Final Report

● Ungraded

1 Day, 23 Hours Late

Group

Erik Wei
Eric Wang
Justina Lam
...and 2 more
View or edit group

Total Points

- / 15 pts

Question 1

Report 15 pts

Question assigned to the following page: 1

Frontline Final Report

Justina Lam, Ethan Ma, Kaily Liu, Erik Wei, Eric Wang

1. Project Information

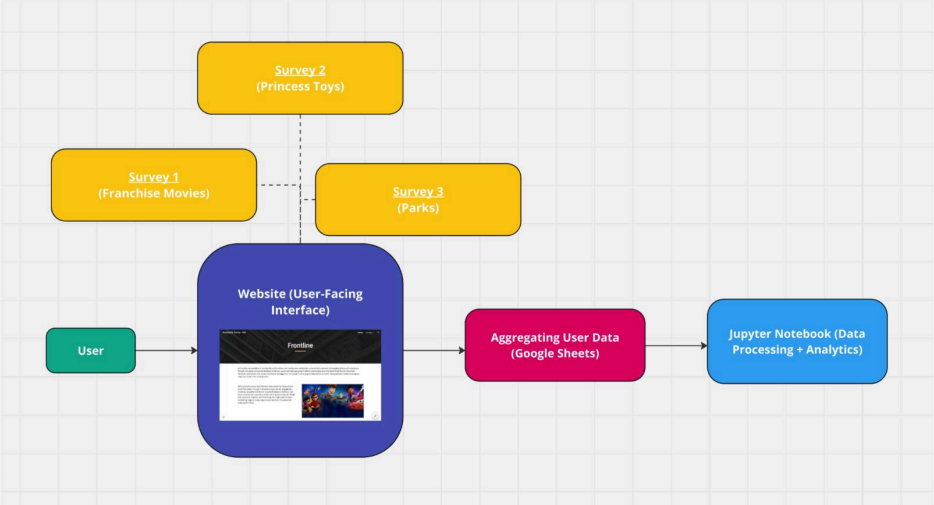
- Ethan - CIS
- Eric - ROBO
- Justina - CIS
- Erik - CIS
- Kaily - NETS

Frontline is a crowdsourcing platform for rating generated ideas! In this day and age, it can be difficult to determine what ideas are valuable and what ideas are bunk! While generative models can assist with idea generation, it's always helpful to have a human in the loop to keep things reasonable. Our project was structured in the following way: users were asked to take a series of surveys which aimed to evaluate their opinions on a variety of different ideas generated by AI. The project also included a gamification aspect where users would be able to bet point values on the projects they found most promising, and to reward them with part of the upside if the idea happened to do well. Though software does exist for facilitating or aiding in idea generation such as QMarkets and Ideanote, there don't seem to be many projects out there which follow a crowdsourcing approach, specifically for evaluating ideas. In essence, Frontline leverages crowdsourcing for evaluating the outputs of idea generation processes. Because the process of data collection was fairly simple, the majority of our effort went into performing in-depth analysis of collected data. We ended up finding that although we were interested in how the ideas performed, we also were very interested in the behavior of users and how they utilized the three rating scales differently.

Question assigned to the following page: 1

The architecture for the entire process was as follows:

- Ideas are generated via an LLM, such as ChatGPT or Claude
- Crowd users are given an allotment of tokens and surveyed for their opinions on generated ideas
    - Users place bets on generated ideas which are deemed to be the most profitable
    - They were also given two other scales to evaluate the ideas on, namely a classic 1-10 rating system, and a price-you-would-pay rating system.
- A quality control system is used to filter out candidates who were not following the rules of the survey or not paying attention (e.g. by betting more than 100 credits), and we used a self-evaluation question to see how qualified user's opinions were, weighting their responses accordingly.
- Aggregations are performed on user data to reveal insights on which idea is likely to profit, as well as perform exploratory data analysis on user behaviors within the scope of these rating systems.
- After the implementation of the product, the idea is that users are compensated monetarily based on a prize pool if the idea is actually implemented and becomes profitable.



Link to video:
https://drive.google.com/file/d/1Q6FwEAQQqdyy0Zrf6mzkvK5n2b4wD3JF/view?usp=sharing

During this report we'll be deep diving into the **"aggregation"** and **"what the crowd gives you"** sections.

Question assigned to the following page: 1

2. The Crowd

The crowd we utilized for our project consisted solely of members of the NETS 2130 class at the University of Pennsylvania. We would have liked to have utilized MTurk to facilitate our crowdsourcing, but unfortunately, we had some difficulty with getting our MTurk accounts verified. As a result, we used a more limited and homogenous crowd which may have resulted in fewer viewpoints when examining ideas and thus reduced the positive impact of leveraging crowd intelligence. After the “Be the crowd” exercise, we were able to accumulate 78 data points across our three surveys, with each data point consisting of a full survey response.

3. Participant Incentives

The primary participant incentive that we envisioned would be the ability for members of the crowd to bet tokens on ideas that they thought would be successful and receive tokens upon the launch of products as a percent of product revenue, with payouts determined by the resulting revenue. Given that none of the products suggested in our example surveys have made it to launch, it’s currently not possible for participants to actually receive any compensation. However, a complete version of the platform would include some service to handle the distribution and accumulation of tokens for each user. Given that risk-based or gambling systems generally drive a huge amount of engagement, even in cases where the resulting payouts aren’t monetary (such as in mobile gaming), we’re sure that such an incentive would be strong enough to drive a high level of engagement.

Additionally, during our in class session we asked several of the participants who filled out our survey whether they enjoyed giving their opinion and if so, why. The general consensus was that people like to be opinionated and judge ideas - it is fun to play with the hypotheticals, *especially* if you are knowledgeable about or a big fan of a brand, franchise, or product. Some people said that they enjoyed this as a form of self-expression, and others liked that it gave them a feeling of expertise and that their opinions were valued. Overall, we were not surprised that there was an inherent enjoyment in exploring such ideas and evaluating them.

Due to the fact that we used primarily qualitative and hypothetical incentive systems, we didn’t feel like it was necessary or very useful to perform an analysis comparing different incentives.

4. What The Crowd Gives You

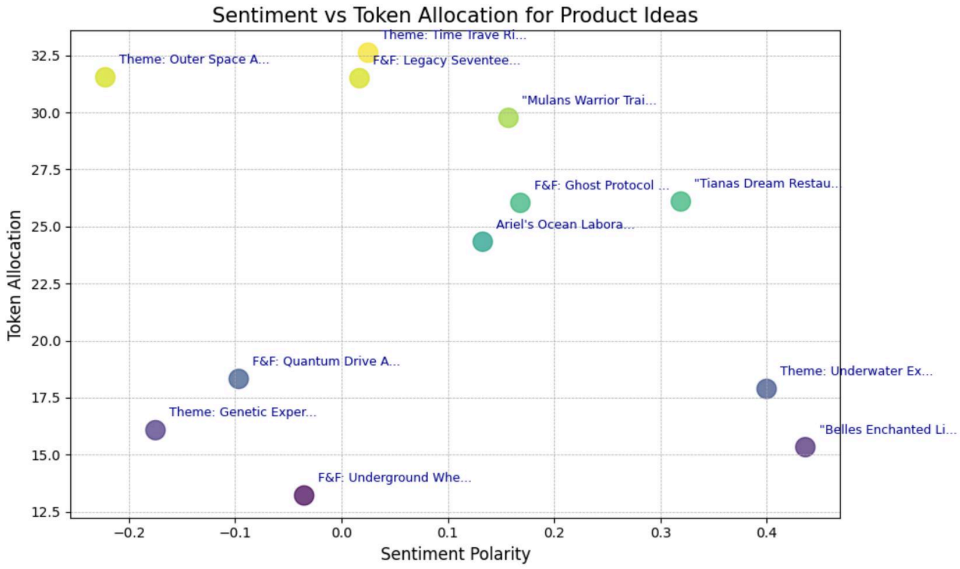
The crowd provides a variety of data including prioritization data (based on their betting amounts on each product), preference ratings (by giving a score 1-10), and qualitative feedback through the general idea feedback. Though we’re doubtful that the process could be entirely automated, there definitely are parts that could be automated. The data aggregation, quality control, and analytics could be automated by running a Collab notebook. The data collection is

Question assigned to the following page: 1

already automated through use of Google Forms. However, the actual feedback portion needs real-feedback from the crowd that cannot be automated.

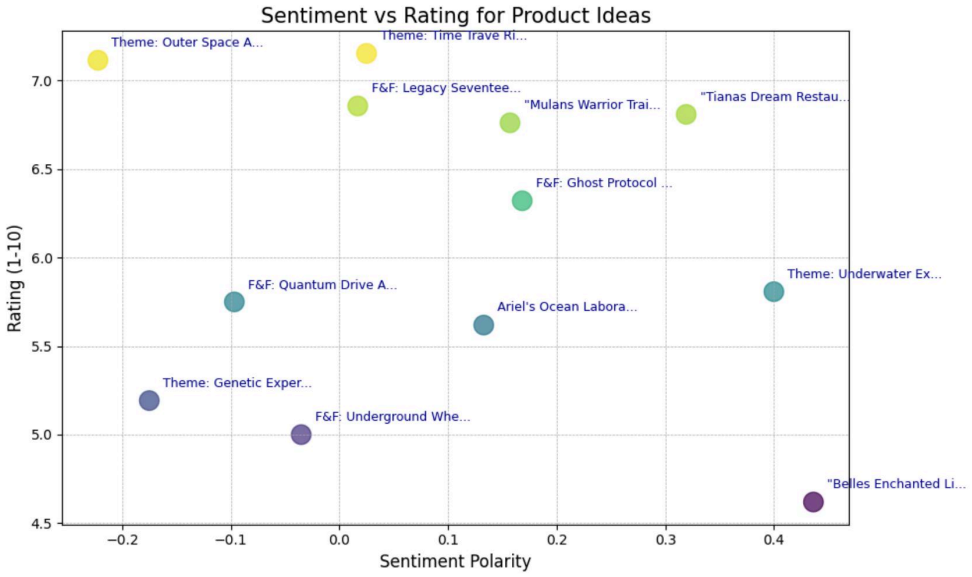
AI/ML Component:

We had 2 AI/ML components. First, we were interested in if the actual wording of the ideas affected how people generally bet. To do this, across every idea we averaged all votes (point betting amount and rating (1-10)), and also used sentiment analysis in order to generate a polarity value for each idea. This was the resulting graph:



Interestingly, the most highly-bet ideas had some negative polarity. In fact, the very positive ideas (>0.3) generally were not in the top 5 highly-bet. However, ideas that had negative polarity were generally more polarizing, either scoring very high or very low. We can observe that in this case, the sentiment that certain ideas were assigned had little to do with the perceived value that users assigned them. Below, we can see that a similar result was observed when comparing sentiment polarity with user ratings as our metric for the true user sentiment. Here, we also saw that ideas with extremely low sentiment polarity received high ratings while there likewise existed ideas with high sentiment polarity and low ratings. In fact, the second highest rated idea had the second lowest sentiment polarity, while the lowest rated idea had the highest sentiment polarity.

Question assigned to the following page: 1



Second, ss a bit of a more fun exploratory analysis, since the nature of rating new ideas is that there is no correct answer, we wanted to explore what we could actually compare our results to, and one of those was ChatGPT:

ChatGPT Rankings vs Crowdsourced Rankings (Movies - Fast and Furious)

ChatGPT	Crowdsourced (weighted)
1. "Ghost Protocol" (Idea 2)	1. "Ghost Protocol" (Idea 2)
2. "Legacy" (Idea 1)	2. "Legacy" (Idea 1)
3. "Quantum Drive" (Idea 3)	3. "Quantum Drive" (Idea 3)
4. "Underground" (Idea 4)	4. "Underground" (Idea 4)

ChatGPT Rankings vs Crowdsourced Rankings (Rides)

ChatGPT	Crowdsourced (weighted)
1. Void Runner (Idea 2)	1. Void Runner (Idea 2)
2. The Time Twister (Idea 1)	2. The Time Twister (Idea 1)
3. Oceanic Odyssey (Idea 3)	3. Oceanic Odyssey (Idea 3) (tied)
4. BioBreak (Idea 4)	4. BioBreak (Idea 4) (tied)

ChatGPT Rankings vs Crowdsourced Rankings (Toys)

Question assigned to the following page: 1

ChatGPT	Crowdsourced (weighted)
1. Tiana's Dream Restaurant (Idea 4)	1. Mulan's Warrior Training Academy (Idea 3)
2. Belle's Enchanted Library Creator (Idea 2)	2. Tiana's Dream Restaurant (Idea 4)
3. Ariel's Ocean Laboratory (Idea 1)	3. Ariel's Ocean Laboratory (Idea 1)
4. Mulan's Warrior Training Academy (Idea 3)	4. Belle's Enchanted Library Creator (Idea 2)

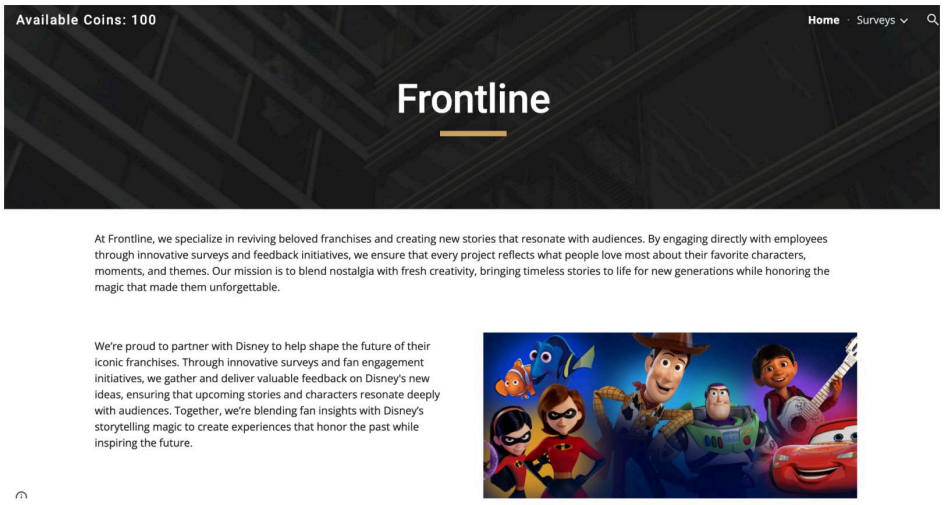
The nature of our project, rating new ideas, assumes that there is no correct answer for the best idea. However, we still wanted to compare our results (seen in the aggregation section) to something. The above results from ChatGPT were done over 10 trials and the final rankings from ChatGPT were calculated from a simple point system where a 1st gives 4 points, 2nd gives 3 points etc. Remarkably, ChatGPT is actually quite consistent in its opinions and would give the same rankings even when retried multiple times. Surprisingly, the ChatGPT rankings also matched up very well with our weighted crowdsourced opinions. In many ways, much of the data that ChatGPT produces results on is based on crowdsourcing of data from across the web. While this was just a low sample size experiment it is interesting and seems unlikely that the rankings would exactly line up for two of our surveys.

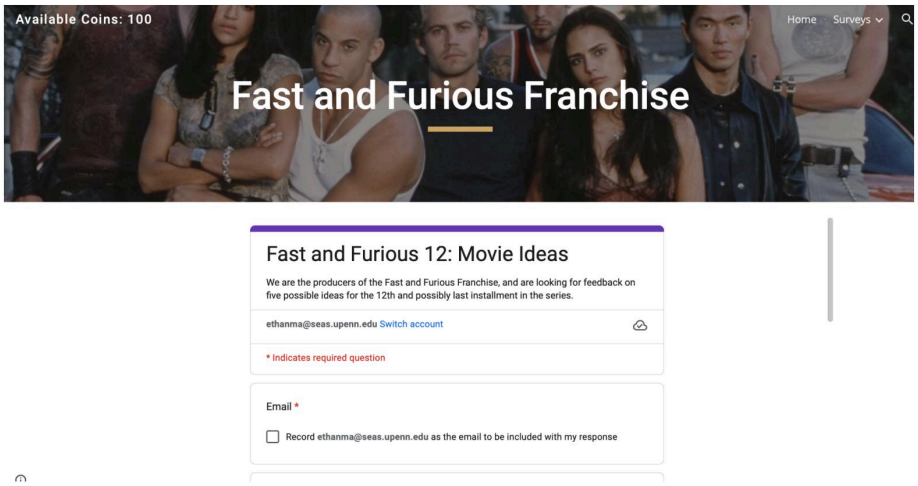
User Interface

Below are some screenshots showing the simple user interface that we created for our crowd workers. Because the primary method for data collection that we utilized was google forms, we really just wanted a single place where we could consolidate the surveys for each of our categories and display simple information like the number of credits available to our users. The surveys themselves would ask users to rate and place bets on four novel ideas along with providing space for users to submit general feedback. Users start from the homepage and can navigate to any of the surveys through the dropdown menu on the top right. Additionally, instructions are clearly displayed on the homepage.

Link to site: https://sites.google.com/seas.upenn.edu/frontline/

Question assigned to the following page: 1





5. Ethics

The ethical considerations of Frontline were carefully evaluated to ensure fairness, transparency, and responsibility. Frontline provides a platform for refining AI-generated ideas, adding value by leveraging human judgment. However, its existence is only justified if it operates ethically and avoids exploitation of participants or unfair evaluation practices. The tasks in this project were designed to be low-risk, involving idea evaluation and token-based betting. However, future expansions might introduce sensitive or harmful

Question assigned to the following page: 1

content to users, necessitating robust content moderation to protect participants. In the current iteration, participants were not compensated beyond the gamification element. A fully implemented version of the platform would ensure fair compensation through the token-reward system, tying participant contributions to potential revenue-sharing opportunities.

Should Frontline expand to create datasets for machine learning, it would be critical to ensure data privacy and compliance with ethical standards. The use of personal or sensitive information must be avoided to respect participant confidentiality. Bias in ML systems is a significant concern, particularly if underrepresented groups are not adequately considered. Such bias could lead to discriminatory outcomes, and care must be taken to ensure that ML algorithms used on the platform do not amplify existing biases. By establishing clear evaluation metrics, mitigating bias through anonymized judging, and incorporating statistical validation, Frontline fosters robust and fair decision-making. Furthermore, regular feedback and reproducibility checks enhance the transparency and accountability of the evaluation process. Ethically, the platform is designed with worker protection, fair compensation, and data privacy in mind, addressing key concerns such as bias and potential harm. However, as Frontline expands, continuous efforts are required to monitor and mitigate risks associated with content moderation and algorithmic fairness.

6. Skills

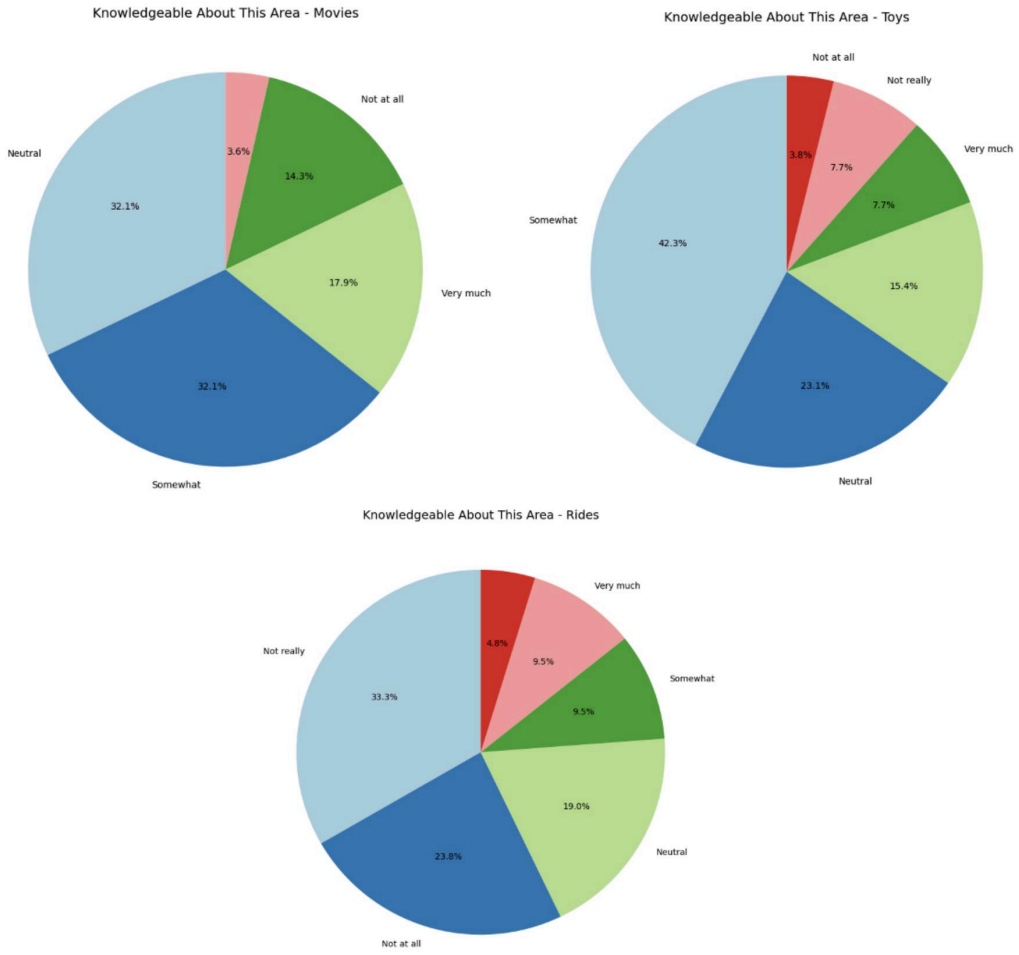
For the Frontline platform, crowd workers do not need highly specialized skills, but they should possess certain abilities to effectively evaluate AI-generated ideas. The necessary skills include critical thinking, the ability to evaluate ideas based on predefined metrics (such as creativity, feasibility, or relevance), and basic knowledge of the platform’s task structure. Workers should also have familiarity with the token-based reward system and an understanding of how to assess and rank ideas fairly.

The skills of individual workers may vary, as some may have more experience in idea evaluation or a better understanding of the criteria, while others may be new to such tasks. Factors such as prior knowledge, familiarity with the platform, and individual biases could cause differences in how well workers perform. Some workers may be better at identifying innovative ideas, while others may excel at evaluating ideas based on feasibility or practicality.

We conducted an analysis of crowd workers' self-reported expertise in the surveyed areas to better understand the distribution of skills within the workers. This analysis was based on workers’ responses regarding their experience and familiarity with the tasks they were asked to evaluate. By examining these self-reported levels of expertise, we aimed to gain insights into how varying levels of expertise might influence the quality and consistency of their evaluations. This allowed us to assess whether more experienced workers performed

Question assigned to the following page: 1

better or if there were discrepancies between self-reported expertise and actual performance on the tasks.



The above pie charts illustrate the ratings that users gave themselves regarding their familiarity with the categories of each of the generated ideas. We can see that the most common familiarity rating varied across each category, with users being similarly familiar with movies and toys but far less familiar with rides. It's possible that familiarity could have a significant impact on the eventual success of the products voted as the most potentially successful by our crowd though without being able to launch our projects, such a hypothesis can't be fully explored.

7. Quality Control

Question assigned to the following page: 1

Since much of the data collected was entered as text inputs by users, its quality and usability as well as formatting cannot be guaranteed, and quality control is necessary. Especially with a limited number of responses, each entry has a non-trivial effect on the ultimate conclusions, analysis, and overall metrics. Therefore, before extracting meaningful results and working with the information we collected, the data needed to be processed to ensure its quality, so that only reliable data would continue on to be used for analysis.

As an overview of the data collection involved in this project, there were 3 categories of crowdsourced data collected, through 3 independent forms filled out by users, corresponding to 3 categories of proposed business ideas: Movies, Toys, and Rides. For each category, participants were presented with 4 ideas for consideration, and asked several sets of questions about each idea.

The primary columns of interest in each category were the numbers of credits participants had bet on each of the ideas. the first step in our quality control module was to parse these responses and remove any non-numeric entries, as well as ignore this section altogether for any users whose credit bets did not sum to at most 100 across the 4 ideas on a form (as the instructions explicitly stated this constraint). This ensured that for all remaining responses, the user had understood the task, and their credit bets were truly meaningful. Additionally, separate questions had asked the users how much money they would pay for the product or experience described by each idea; for these questions, we processed the responses and removed any entries that were left empty or did not specify a number, and extracted the numerical dollar amounts from the rest to be used in aggregation. To confirm that this process was successful, we verified that all responses that passed through the QC module had credit bets that summed to at most 100, and all columns for later quantitative analysis consisted of numerical or enumerable entries.

8. Aggregation

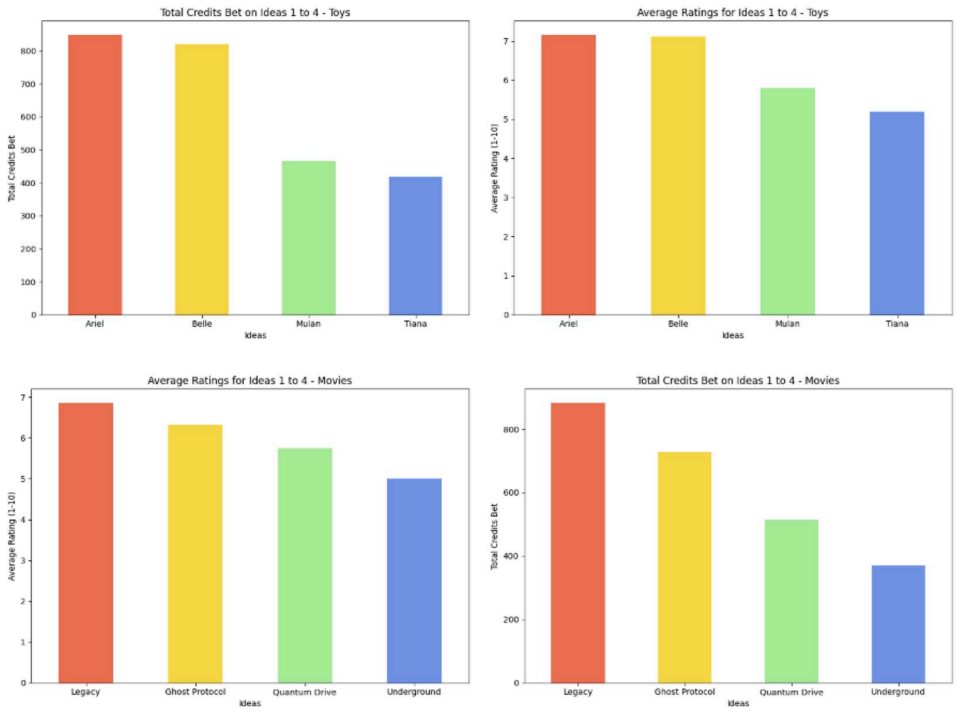
Taking this preprocessed data, the Aggregation module then performed a series of analyses on this reliable subset of data crowdsourced from those who participated on our platform. This section of code served to answer numerous questions of interest regarding the comparison and ranking of ideas within each category, correlations between participants' answers to different questions about the ideas, patterns in the data and distributions of responses with various weights applied according to user behavior, and more.

The primary metrics of interest were measures of the respective popularity and caliber of each idea, focusing on the credit bets as the most reliable expression of a user's opinion. These included a simple sum of the credits received towards each idea, a sum weighted by the user's self-reported level of confidence and knowledge of the relevant area, and further calculations that took into account the user's other answers (e.g. 1-10 rating of the idea and amount of money they would spend on it as a consumer). In each category, the four ideas

Question assigned to the following page: 1

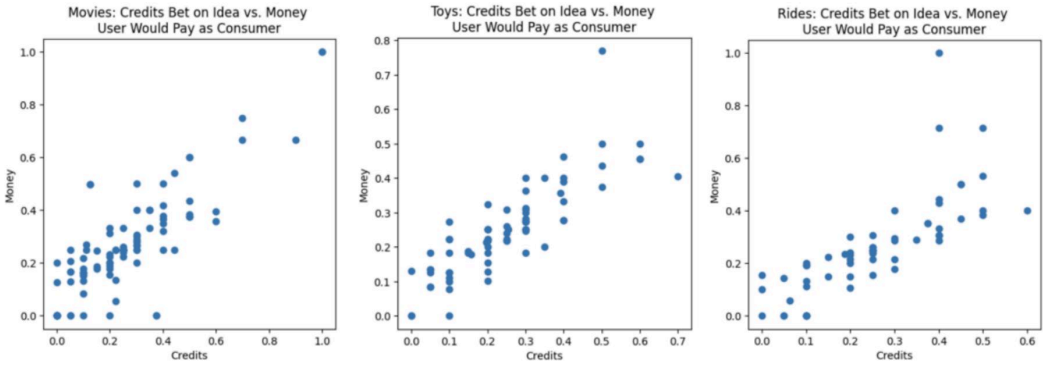
could be narrowed down to one or two winners according to the crowdsourced information, each corroborated by multiple analyses of ratings and credit bets.

We also investigated several questions involving the combination of multiple columns and answer types, and correlations between them. The correlations between participants' answers and consistency of their responses to different questions speak to the quality of the data and efficacy of the QC module. Both within the same participant's submission and evaluation of 4 ideas, and on average over all responses, answers to different questions reflected consistent opinions of the ideas and relative interest in each.

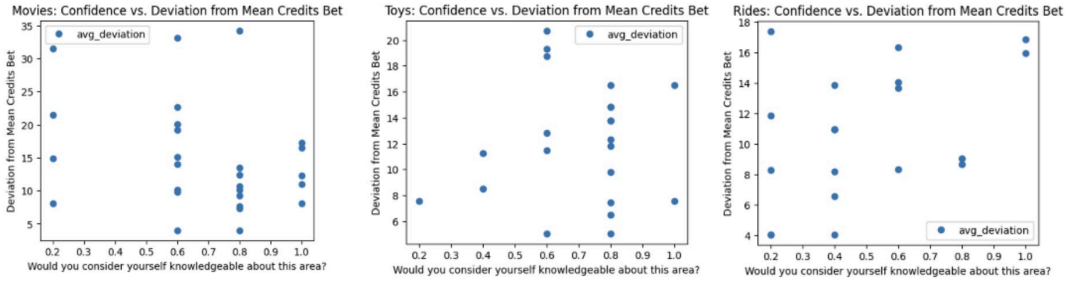


As an example, the bar charts above show a resemblance between aggregated opinions of 4 movie ideas, as expressed through ratings and credit bets. We can observe that while the top two toy ideas were fairly evenly matched, the top scoring movie idea was far in the lead compared to its competitors.

Question assigned to the following page: 1

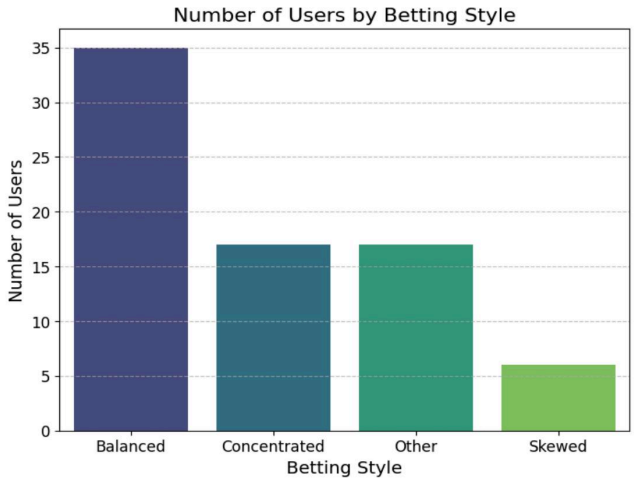


The above row of scatter plots shows the consistent correlation between credits bet and reported dollar amounts that a participant would pay for the product—another interesting question that we sought to answer through our aggregation module.

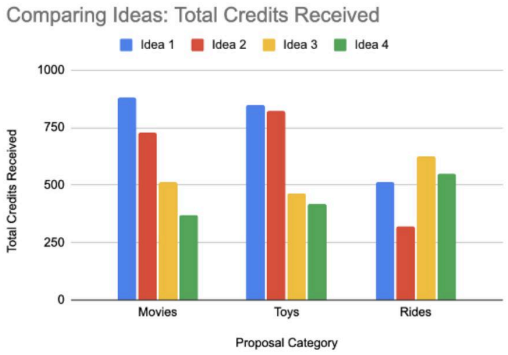


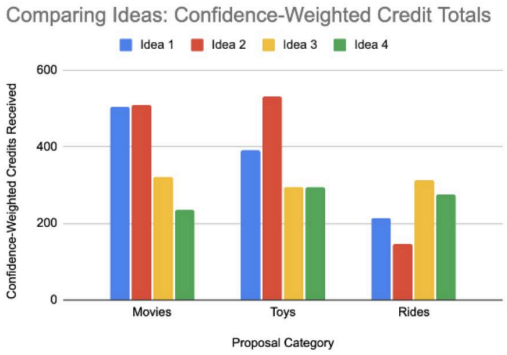
Furthermore, we looked into patterns among participants themselves as evidenced by the scatter plot above. Here, we looked at the correlation between a user's self-reported knowledge of the relevant area for a product category, and the distance from their credit bet to the mean. These experimental results were less clear, but seemed to follow the expected pattern that those familiar with the area distributed their credit bets in a similar manner.

Question assigned to the following page: 1



The above bar chart characterized the betting styles of our users when presented with the opportunity to bet on the four ideas that were generated for each categories. Users were assigned to their respective categories based on the threshold of their balance, concentration, and skewness metrics. Balance was calculated using the standard deviation of user bets, while skewness captured the asymmetry of the distribution of their bets, and concentration evaluated the proportion of their maximum bet when compared to the total bet amount. We can observe here that most of our users placed their bets fairly evenly, resulting in small standard deviations in the amounts bet between each of the four categories.

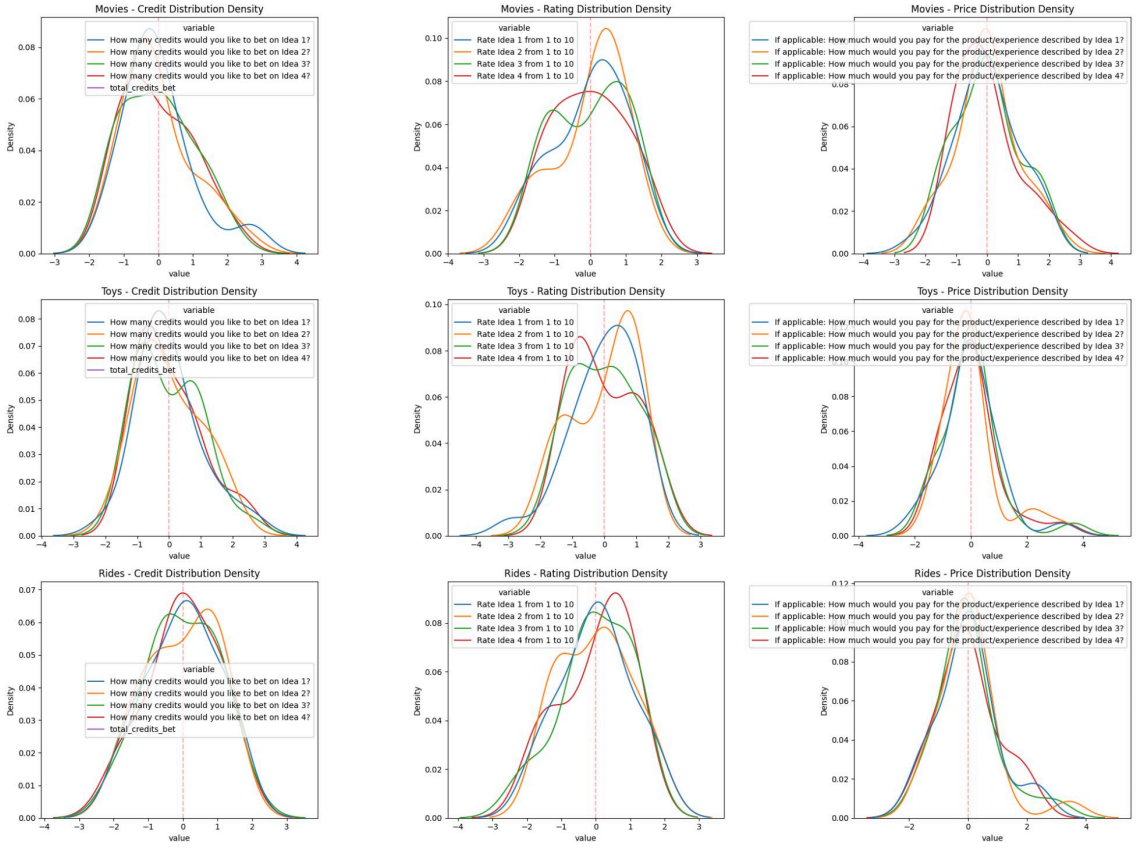




The above grouped bar chart for the total credits each idea received shows clear insights into the popularity of ideas within two categories, Movies and Toys. Movie ideas elicited a clear ranking of preferences, with users spending the most credits to bet on the first proposal. Toys, on the other hand, exhibited a different distribution, with most credits going to the first two ideas equally, while the Rides category showed much more uncertainty in users' observed preferences.

Question assigned to the following page: 1

The confidence-weighted results, on the other hand, display a more meaningful measure of the ideas' relative predicted success. By linearly weighting each user's credit bet according to their self-reported knowledge of the relevant business area, this provided reliable clarity on the more ambiguous results of ranking by total credits. For example, in the Toys category, while total credits favored ideas 1 and 2 equally, the confidence-weighted credit totals showed a preference for idea 2 among those more knowledgeable about this area. Meanwhile, in the Movies category, this analysis also prevented idea 2 from being overlooked, as it was surpassed by idea 1 in total credits, but was more popular among these likely more reliable respondents.



Another pattern we were interested in regarding the user ratings was the distribution (above) of the scores given of different rating scales. The values on the x axis represent what people rated, and are normalized to be standard deviations, which is more useful than plotting raw data. We thought that the three scaling systems would perhaps prompt different thought processes when judging the quality of the provided ideas, and we were correct! We can see that in the three distributions on the left, two of them are positively skewed right (the tail is on the right). This is particularly interesting because it indicates that people are

Question assigned to the following page: 1

betting more of their credits on *one idea*, rather than placing even bets or splitting up their credits between two ideas. An example of this would be something like a set of [20, 20, 20, 40] so the higher values are rarer but skew the graph positively (and make the mean higher than the median as seen on the graphs). We also see that the lines of the three surveys are more or less quite close together.

The classic 1-10 rating distributions are more symmetrical, which is expected as the nature of rating a movie does not require nor inspire allocating resources to the ideas asymmetrically (as is with betting credits). It is interesting to see that while both the 1-10 rating distribution and the price distribution are more or less symmetric, there is a higher discrepancy between the lines in the rating distributions. A possible explanation for this is that there are usually standardized price points for products, for example at $14.99 or $19.99, which most of the ideas fall into.

9. Scaling Up

The problem we are trying to solve is generally specific to a single organization or company. For our purposes, this was our class (so about 30-50 people). For larger organizations, this could be in the 10-thousands scale. Since our solution relies mostly on Google products like Forms and Sites, we could certainly scale to these larger organizations. Though our project certainly would benefit from far more contributions, given that the more people are answering, the more diversity of opinion we can capture, this is also largely contingent on the crowd which we're able to assemble. Our project could perform equally well with a small but well-informed and diverse crowd or a crowd of tens of thousands. However, it is true that some of our analytics rely on larger scales of answers. For example, when trying to see broader betting trends, then it would be very useful to have thousands of people to collect data from.

Furthermore, for the organization using Frontline, thousands of people would allow for more accurate feedback on potential product ideas. Given that one of the questions on our form is general feedback, if thousands of people were to all give their own answers, it would be difficult to go through all of the data. Instead, we would either have to rely more on the points or 1-10 scale, or introduce some other metric for general feedback in order to facilitate scalable transmission of large amounts of unstructured text data.

While investigating the scalability of our project, we performed a brief cost analysis. To this end, we researched the capacity limits of gathering responses through Google Forms, storing data in Google sheets, and the need of compute resources to power any complex analyses. Since our project is built primarily on Google services, which are mostly free for any organization or person with a Google account, it is completely possible to gain thousands of responses in Google forms without paying a single cent.

We found that the limit for Google Forms is 5 million responses, which is also the limit for Google sheets. For even a form with 20 questions, which is more than that of the forms we

Question assigned to the following page: 1

used, it can support 100,000 responses which is more than enough. Additionally, aggregating the data as a spreadsheet in Google sheets does not cost any money at all. The only part of our crowdsourcing pipeline that requires more scalability would be computing units through Google Colab, which would speed up the generation of complex analysis especially with thousands of responses. We believe that 100 compute units on higher GPUs, which costs $9.99 can easily satisfy our compute requirements, and is not even needed if compute time is not an issue.

10. Project Analysis

The inherent difficulty of our project lay in the fact that without actually being able to evaluate the performance of our user predictions by launching actual products, truly evaluating success would prove difficult. However, given the user data that we were able to aggregate, we were able to derive a pretty huge amount of insights from our aggregation process which I think would be sufficient to help inform someone in the case that they actually did want to implement these ideas as actual projects. When we originally proposed our project, we had envisioned that we would mostly be building out a platform for facilitating the collection of user data and the handling of the betting system. However, partway through we realized that without a robust aggregation process, a platform wouldn't be sufficient to actually provide users of the platform with actionable information. As a result, we did a pivot to focusing primarily on the depth of the insights produced by our aggregations.

While the scalability of our project is predicated primarily on the scalability of the Google ecosystem, which we believe to be sufficient for our purposes, the largest limitation of our product is the fact that there currently doesn't exist a process or any structure at all for converting our aggregated results into actual product launches. We're limited to providing users with insight on which product could be most successful, but the rest of that process is largely up to them. Our results did in fact align with some of the insights that were covered during class, particularly regarding both the need for quality control and the likely fact that without sufficient incentive for users to produce work of high quality, ie. inputting bets summing to a valid total, the quality of data collected would suffer such as during the "Be the Crowd" exercise where our classmates had little incentive towards producing data of high quality.

11. Technical Challenges

Before beginning implementation, our team looked for a simple, cost-effective solution to collect and analyze data from different surveys. We chose to use Google Forms, Google Sites, and Google Colab because they offered a free and straightforward way to implement the entire process without the need for expensive hosting or complex infrastructure.

Question assigned to the following page: 1

Creating the surveys in Google Forms was simple, and the responses were automatically stored in Google Sheets, which allowed us to easily access and manage the data. Google Sites provided a simple platform to embed the forms and create a cohesive user experience without any additional costs or technical overhead.

For analysis, we used Google Colab, which provided a free, cloud-based environment for running Python code and performing data analysis with libraries like pandas and matplotlib. The biggest challenge was ensuring smooth integration between the Google Forms, Sheets, and Colab. However, by using APIs like gspread, we were able to pull the form responses into Colab and perform our analysis efficiently. This combination of Google tools made the project both cost-effective and streamlined implementation.