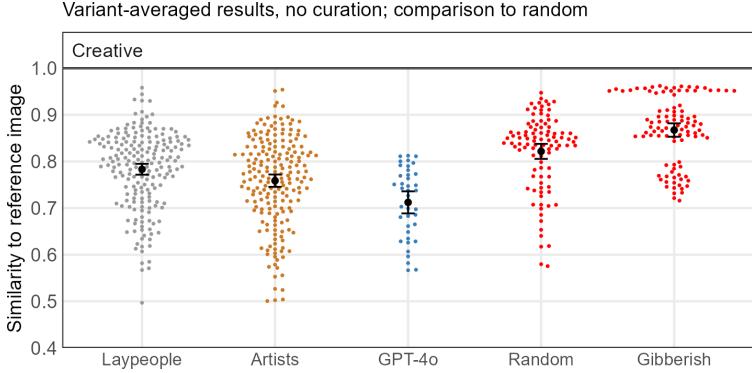


Figure S2. Experiment results for the creative task, where the goal was to create images with a low similarity to the reference (variant-averaged, no curation), compared to two random baselines: prompts consisting of randomly sampled English words, and of randomly constructed gibberish pseudo-words. 95% confidence intervals (black) added for reference.

S 3 Human and GPT results trial by trial

Figure S3 expands the main results figure of the main text, showing results for each trial separately (without curation). While there is variation between the trials (which we control for in statistical modeling), the overall trends remain the same. Some trials were more difficult than others. Creating a copy of the apples still life painting was comparatively easier than copying the couple on the bench, but the former proved challenging still for about half the laypeople (notice the bimodal distribution). Making the castle or penguin different from the reference was easier than doing the same with the Mini Cooper.

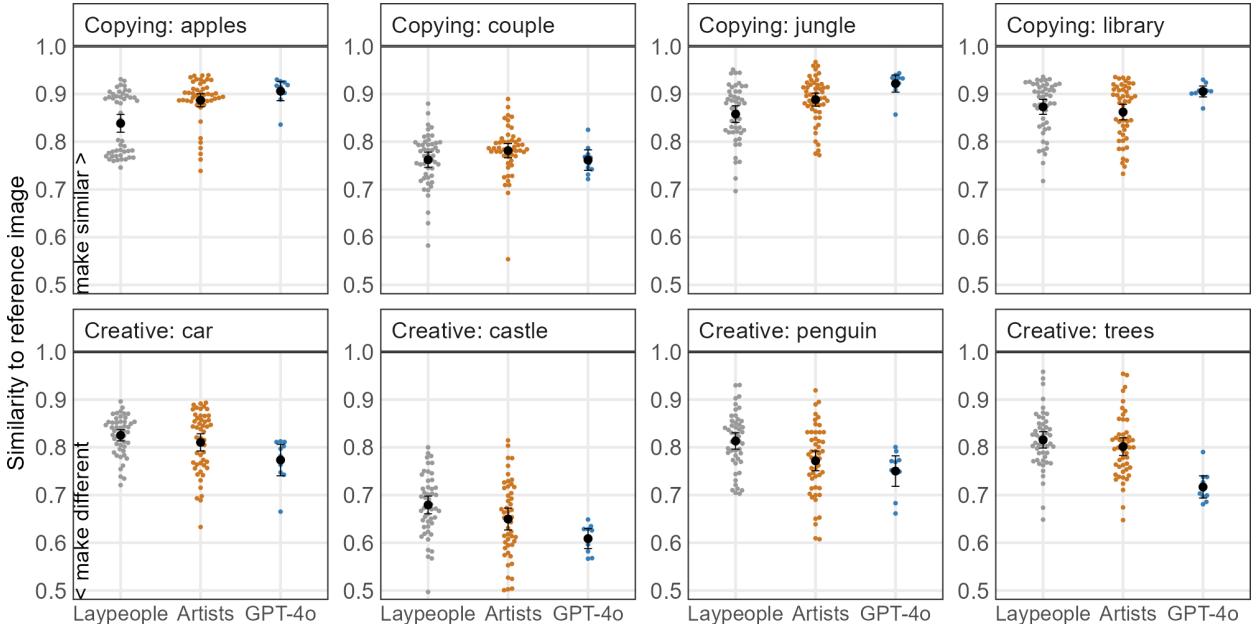


Figure S3. Experimental results by trial. As in the figure in the main text, each point is the averaged cosine similarity of the generated image quartet to the reference image, and each point represents the attempt of one participant.

S 4 Human and GPT results illustrated

This section serves to further illustrate the variable results produced by our human participants as well as the semi-autonomous image generation agent. Figure S4 depicts the reference images and all results from one example participant to illustrate the (little) variability between the variants, as well as the solutions to the task by one creative individual.

(A) One participant's generations in Copy task



(B) The participant's generations in Creative task

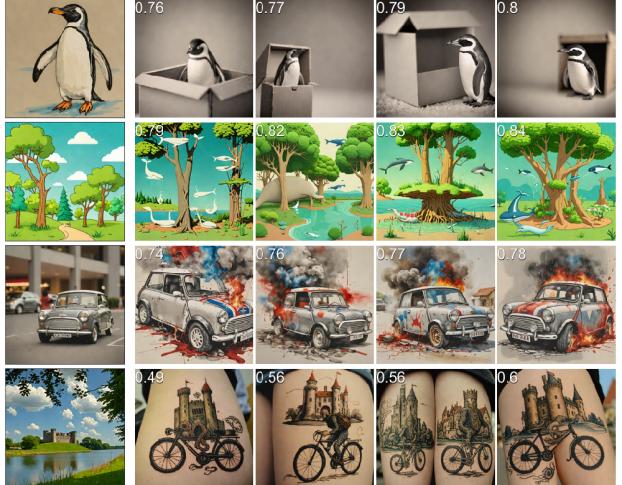


Figure S4. All creations by one artist participant. The column on the left are the reference images; cosine similarity in the corners of the results. The participant used the following prompts for the copying task (A): "abstract painting of green and red apples in a basket. Orange and grayish blue checkered background", "two persons of color sitting on a bench talking. man sitting to the left. female in yellow dress to the right. blurred builing in the background", "fantasy painting of waterfall. exotic flowers in the foreground. mountains in the back. sunset and a few clouds", "cartoon tree with glasses reading a book at a library. Cartoon drawing". The creative task (B) - with the fixed prefix of the prompt highlighted in italics here: "*Painting of a penguin* in a box hiding. black and white photo motionblurred", "*A cartoon landscape with trees* negative and upside down. skulpture and performed by whales wearing hats", "*A photo of a gray classic Mini Cooper* exploded by a carbomb. watercolor with red and blue colors only. painted by a monkey", and "*Photograph of a medieval castle*, small tattoo on a octopus wearing pants and on a bycycle". In the latter, the participant managed to refocus the generator's attention, successfully backgrounding the original concept.

Figures S5 and S6 expand the top examples graph in the main text, showing the best and worst attempts by both our human participants as well as those in the additional LLM experiment.

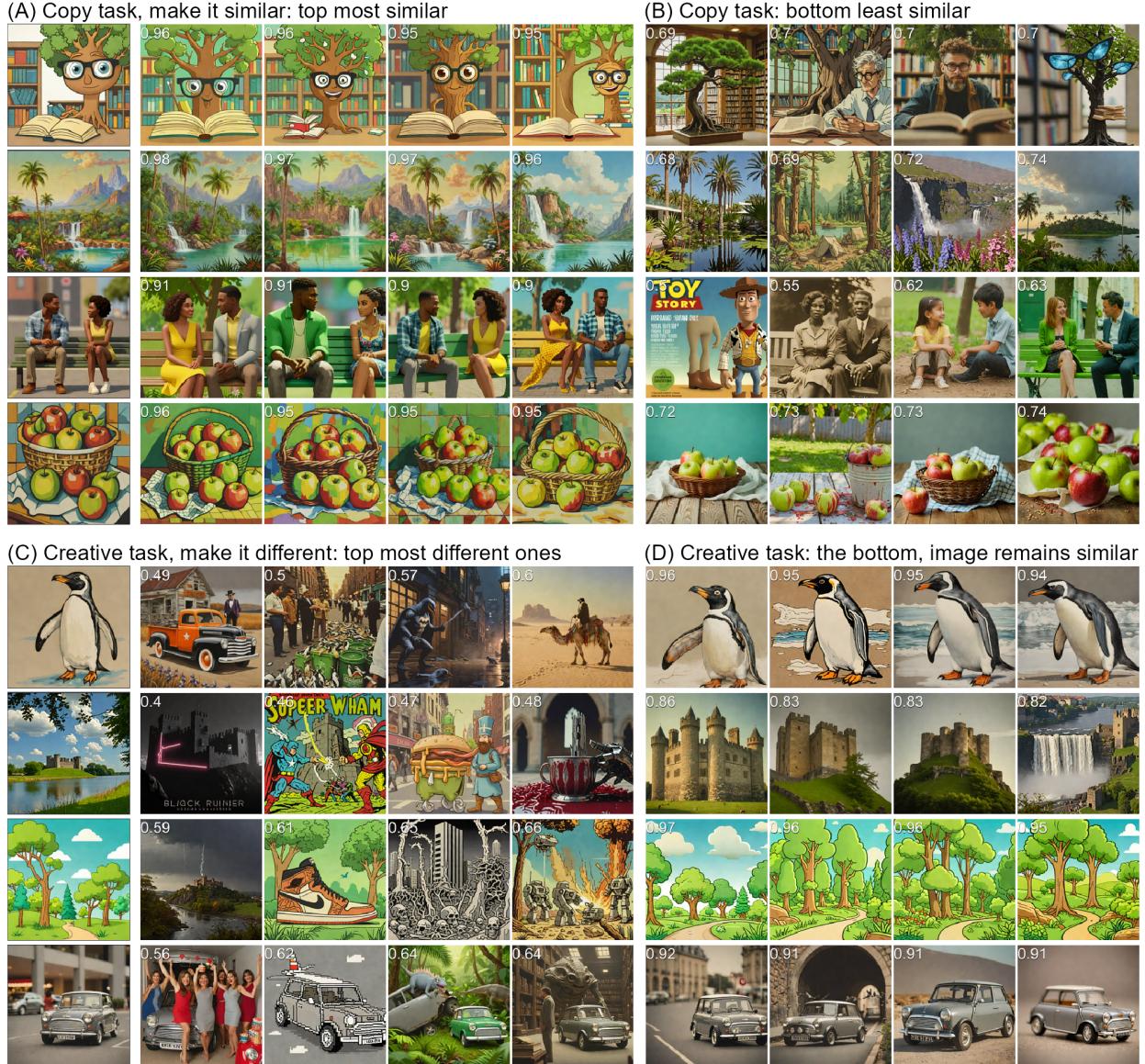


Figure S5. Examples of most and least successful trials by human participants across the experiment; cosine similarity in the corners. The reference images are in the leftmost column. This supplements the figure in the main text, providing not only the best but also the worst attempts for comparison.

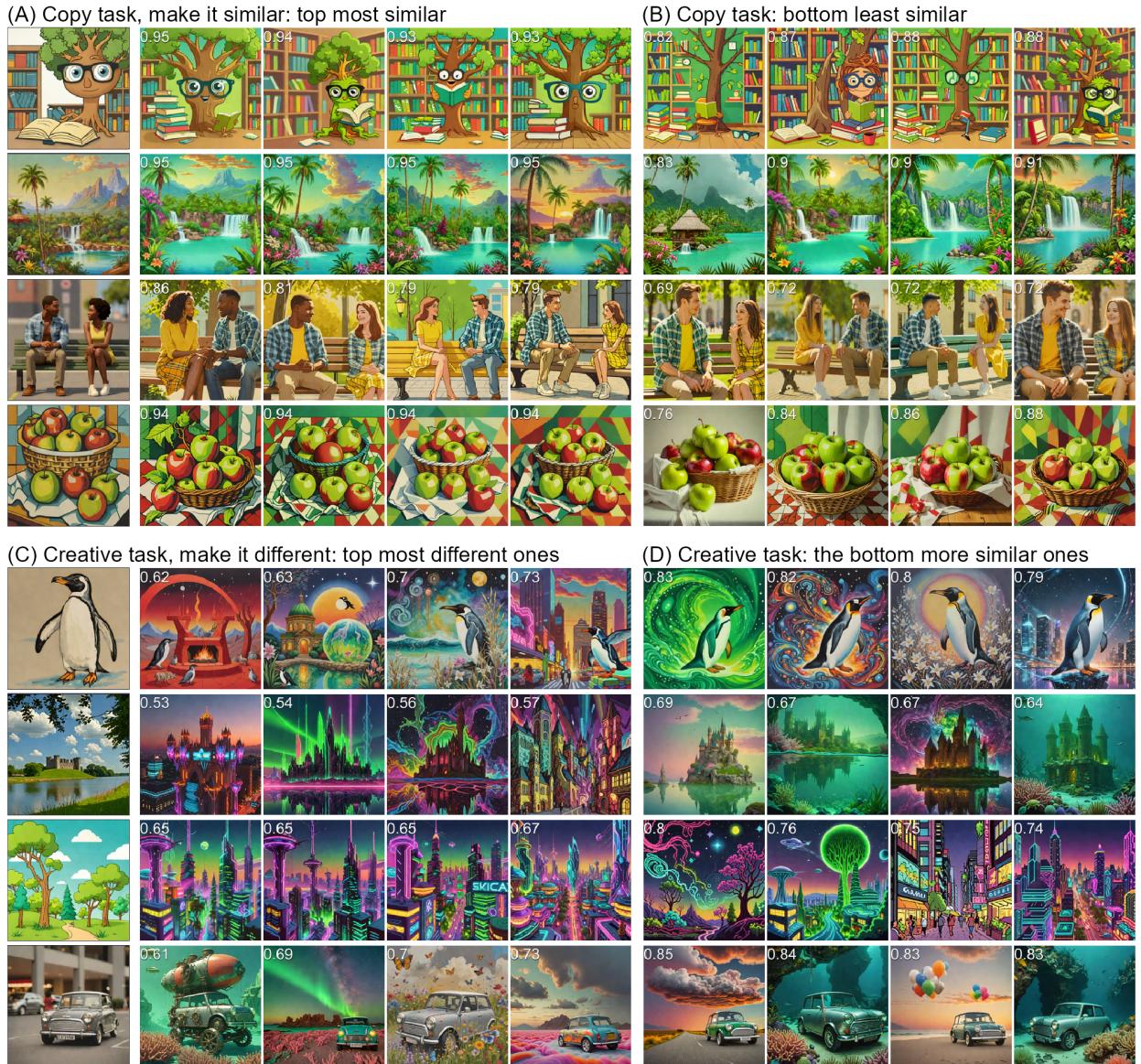


Figure S6. Examples of most and least successful trials by GPT-4o; cosine similarity in the corners. The LLM was prompted to complete the same trials in a comparable setting, and performed on par (or even slightly better) than humans. GPT's worst attempts are often better than the worst of human participants, demonstrating good adherence to instructions. In the creative task, the prevalent strategy seems to have been to overwhelm the image with new objects and colors. The reference images are again in the leftmost column.

S 5 Stimuli generation prompts

The copying task reference image were created with Stable Diffusion using the following prompts. See the figures above for the images themselves, here referenced to using single-word shorthands:

Library: cartoon style tree character with black glasses and big blue eyes and a book in the left in a library

Jungle: oil painting of a colourful African landscape with a river, a jungle with palm trees, flowers, a waterfall and a mountain on the horizon

Couple: A 3D-rendered game screenshot featuring an African American man on the left and a woman on the right, sitting on a bench at a bus stop. They are facing each other in a friendly manner, engaged in conversation, set against an urban street backdrop.

Apples: An oil painting of a full basket of apples in the Cubist style, inspired by Picasso

The creative task reference image were created with Stable Diffusion using the following prompts; the fixed parts that the participants could see but not edit are shown in bold:

Penguin: **Painting of a penguin** in a student-like style

Castle: **Photograph of a medieval castle** with four watch towers centered by a lake, framed by lush grass and scattered trees. The castle stands prominently under a bright blue sky, dotted with impressive cumulonimbus clouds.

Trees: **A cartoon landscape with trees** and blue sky

Car: **A photo of a classic gray Mini Cooper** in a parking lot of a shopping mall

S 6 Prompts used in the LLM experiment

The prompts used to instruct GPT-4o to complete the experiment were as follows. The reference image, mentioned in the prompt, was sent to the API along with the prompt.

The copying task: *You are an expert at visual art, photography, drawing, painting, art history, and an expert prompter. Your task is to create an image by making use of a generative AI model, our "image machine". To do so, you need to provide text inputs, so-called "prompts". You are being shown an image. Your goal is to write a detailed prompt that would result in an image that is as close as possible to this image in all aspects, including main subjects and where they are on the image, and image composition, colors, style or genre, etc. Describe the shape, size and color of prominent objects, and relevant characteristics (gender, age, appearance, ethnicity, etc) of any focal human subjects. The prompt should be up to 30 words in length, so be concise. Do not comment, just provide the 30-word prompt without quotes.*

The creative task: *You are a creative visual artist and an expert prompter. Your task is to create an image by making use of a generative AI model, our "image machine". To do so, you need to provide text inputs, so-called "prompts". You are being shown an image. Your goal is to write a prompt that will make a new image as different from the original as possible. Pro tip: Ideally, you should try to aim for creating an image with different content as well as different visual style, as unrelated to the original as possible. Also, negation does not work very well here - if you state "an image with no elephants", you are actually more likely to get an image with elephants! So do not use negation in the prompt.*

Now look at this image. Try to write a prompt that would create an image that is as different as possible from this image. The prompt should be up to 30 words in length in total but not longer. However, you are constrained in that your prompt MUST start with this phrase describing the original image:

/the fixed lead of the given trial, e.g. "A cartoon landscape with trees"/

Start with that phrase, but then try to write the rest of the prompt so that its resulting image would actually not contain or would obscure anything mentioned in that phrase. Use clever wordplay and think outside the box to get away from the meaning and appearance of the reference image and its descriptive phrase. Your prompt, while containing this initial phrase, should describe a new image that would look as different as possible from the reference image you are seeing right now, and would be nothing like /the fixed lead/. Do not comment, just provide the 30-word prompt without quotes.

S 7 Prompts used in zero-shot analytics

The zero-shot instructions for the prompt components feature analysis were the following:

Analyze this short quoted Text describing an image and output a comma-separated list of elements that are mentioned. Output only these keywords where relevant, as explained and exemplified (in brackets) here. These are the only valid output terms:

style = if Text mentions artistic style or period (like cubist, expressionism)
genre = mentions genre (like landscape, still life, fantasy)
medium = mentions type or medium of image (like photo, painting, cartoon)
colors = mentions any colors (including phrases like colorful or bright colors)
adjectives = contains any adjectives or similar descriptives (like pleasant, misty, dystopian, destroyed)
objects = mentions one or more tangible subjects or objects (like car, trees, buildings, people)
setting = mentions a setting or background (like library, sea, jungle, wilderness, night) action = if Text mentions some action (like reading, sit, standing).

How to respond: only a comma-separated keywords from the above if relevant, but do not describe the actual colors or styles, and do not comment! If no none of the above a present, just say None. Here is the Text to analyze:

"""/prompt text/ """

This prompt was used to infer sentiment:

Below is a quoted Feedback from an experiment where people used gen AI to create images. We asked them how they feel about AI. Summarize the Feedback sentiment as one of these categories that best matches its primary sentiment:

Positive = mostly positive, excited, praising or enthusiastic attitude towards AI.

Critical = a critical, concerned or dissatisfied attitude or talks about flaws or worries or implies AI is not good for art.

Mixed = if Feedback expresses mixed or ambivalent feelings or mentions both positive and negative aspects (eg "fun but boring").

Neutral = pragmatic or indifferent Feedback, or neutral discussion about technical aspects without sentiment or emotion (eg "task was challenging" - this is not negative about AI as such discusses the task they completed beforehand).

Output a single category, do not comment! This is the Feedback:

"""/Opinion on generative AI: /the opinion/ """