Final Report

● Ungraded

1 Day, 23 Hours Late

Group

John Majernik
Jason Figueroa
Via Liu
...and 2 more
View or edit group

Total Points

- / 15 pts

Question 1

Report 15 pts

Question assigned to the following page: 1

Basic project information

Name of your project

PixelPatchwork

Name of your teammates

Dan, Jason, Ryan, Via, John

Give a one sentence description of your project. Please use the name of the project in your description.

PixelPatchwork is a collaborative digital art platform that enables users to iteratively refine AI-generated images through masking, prompting, and a voting system to collectively curate high-quality visuals.

Logo for your project.

What problem does it solve?

Question assigned to the following page: 1

We aimed to highlight the limitations and possibilities of AI art and examine worker skills (prompt writing for determining winning images).

What similar projects exist?

- Wikipedia, open source contributions - both of which use the iterative human computation approach
    - https://crowdsourcing-class.org/readings/downloads/programming/iterative-and-parallel-processing-in-hcomp.pdf
- Google Arts and Culture Experiment
    - https://artsandculture.google.com/experiment/say-what-you-see/jwG3m7wQShZngw?hl=en
- art98: a real-time collaborative pixel art canvas supporting multiple user types with varying privileges. Features include zooming, panning, and real-time drawing.
    - https://art98.vercel.app/
- Magma: a collaborative drawing website where artists can paint together on a shared canvas. It supports sketching with friends online, fostering a community of artists.
    - https://magma.com

What type of project is it?

Social science experiment. We use the crowd to iteratively generate AI art.

What was the main focus of your team's effort

Our focus was on evaluating both the collective final output of the crowd's efforts and the prompts generated to produce a winning image. We engineered a complex system to enable branching tree scenarios, where users could fork from a starting point and create multiple versions of artwork, promoting diverse creative directions. This iterative approach involved sequential creation tasks, with each task building on the best output from previous iterations. AI-generated art, such as DALL-E seed images, served as starting points or modification suggestions, allowing users to vote or propose changes.

How does your project work? Describe each of the steps involved in your project. What parts are done by the crowd, and what parts will be done automatically.

Our project operates as an iterative and collaborative platform for image generation and refinement, combining user input with AI image generation. Data is collected from classmates and Amazon Mechanical Turk workers, who interact with a seed image presented on a canvas. Workers use a masking tool to indicate areas of the image they wish to modify and provide detailed prompts describing the desired outcome. These inputs are processed using DALL-E, generating customized images based on the user's edits and prompts. After contributing, workers participate in a voting system, where they upvote or downvote images created by other users on the same day, acting as a quality control

Question assigned to the following page: 1

mechanism. At the end of the day, votes are aggregated, and the highest-voted image becomes the seed image for the next day, creating an iterative “feedback-loop” that incorporates community preferences. Additionally, workers can view previous contributions through a "show history" feature, fostering collaboration by enabling inspiration from prior responses. This system seamlessly combines independent creation, iterative refinement, and community-driven evaluation to produce evolving interpretations of the original seed image over multiple cycles.

Provide a link to your final presentation video. Give the full URL to your YouTube video or your Google Drive video and make sure it is publicly available (OK to keep it unlisted).

https://drive.google.com/file/d/16jjDGjSW1skFFihlSFL2iqCjEpGfvOBN/view?usp=drive_link

Which two sections below did you pick for your in-depth analysis?

We focused on skills and project analysis.

The Crowd

Who are the members of your crowd?

The members of our crowd included classmates as well as workers from Amazon Mechanical Turk (MTurk).

For your final project, did you simulate the crowd or run a real experiment?

We used a real crowd for our experiment.

If the crowd was real, how did you recruit participants?

We recruited classmates through the “Be the Crowd” in-class assignment. We also recruited Turkers through HITs posted on Amazon MTurk using small monetary incentives.

How many unique participants did you have?

By the end of the experiment, we recorded 133 unique participant sessions over the course of 16 days.

Incentives

What motivation does the crowd have for participating in your project?

Question assigned to the following page: 1

Turkers were motivated by pay; HITs were published offering workers $0.05 per task. Classmates were motivated by class participation credit.

How do you incentivize the crowd to participate? Please write 1-3 paragraphs giving the specifics of how you incentivize the crowd. If your crowd is simulated, then what would you need to do to incentivize a real crowd?

Workers were incentivized to participate through pay and enjoyment. A portion of our workers were recruited from MTurk and offered $0.05 per HIT completed. Our classmates were incentivized to participate through another form of pay– their class grade. The NETS 2130 staff provide the incentives for us via class participation credit.

We also aimed to have participants use the platform out of enjoyment and a sense of creative fulfillment. Many of the projects that inspired our experiment were able to successfully harness the creativity of individuals interested in participating in community art. Many workers participated to see what creative ideas other artists generated, and how they could build on these ideas. To achieve this, the interface was made to be user-friendly to provide an intuitive experience to the user. The platform was designed to highlight the collaborative nature of the artwork by showcasing community contributions, reinforcing a sense of shared purpose and achievement. Users were able to see how the image evolved over time, as well as people's current ideas on how to improve the image.

We also included a tutorial link, detailed instructions, as well as different ways to contribute to the project.

Did you perform any analysis comparing different incentives?

No, our group did not perform any analysis comparing different incentives.

What the crowd gives you

What does the crowd provide for you?

The crowd gives us different perspectives on what art is, as well as the skills required in prompt writing for generating winning images. Although users were given the same image each day, all workers produced variable results and focused on different parts of each image. This opened a diverse selection of images to vote on during the voting stage, in which users were tasked with determining the “best image of the day”. In addition to providing images, users also provided prompts, as we were particularly interested in analyzing the chain of prompts used to get from one image state to another. We also collected various metadata on each image such as the date, creator_id, number of upvotes and downvotes, and the number of images generated per user. This data was helpful in doing preliminary exploratory data analysis such as user engagement behavior analysis as well as prompt comparisons.

Question assigned to the following page: 1

Is this something that could be automated?

While prompt writing can be outsourced to LLMs, the entire workflow still requires a human to manually input answers and earn rewards. Both of the main tasks—image generation and voting—strictly relied on human creativity/attention to detail and human judgment.

If it could be automated, say how. If it is difficult or impossible to automate, say why.

It's important to note that while prompt writing could be outsourced to an LLM, we still require the user to manually interact with the canvas to specify the parts of the image to erase, which would be difficult to automate without external/third-party browser tools. For the voting stage, it is difficult to automate votes when there is no specific metric to measure what makes an image better or worse. We trust the crowd to use their best judgment to select the best images.

Did you train a machine learning component from what the crowd gave you?

Yes, we trained a Random Forest Classifier on predicting the winner based on features and embeddings.

If you trained a machine learning component, describe what you did.

We implemented a Random Forest Classifier to predict winning prompts. The feature space combined 1,536-dimensional text embeddings generated using OpenAI's text-embedding-ada-002 model with creativity scores derived from GPT-4 analysis. We trained the model on 80% of the data while preserving 20% for evaluation. The classifier utilized default scikit-learn RandomForestClassifier parameters

Did you analyze the quality of the machine learning component? For instance, did you compare its quality against crowd workers using an n-fold cross validation?

Our Random Forest model achieved 93% overall accuracy. While the model perfectly predicted non-winning prompts (100% recall), it completely failed to identify winning ones (0% recall and precision). Due to a severe class imbalance and limited positive samples (n=3), we determined that further validation methods like n-fold cross-validation or comparisons against human performance would not yield meaningful insights. We instead focused on alternative analytical approaches.

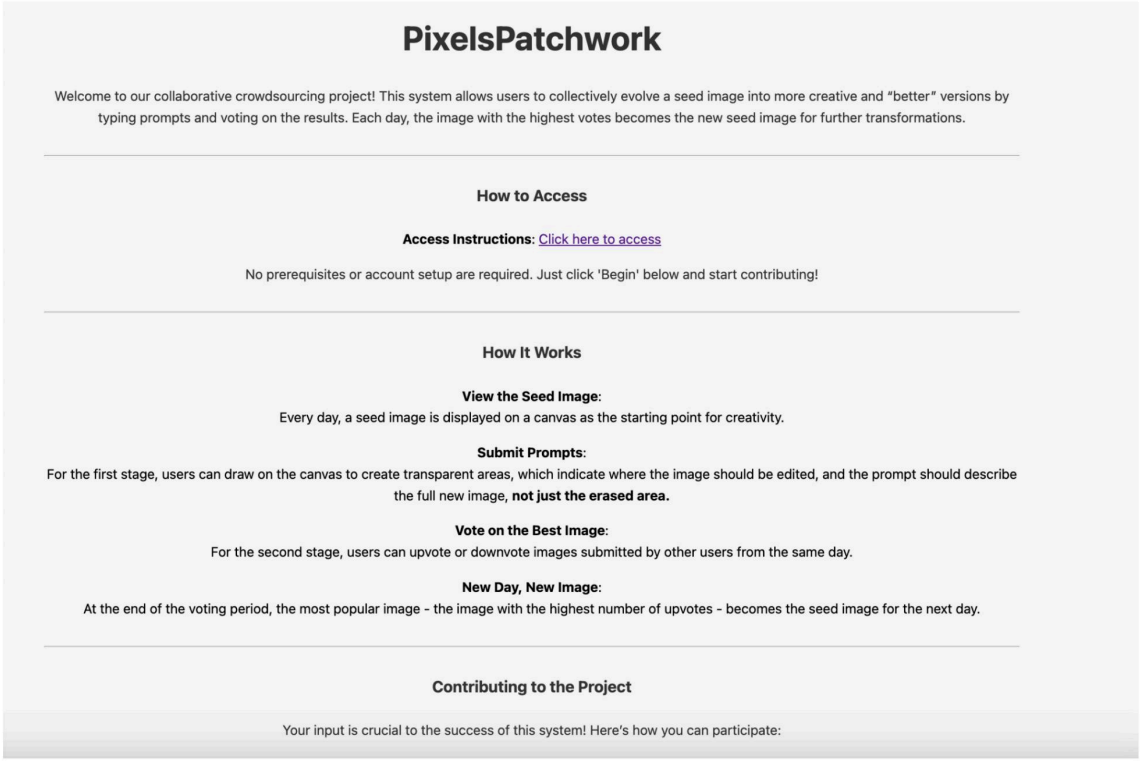
Did you create a user interface for the crowd workers? Answer yes even if it's something simple like a HTML form on MTurk.

Yes, we built a user interface using HTML and CSS. We handled various logic and interactivity functionalities with Javascript.

If yes, please include a screenshot of the crowd-facing user interface in your report. You can include multiple screenshots if you want.

Question assigned to the following page: 1

Landing page



PixelsPatchwork

Welcome to our collaborative crowdsourcing project! This system allows users to collectively evolve a seed image into more creative and "better" versions by typing prompts and voting on the results. Each day, the image with the highest votes becomes the new seed image for further transformations.

How to Access

Access Instructions: Click here to access

No prerequisites or account setup are required. Just click 'Begin' below and start contributing!

How It Works

View the Seed Image:
Every day, a seed image is displayed on a canvas as the starting point for creativity.

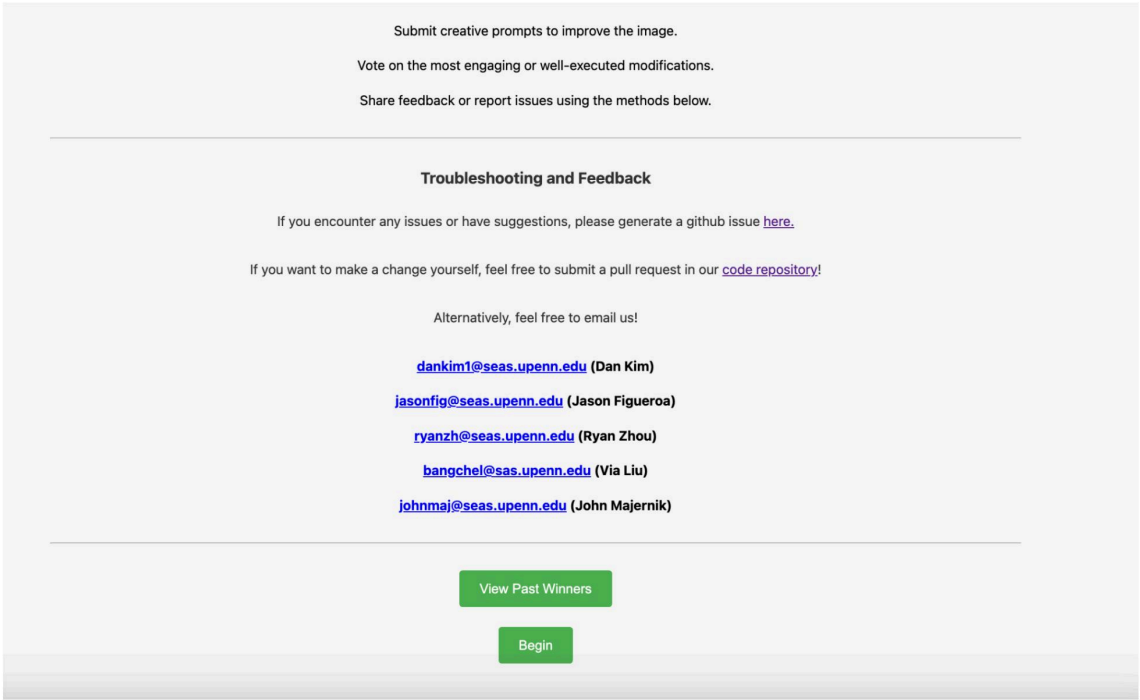
Submit Prompts:
For the first stage, users can draw on the canvas to create transparent areas, which indicate where the image should be edited, and the prompt should describe the full new image, **not just the erased area.**

Vote on the Best Image:
For the second stage, users can upvote or downvote images submitted by other users from the same day.

New Day, New Image:
At the end of the voting period, the most popular image - the image with the highest number of upvotes - becomes the seed image for the next day.

Contributing to the Project

Your input is crucial to the success of this system! Here's how you can participate:



Submit creative prompts to improve the image.

Vote on the most engaging or well-executed modifications.

Share feedback or report issues using the methods below.

Troubleshooting and Feedback

If you encounter any issues or have suggestions, please generate a github issue here.

If you want to make a change yourself, feel free to submit a pull request in our code repository!

Alternatively, feel free to email us!

dankim1@seas.upenn.edu (Dan Kim)

jasonfig@seas.upenn.edu (Jason Figueroa)

ryanzh@seas.upenn.edu (Ryan Zhou)

bangchel@sas.upenn.edu (Via Liu)

johnmaj@seas.upenn.edu (John Majernik)

View Past Winners

Begin

Question assigned to the following page: 1

Top images per day modal

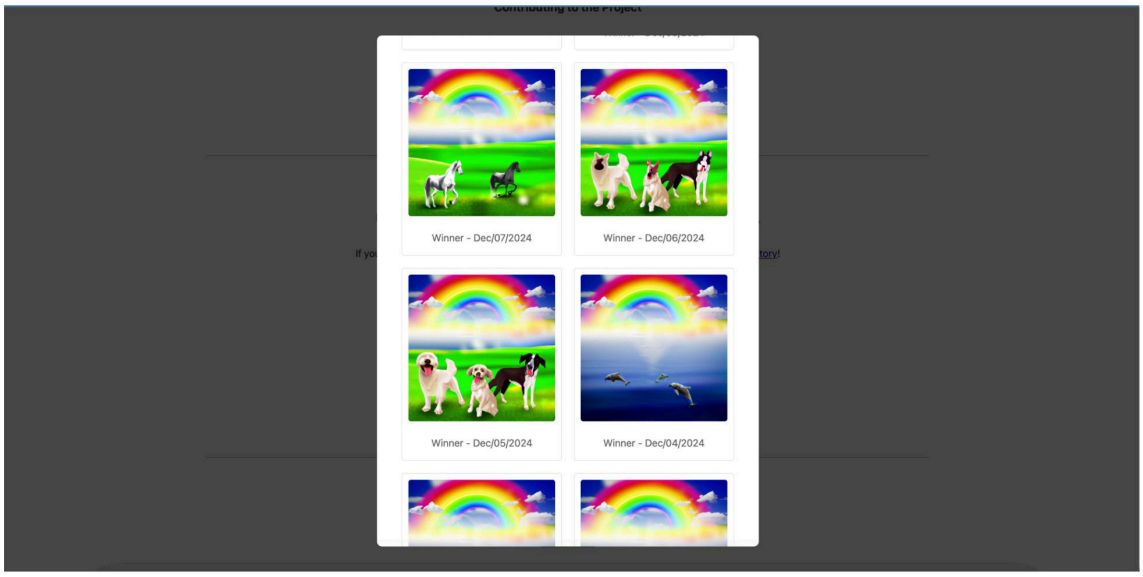


Image generation stage

Question assigned to the following page: 1

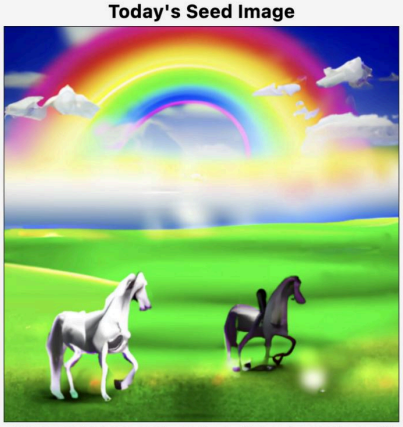
Generate a New Image

Please avoid generating more than 10 times. You may have to wait a few minutes.

The transparent areas of the mask indicate where the image should be edited, and the prompt should describe the full new image, **not just the erased area.**

You will need to play around with the prompt a few times to get the output you want. Get creative!

Today's Seed Image

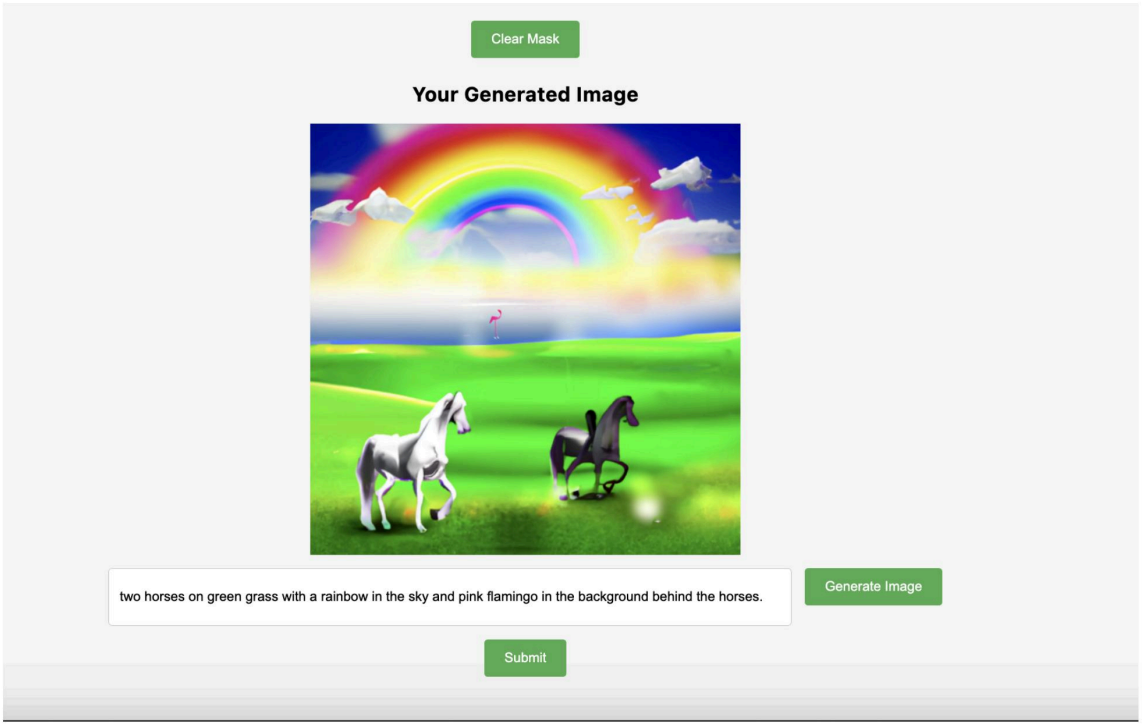


Use your mouse to draw a mask over the seed image (areas to be modified).

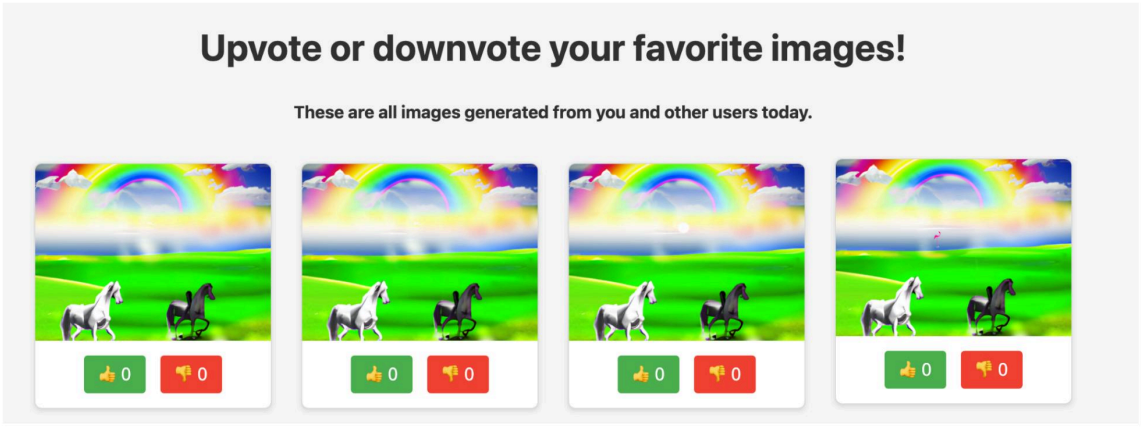
Clear Mask

Your Generated Image

Question assigned to the following page: 1

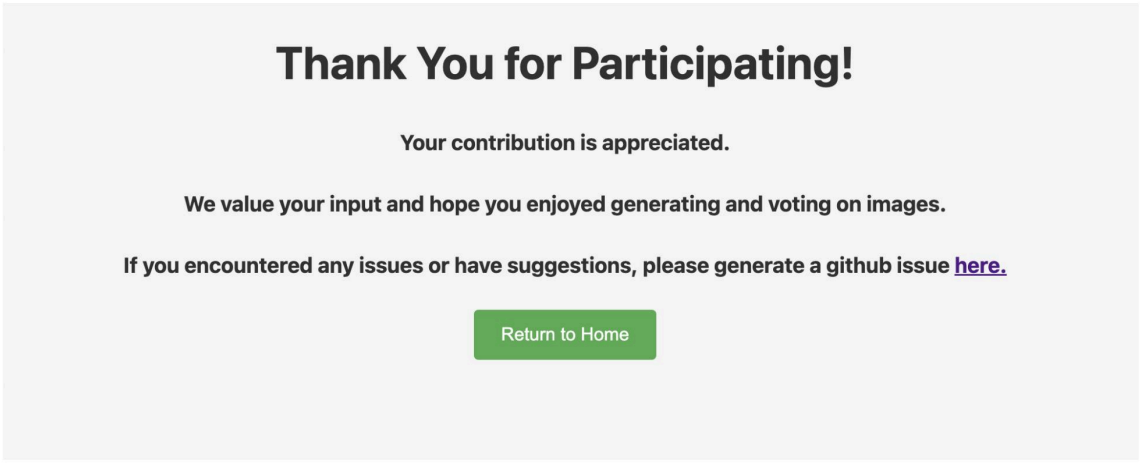


Voting stage (please note: these are not real results from the experiment!)



Thank you screen

Question assigned to the following page: 1



Describe your crowd-facing user interface. This can be a short caption for the screenshot. Alternatively, if you put a lot of effort into the interface design, you can give a longer explanation of what you did.

- Our first page is an instructions page. This page gives crowd users instructions on how to complete our experiment. Along with the instructions, the first page gives a link to a video demo of the experiment. Then there are 2 buttons, one to start the experiment and one to view past winners.
- Our second page allows users to generate images, erase part of the image, and write in a prompt to change that part of the image. The users can then see what their generated image is and either continue generating images or move on to the next step.
- On the second page, users are shown the list of images from all users of the day and can upvote images they like and downvote images they dislike.
- On the last page, users see a thank you for participating.

Ethics

Should my application exist at all?

Yes, our application PixelPatchwork has a clear purpose of fostering creativity and exploring human-AI collaboration making it a meaningful and ethically justifiable application. Its existence is valid as long as it upholds fairness, transparency, and respect for contributors' rights and data.

Does this task potentially expose workers to harm (for example, content moderation)? What effect can it have on them?

Question assigned to the following page: 1

No, this task does not expose workers to significant harm. The content involves creative image modifications and voting, which are unlikely to include harmful or sensitive material.

Are you fairly compensating the workers for their time?

Workers were compensated fairly for their participation in our experiment. Workers recruited from MTurk were given $0.05 per task completed. Classmates who participated were incentivized through the opportunity to improve their overall grade in the class. Workers who were not incentivized monetarily or by grades were rewarded through enjoyment– the project offered a unique opportunity to participate in a collaborative art project, allowing them to contribute creatively and be part of a shared artistic experience

Is your evaluation sound? Do the conclusions you reach stand up to scientific scrutiny?

Yes, the evaluation is sound and grounded in systematic data collection and analysis. Our conclusions are based on clear metrics such as voting patterns, prompt quality, and iterative improvements, aligning with established methodologies in crowdsourcing research.

Skills

Do your crowd workers need specialized skills?

No specialized skills were needed; however, English proficiency is helpful.

What sort of skills do they need?

A basic requirement would be for workers to read English in order to be able to understand what is expected of them in the task. Beyond that, workers need to be able to write complete and detailed prompts in English.

Do the skills of individual workers vary widely?

By observing the distribution of votes given to user-generated images, it is clear that the skills of the workers varied by a large margin, possibly due to the quality of the prompts generated.

If skills vary widely, what factors cause one person to be better than another?

One possible factor could be that the quality and creativity of the prompt could make one output higher in quality than another generated by a different user.

Did you analyze the skills of the crowd?

Yes, we analyzed the skills of the crowd.

Question assigned to the following page: 1

If you analyzed skills, what analysis did you perform? How did you analyze their skills? What questions did you investigate? Did you look at the quality of their results? Did you analyze the time it took individuals to complete the task? What conclusions did you reach?

We specifically performed analysis on prompt writing for winning images from different workers. We first performed exploratory data analysis by analyzing user engagement and comparing prompts. To examine how prompt quality leads to a winning image, we loaded the data, calculated net votes, identified winning images, and used OpenAI to examine prompt sentiment (negative, neutral, positive), keywords, and creativity ranking (1-10), grouped similar prompts by sentiment embeddings and quantified their underlying characteristics, used clustering to identify types of prompts and their association with winning images, and then we used a random forest model to quantify how strongly different features influence the likelihood of a prompt generating a winning image.

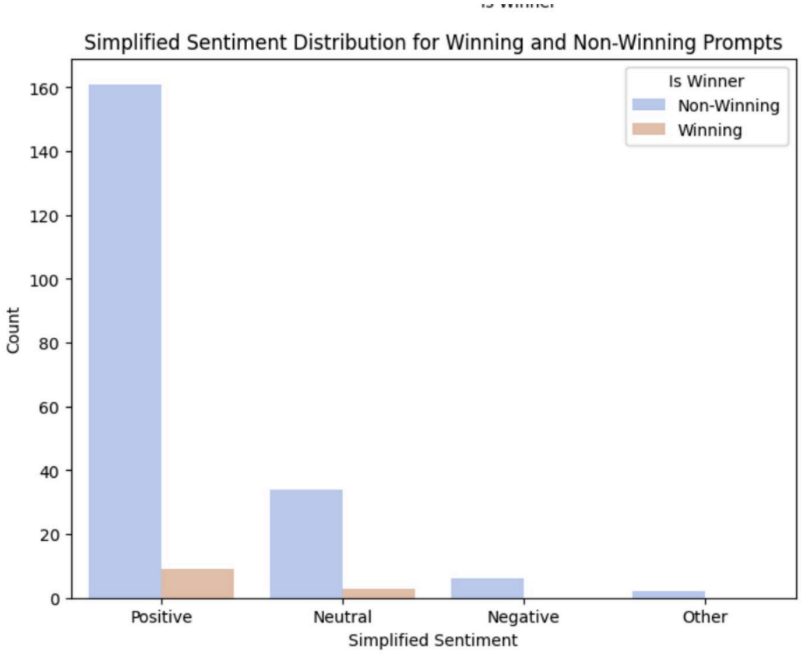
We were particularly interested in exploring specific prompts that led to the winning image, the correlation between winning prompts and general sentiments of the prompts, the correlation between specific words used and winning prompts, as well as the evolution of most commonly used words. While these questions are not entirely exhaustive for determining a strong correlation between worker skills and winning images, they provide us a fairly good understanding of the types of prompts, sentiment, and word choices that tend to be associated with successful contributions.

For the initial user engagement analysis we found that most people submitted 1 image which was expected. That being said, it was surprising to see that ~50 users did not submit any images. It's important to note that there is a difference between the number of users who visited our platform but did not complete the HIT and the number of workers who successfully submitted valid contributions, as the actual number of the former is ~180 while the latter is 133. We found that the most common words in the winning prompts had words that directly described what was in the image. This makes sense to us as prompts that use words directly in the image will be more likely to relate back to the original, which is important for the image generation model to produce desired output. The features we selected did not indicate a clear way to distinguish between photos that did well and those that did not. However, we believe that it may be possible to perform a more meaningful analysis with smarter feature engineering.

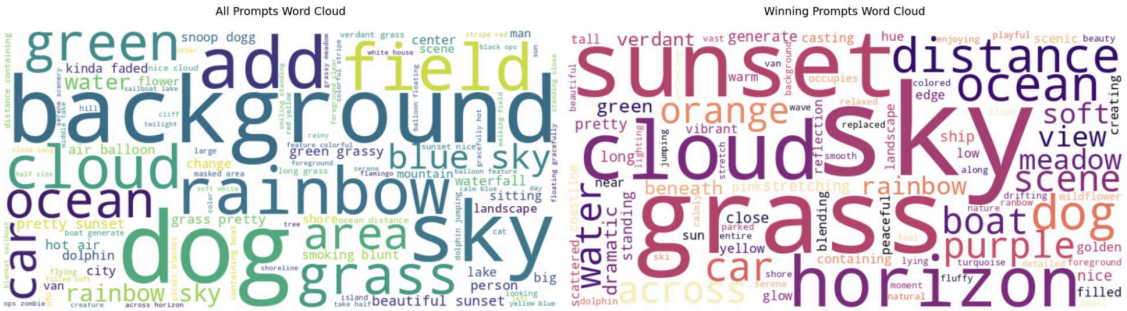
So far, we concluded that creativity alone does not predict success. We also found that winning and non-winning prompts are scattered across the embedding space, showing weak separability based on semantic characteristics, and positive prompts tend to win, but this effect is minimal given the bias towards positive responses.

Do you have a graph analyzing skills? If you have a graph analyzing skills, include the graph here.

Question assigned to the following page: 1

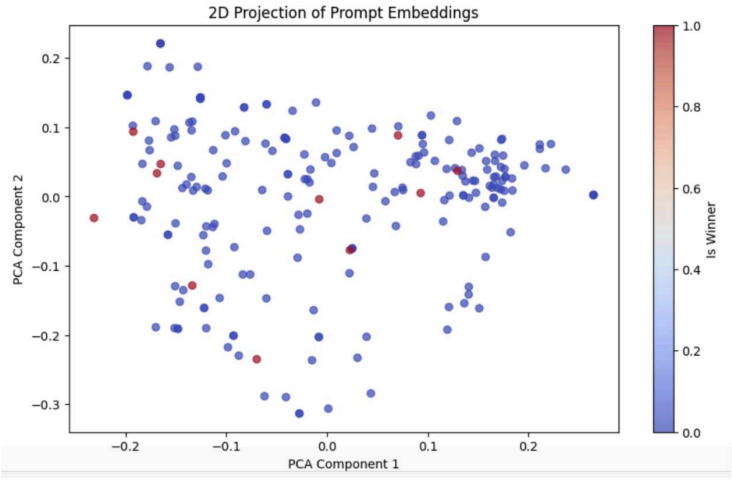


Correlation between winning images and prompt sentiment

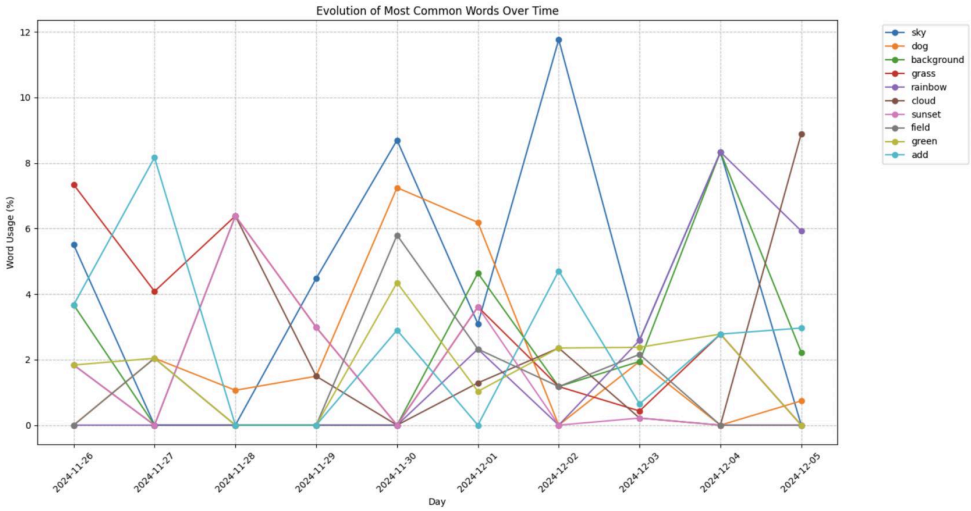


Correlation between words and winning prompts

Question assigned to the following page: 1



2D clustering of prompt embeddings



Evolution of most commonly mentioned words.

Quality Control

Is the quality of what the crowd gives you a concern?

Yes, the quality of user contributions is a concern, especially in a collaborative and iterative system where subsequent contributions rely on the quality of prior work.

How do you ensure the quality of the crowd?

We ensure quality by incorporating mechanisms such as:

Question assigned to the following page: 1

- Allowing users to vote (upvote or downvote) on images to help surface the highest-quality contributions.
- Providing transparency through the display of previously generated images to inspire better outputs.
- Limiting the number of submissions each user can make to encourage users to prioritize quality over quantity.

If quality is a concern, then what did you do for quality control? If it is not a concern, then what about the design of your system obviates the need for explicit QC? This answer should be substantial (several paragraphs long).

First, we implemented a voting mechanism that allows users to upvote or downvote images, serving as a peer review system. This mechanism empowers the community to collectively assess the quality of contributions, amplifying the visibility of superior submissions while de-emphasizing those of lower quality. By incorporating this system, we create an environment where users are incentivized to produce thoughtful and high-quality contributions that resonate positively with others. Additionally, the votes provide valuable data for identifying trends and user preferences, which can inform future refinements to the platform.

Second, we introduced sequential contributions, enabling users to iteratively refine and enhance prior submissions. This iterative approach encourages collaboration by allowing participants to build on the work of others, fostering a sense of shared ownership and continuous improvement. For example, users can modify existing images or refine associated prompts, leveraging prior creativity to achieve even better results. This design not only aligns with the principles of iterative content generation but also mitigates the risk of starting from scratch with every contribution, resulting in a more cohesive and polished output.

Lastly, we implemented a feature to prioritize image display of previous winners. By letting users view highest-voted images from the past, we ensure that users are presented with the best contributions, inspiring further creativity and engagement. This approach is particularly effective in a collaborative system, as it provides visibility to exemplary work and sets a benchmark for future submissions. Users are more likely to engage meaningfully with the platform when they can easily identify and interact with high-quality outputs.

Together, these quality control measures not only maintain the integrity of the contributions but also enhance the overall user experience by promoting the highest standards of creativity and engagement.

Did you analyze the quality of what you got back? For instance, did you compare the quality of results against a gold standard? Did you compare different QC strategies?

Question assigned to the following page: 1

While we didn't explicitly compare results against a gold standard, we relied on user voting data (upvotes and downvotes) as a proxy for quality. This data allowed us to evaluate the community's perception of quality and refine future contributions iteratively.

Given the subjective nature of art, it was difficult to assess the quality of a submission without crowd input. However, we encouraged quality contributions by providing clear instructions and prompts, implementing voting and moderation systems, highlighting exemplary contributions, and providing users the “best” image (as determined by the crowd) as a starting point for their submission.

Success of these quality control mechanisms can be observed in a comparison with an iterative approach to generating a better image. When comparing the final image generated by our platform to the efforts of an individual iteratively improving on an image in the same way, we found that people preferred the image generated by the crowd by a substantial margin.

What analysis did you perform on quality?

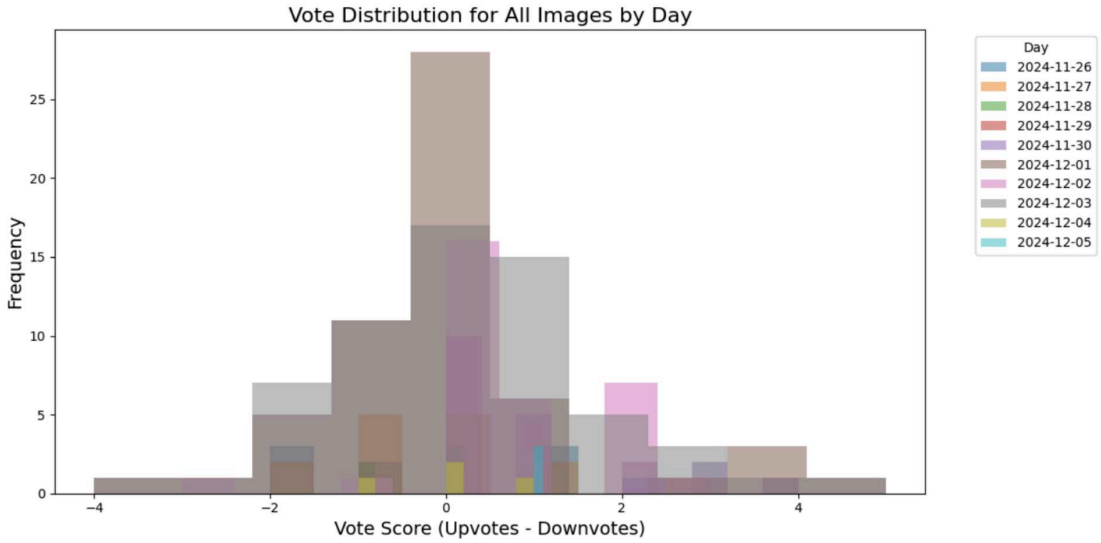
We analyzed user engagement metrics, such as voting activity, the distribution of votes across all images, and number of users over time, to measure the effectiveness of our quality control mechanisms.

What questions did you investigate? What conclusions did you reach?

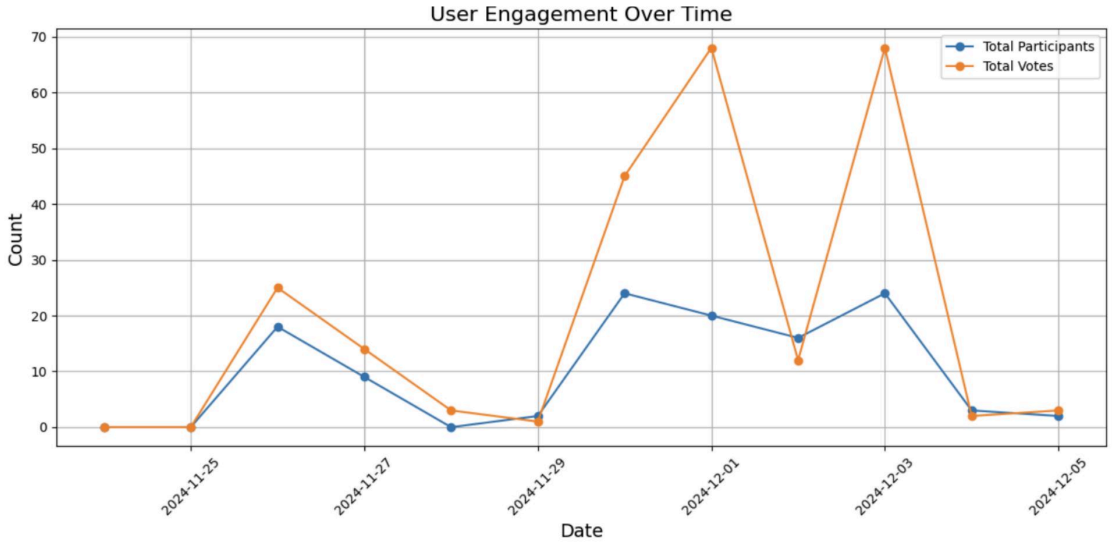
We sought to investigate the effectiveness of our voting mechanism in surfacing the highest-quality images. Based on the vote distribution, we observed that a significant number of images did not receive any upvotes or downvotes. This may indicate worker neglect, as voting was encouraged but not mandatory. As a result, many images went unrated, raising questions about whether the winning images truly reflect the crowd's collective opinion. The graphs show a skewed distribution of votes, with most vote scores clustering near zero and limited engagement on certain days. This lack of comprehensive voting suggests that the voting mechanism may need refinement to ensure more reliable and representative results.

Do you have a graph analyzing quality? If you have a graph analyzing quality, include the graph here.

Question assigned to the following page: 1



- Engagement Over Time: Tracking user activity (votes and contributions) across days



Aggregation

How do you aggregate the results from the crowd?

The results from the crowd were aggregated using the voting mechanism, where users provided upvotes or downvotes for generated images. Each image's score was calculated by tallying its total upvotes and downvotes, with the final score reflecting the community's

Question assigned to the following page: 1

consensus on its quality, or a proxy of its quality. We also aggregated the winning images across days to display the history of winning images to the user.

Did you analyze the aggregated results?

The aggregated results were analyzed to identify trends in user preferences and evaluate the quality of submissions. Specifically, we examined which types of prompts or image attributes tended to receive the highest votes. This analysis aimed to uncover patterns in community preferences, such as themes, styles, or image features that consistently resonated with users.

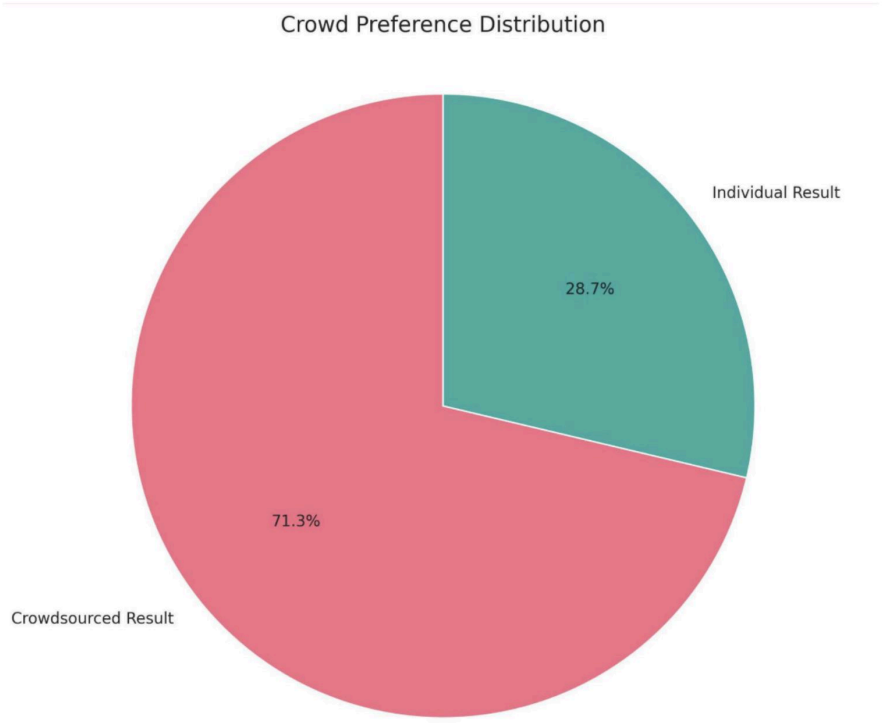
What analysis did you perform on the aggregated results? What questions did you investigate? Did you compare aggregated responses against individual responses? What conclusions did you reach?

We performed several analyses on the aggregated results, including:

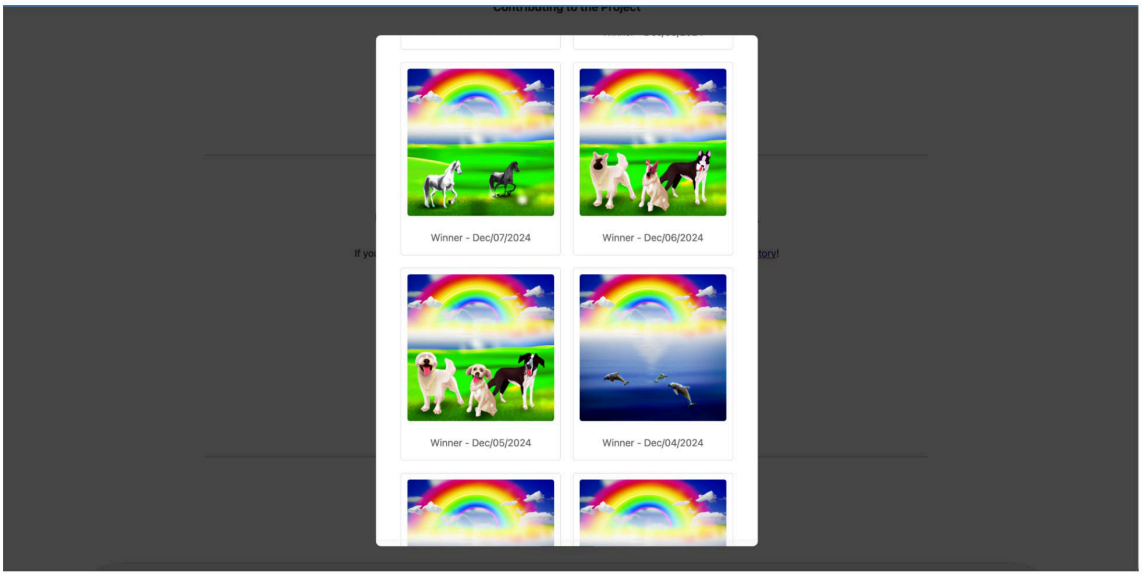
- Past winners: We aggregated a list of past winning images so users could see.
- Comparison of Aggregated vs. Individual Responses: We compared aggregated results with individual vote patterns to understand how collective feedback differed from personal preferences. This comparison highlighted the value of aggregation in surfacing the most broadly appreciated contributions.
- Vote Distribution: We investigated the distribution of upvotes and downvotes across all images to identify patterns in user engagement

Do you have a graph analyzing the aggregated results? If you have a graph analyzing the aggregated results, include the graph here.

Question assigned to the following page: 1

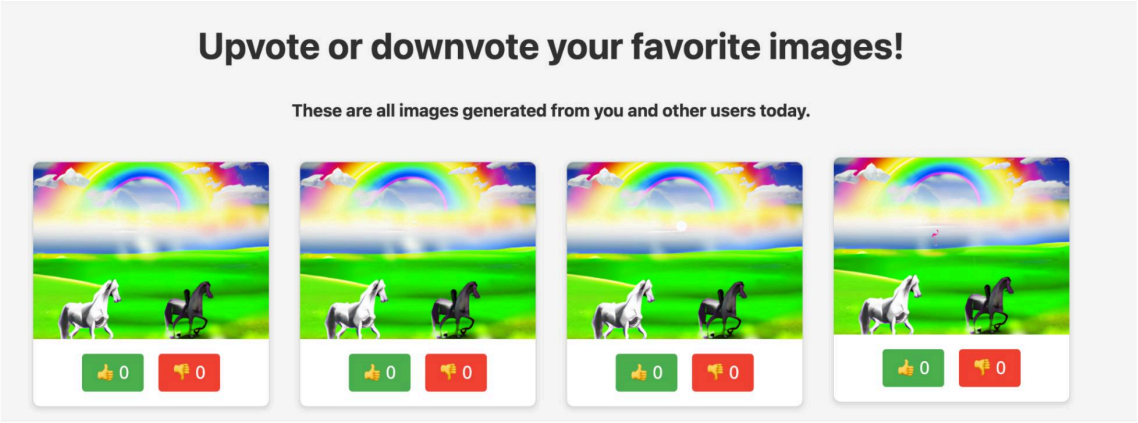


Did you create a user interface for the end users to see the aggregated results? If yes, please include a screenshot of the user interface for the end user in your final report. You can include multiple screenshots, if you want.



Aggregated past winners.

Question assigned to the following page: 1



Voting stage aggregates votes.

Describe what your end user sees in this interface. This can be a short caption for the screenshot. Alternatively, if you put a lot of effort into the interface design, you can give a longer explanation of what you did.

Captions are under each respective screenshot.

Scaling Up

What is the scale of the problem that you are trying to solve?

The scale of our project involves crowdsourcing creativity and decision-making in a structured and iterative manner from different users. Specifically, the project involves a continuous influx of user-generated images, sequential edits, and feedback through voting. As the platform scales, this could mean thousands to millions of images contributed and evaluated over time, which would require more thoughtful design for data storage management, retrieval, and processing. While our platform caters to any audience, our application is mostly well-suited for English-speaking workers as the UI's text is only in English; however, the image generation model can likely handle different languages (though the quality may vary depending on language used). As the number of participations per day grows, so does the need for scalable infrastructure to support real-time interactions, storage and retrieval of large image datasets, and computationally intensive operations like api calls to AI image generation models.

Would your project benefit if you could get contributions from thousands of people?

Yes, as this would increase our sampling size and provide more data points.

If it would benefit from a huge crowd, how would it benefit?

Question assigned to the following page: 1

It would allow us to draw more comprehensive and generalizable conclusions about the correlation between the quality of user prompts, the output of said prompts, and the scalability and robustness of our system for hundreds of thousands of users.

What challenges would scaling to a large crowd introduce?

Scaling to a large crowd introduces challenges in maintaining system performance, data quality, and effective user engagement. With an increase in users, the infrastructure must handle higher volumes of requests, potentially leading to slower response times and server overloads. For example, if there are thousands of images generated in a day, then our database would need to constantly perform full scans over tables to retrieve the image_ids, which would be excessive and computationally expensive at scale. Ensuring that the platform remains responsive and scalable may require optimized database queries, caching frequently used results, indexing, load balancing, and distributed architecture.

Additionally, a larger crowd amplifies issues with data quality, such as spam, malicious inputs, or noise in aggregated results. Better moderation mechanisms and robust algorithms to filter low-quality contributions become critical. Finally, keeping users engaged while preserving a sense of community and individual contribution requires thoughtful UI/UX design and incentives, as the experience may feel impersonal or overwhelming in a larger crowd. Scaling the frontend would require additional consideration and possibly a migration to React instead of HTML to facilitate development of new features and reusable components.

Did you perform an analysis about how to scale up your project? For instance, a cost analysis?

No, we did not perform any explicit quantitative analysis on how to scale up the infrastructure. However, we would need to invest more in OpenAI and possibly migrate beyond the AWS free tier if we were to support more users, and this would require a substantial upfront investment.

What analysis did you perform on the scaling up?

This analysis is not applicable for our project.

What questions did you investigate? What conclusions did you reach?

While we did not investigate this portion of the project deeply, we did consider scalability constraints such as database load and optimizing SQL queries to handle large throughput, concurrent operations, and strains due to increased database loads. Due to the technical feasibility of the tasks, our group was unable to devote the necessary time to enhance the system design and therefore, cannot make comprehensive conclusions about this.

Question assigned to the following page: 1

Do you have a graph analyzing scaling? If you have a graph analyzing scaling, include the graph here.

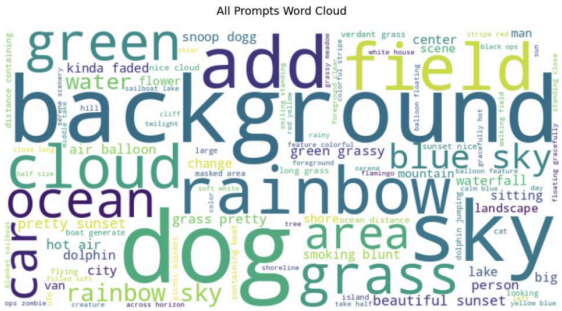
Not applicable.

Project Analysis

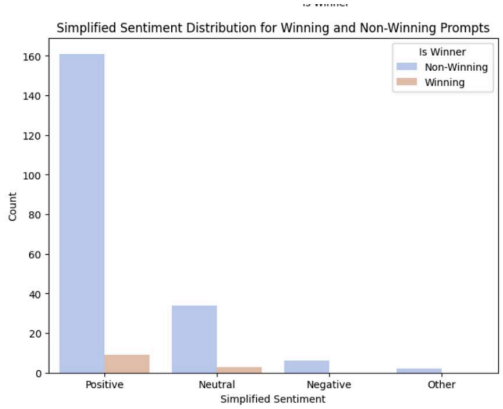
Did your project work? How do you know? Analyze some results, discuss some positive outcomes of your project.

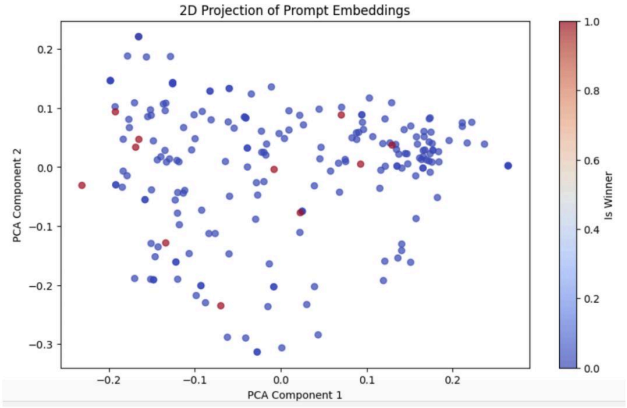
Yes, the project's goal of achieving AI-generated art through crowdsourcing was a success. Our participation and engagement were a huge indicator as the project recorded 133 unique responses and successfully engaged participants in both generating and voting on images. Our workers created 216 unique images with various prompts leading to diverse interpretations of the seed image. The voting mechanism worked flawlessly in selecting a "winning image" daily which iteratively shaped the profession of the artwork. Analyzing the results and looking at the positive outcomes of the project, participants generated varied prompts which resulted in highly diverse outputs that are representative of human creativity in modifying AI-generated art. Analyzing the prompt quality revealed that descriptive and detailed prompts often correlated with winning images which highlighted how clarity and specificity enhanced winning outcomes. From doing NLP analysis such as clustering and sentiment evaluation revealed trends such as the impact of positive wording on success rates. Finally, the project demonstrated that iterative contributions coupled with community-driven voting could refind and enhance the quality of the final art output.

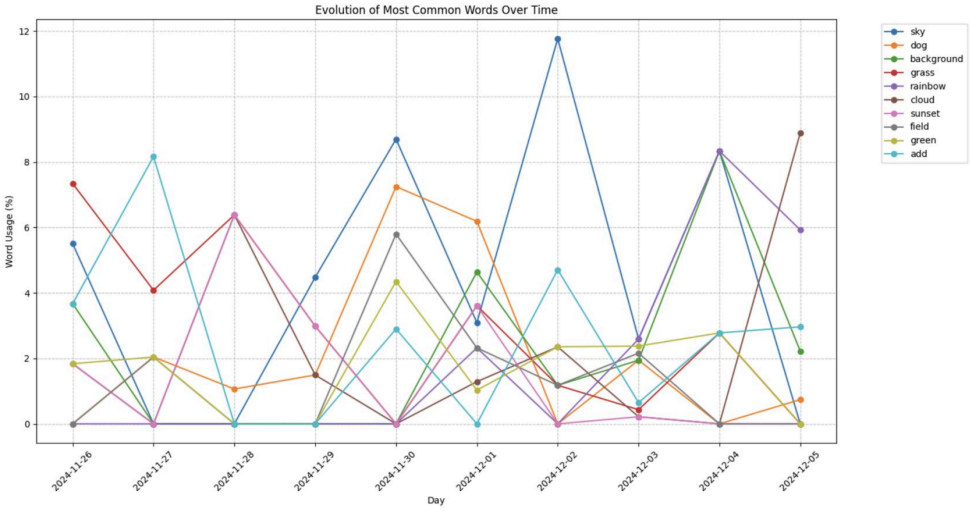
Do you have a graph analyzing your project? If you have a graph analyzing your project, include the graph here.



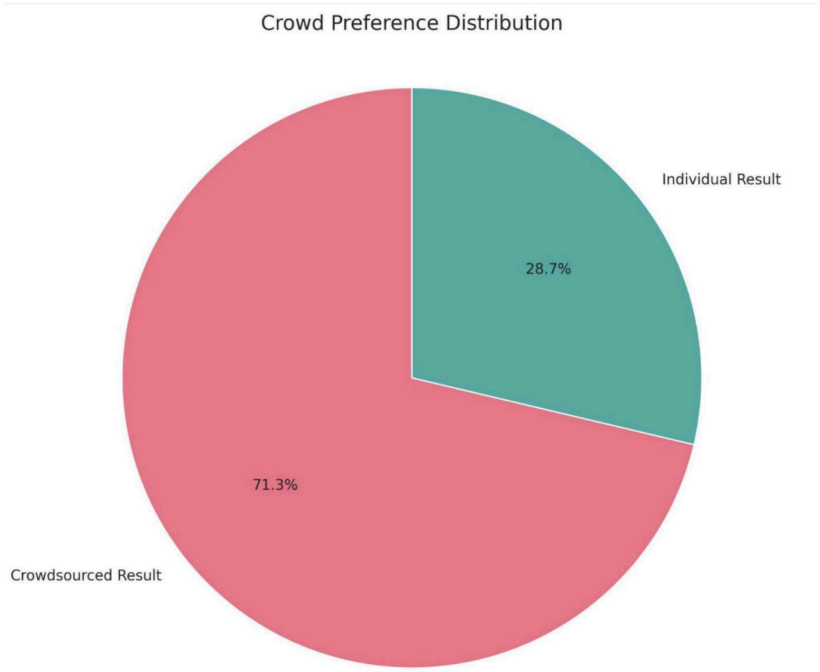
Question assigned to the following page: 1


Simplified Sentiment Distribution for Winning and Non-Winning Prompts
Is Winner
Non-Winning
Winning
Count
Positive
Neutral
Negative
Other
Simplified Sentiment


2D Projection of Prompt Embeddings
PCA Component 2
PCA Component 1
Is Winner


Evolution of Most Common Words Over Time
Word Usage (%)
Day
sky
dog
background
grass
rainbow
cloud
sunset
field
green
add

Question assigned to the following page: 1



What were the biggest challenges that you had to deal with?

While doing this project, one of the biggest challenges we had to deal with was precise image modification. We wanted to enable the AI-generated image to modify only specific portions of the image (up to the user's choice) instead of the entire image. This required significant coding and debugging efforts to implement precise masking features. We had to design a custom canvas that allowed users to highlight the area they wanted to change and integrate this input with the AI models to ensure that the output respected these constraints. The process involved fine-tuning API calls, testing the interaction between the masking tool and the AI model, and resolving unexpected errors in image processing pipelines. Another challenge was processing contributions and votes in real time. We wanted to generate the next day's seed image in real-time and that required more efficient data handling.

Were there major changes between what you originally proposed and your final product?

Yes, there were notable changes between the initial proposal and the final product.

If so, what changed between your original plan and your final product?

Initially, the project proposed modifying the entire image with each iteration, but the final product implemented a more sophisticated system that allowed users to mask specific portions of the image for targeted edits. This enhancement added complexity but was integral in improving the user experience. Additionally, the original plan envisioned using

Question assigned to the following page: 1

entirely new seed images for each iteration; however, the final implementation used the "winning image" from the previous round, as determined by the voting process, as the next day's seed image. This shift created a dynamic, iterative feedback loop where the community's preferences directly influenced the progression of the artwork. While the initial idea focused on individual contributions, the final product emphasized iterative collaboration, enabling the crowd to refine and build upon previous outputs collectively. To support this evolution, the workflow was redesigned to include mechanisms for aggregating, displaying, and reintroducing contributions. Lastly, as the project evolved, a stronger focus on quality control emerged, leading to measures such as limiting user submissions and prioritizing high-voted images to ensure the quality of iterative outputs.

What are some limitations of your product? If yours is an engineering-heavy project, what would you need to overcome in order to scale (cost/incentives/QC…)? If yours was a scientific study, what are some sources of error that may have been introduced by your method.

Our product has a few limitations, mainly related to scalability and quality control. Scaling to a larger audience would require upgrades to backend systems, storage, and processing to handle increased participation smoothly. While costs for incentives and AI model usage may grow with scale, these are manageable with strategic budgeting and efficient use of resources. Maintaining consistent quality is another focus area; while the voting mechanism effectively surfaces the best outputs, further enhancements like AI moderation or gamified contributions could refine the process. Some variability in AI responses and voting preferences is natural in a creative project but does not detract from its overall success. Expanding support for multiple languages and cultural contexts could broaden accessibility, while refining feedback loops would ensure high-quality outcomes as the project grows.

Did your results deviate from what you would expect from previous work or what you learned in the class?

Our results aligned with many principles from class and prior work but also revealed some deviations. The iterative structure and voting mechanism successfully demonstrated how crowdsourcing can produce diverse and creative outputs, consistent with theories of collaborative work. However, the impact of prompt specificity on output quality was stronger than anticipated, emphasizing the importance of clear instructions for contributors.

If your results deviated, why might that be?

Unexpectedly, the iterative process showed slower convergence toward a consensus "best image" than predicted, likely due to diverse user preferences and varying interpretations of quality. Additionally, while previous studies suggest that iterative editing often improves outcomes, we observed that early low-quality contributions occasionally propagated through iterations, affecting the final result. These deviations highlight the unique challenges that

Question assigned to the following page: 1

come with integrating human creativity with AI-driven tools in a collaborative iterative framework.

Technical Challenges

Did your project require a substantial technical component? Did it require substantial software engineering? Did you need to learn a new language or API?

Yes, it did require a substantial technical component, which involved building a web platform using Flask, HTML/CSS/Javascript, OpenAI API, and a cloud database.

For the frontend, we built 4 pages, a landing page which instructs the user on how to use our platform, image generation page, voting stage page, and a goodbye page. The image generation page included a custom canvas that a worker could interact with. All pages and elements were built using simple HTML and CSS, and all asynchronous functions and other interactive features were implemented using javascript.

For the database, we spent considerable time thinking about what specific data we wanted to collect from the user. Our schema included 3 tables, including User(user_id, created_at), Day(date, seed_image_id, total_votes, total_participants), and Image(image_id, s3_path, prompt_text, creator_id, day, upvotes, downvotes). In the Image table, the creator_id was a foreign key referencing user_id in User and day was also a foreign key referencing the date in Day. Our database was hosted on an AWS instance.

For the backend, setting up the Flask backend required reading documentation to become familiar with the most optimal architecture and software design. We also needed to learn how the dall-e 2 image generation model used both a mask and a prompt to modify an image through the OpenAI documentation.

If the project required a substantial technical component, describe the largest technical challenge you faced.

The biggest technical challenge we faced was ensuring the image modification/generation pipeline enabled direct image modification with the user interface while maximizing user experience. Specifically, creating the canvas so the seed image would appear on render and allow the user to paint over required a custom solution, as not many open-source solutions satisfied our use case.

How did you overcome this challenge? What new tools or skills were required? Feel free to nerd out a bit, to help us understand the amount of work that was required.

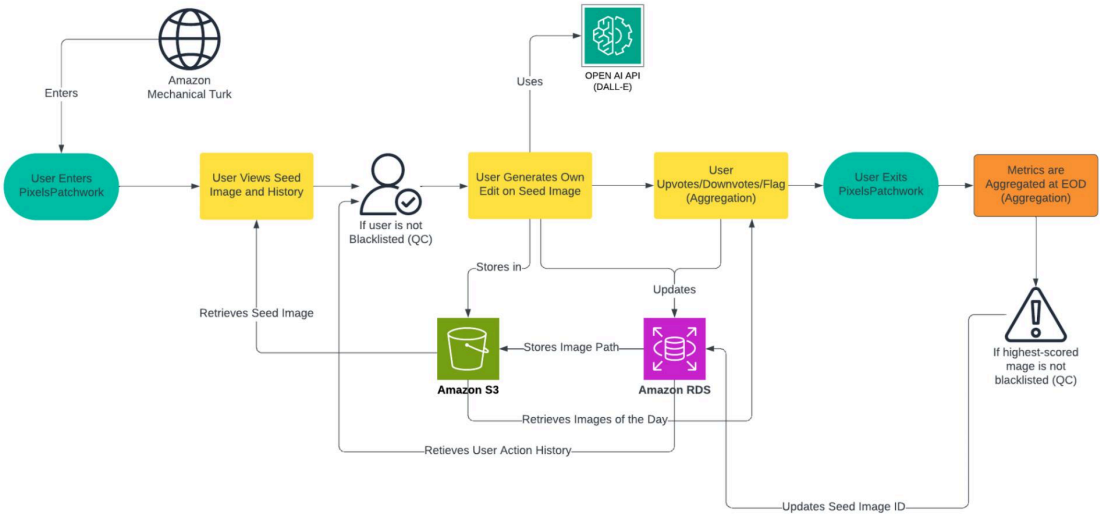
To overcome this challenge, we developed a custom implementation for the canvas functionality, leveraging HTML5 Canvas APIs to ensure seamless rendering and interaction. We used a combination of JavaScript and advanced event handling to allow users to draw

Question assigned to the following page: 1

directly on the canvas, enabling them to mask areas for editing with precise control. Additionally, we integrated DALL-E's API for image generation, which required adapting the output from the canvas to a format compatible with the API's requirements. This involved capturing the user's masking data as a PNG file and ensuring it aligned correctly with the seed image dimensions. Debugging the canvas rendering was particularly challenging due to inconsistencies in browser behavior, so we utilized tools like Chrome DevTools for real-time performance debugging and ensured cross-browser compatibility.

This process also required a deep dive into working with image manipulation libraries and learning how to handle base64 encoding for transferring the masking data. We iterated on the design through UI testing to fine-tune the experience, balancing performance and usability. While the open-source libraries we evaluated didn't meet our needs out of the box, this challenge became an opportunity to enhance our skills in building custom user interfaces and bridging the gap between creative tools and AI-powered workflows.

Do you have any screenshots or flow diagrams to illustrate the technical component you described? If so, include the graph here.



Project Technical Workflow