Final Report

● Ungraded

Group

Ahmed Muharram
Eric Zou
Akash Kaukuntla
...and 1 more
View or edit group

Total Points

- / 15 pts

Question 1

Report 15 pts

Question assigned to the following page: 1

Milestone 5 - Final Report

Basic project information

Name of your project

DataLabeler

Name of your teammates

Ahmed Muharram, Akash Kaukuntla, Eric Zou, David Zhan

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

DataLabeler aims to provide dense captions for a variety of images using crowdsourcing as a source of training data for multimodal LLMs and other machine learning models.

Logo for your project.

Question assigned to the following page: 1

What problem does it solve?

It provides dense image captions which is good for machine learning models on images or multimodal LLMs like Molmo.

What similar projects exist?

Molmo probably has a similar application that they used to gather training data to make Molmo. There are also various sources of dense image captions online but these are sparse.

What type of project is it?

This is a crowdsourcing project that accomplishes its goal through paying workers for short tasks. On a technical level it works as a full stack web application that workers access after accepting a task on mechanical turk.

What was the main focus of your team's effort

Our main focus was to create a comprehensive application that workers would have an easy time with. Making as little roadblocks as possible made it more likely that we would get great data.

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.

First workers will accept a task on Amazon Mechanical Turk. This task will give them a link to our application. From there, they must read instructions on how to complete the task, they just record a minute of audio describing the shown picture. From here, we will automatically transcribe the audio and do quality control on the transcription. If the quality control fails, the worker will have to submit their audio again with better quality. Finally, when they submit they will be given a unique confirmation code which they will put back on Turk. If this code is valid, our system will automatically accept or reject the HIT, providing them with payment.

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://www.youtube.com/watch?v=LLBAH43Sk9o
Enable Captions!

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

Project Analysis and Aggregation (Ethics section also went very in-depth)

Question assigned to the following page: 1

The Crowd

Who are the members of your crowd?

In practice the members of our crowd will be random Amazon Mechanical Turk workers. We do not put restrictions on who can participate so it generally would be the general public.

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

For our final project we used a combination of friends and family to collect data as well as classmates.

If the crowd was simulated, how did you collect this set of data?

Classmates, friends and family filled out these forms for us.

If the crowd was simulated, how would you change things to use a real crowd?

People weren't trying to pass our quality control in order to get paid so this wasn't an issue in our simulation but in reality if we are paying people they will probably try to game the system. In practice, we may want more strict quality control methods but this could in turn deter some workers if their audio doesn't pass on the first try.

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

We just asked people to participate.

How many unique participants did you have?

We had around 35-40 unique participants.

Question assigned to the following page: 1

Incentives

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

The main motivation the crowd has to participate in our project is the pay. It is hard to consistently motivate new workers to participate in our project without pay because it is a repetitive task that doesn't induce learning or some other reward easily. Additionally, a small subset of participants would have an incentive to participate just because they want to see machine learning as a whole succeed and simply want to contribute to training data.

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?

To incentivize the crowd to participate, we rely primarily on monetary compensation. Given the repetitive and task-oriented nature of the work, financial rewards are the most effective way to consistently attract and retain participants. By offering competitive pay, we ensure that workers feel their time and effort are valued, which keeps them motivated and engaged. Without this financial incentive, it would be challenging to maintain steady participation over time, as the tasks don't naturally offer learning opportunities or personal growth.

Beyond monetary rewards, we recognize that a small subset of participants may be intrinsically motivated by a desire to contribute to the advancement of machine learning. These individuals might participate simply because they believe in the larger goal of improving AI and training data. While this group is valuable, relying solely on altruism is not scalable or practical for maintaining a large, consistent workforce.

When transitioning to a real crowd in addition to offering pay, we could explore other motivators, such as gamification elements to make tasks more engaging, recognition systems to reward top contributors, or even opportunities for participants to learn new skills. These additions could help us appeal to a broader audience while maintaining productivity and engagement.

Did you perform any analysis comparing different incentives?

We did not perform much analysis with regard to differing incentives because it was clear to us that for this project the main motivator should be monetary rewards.

If you compared different incentives, what analysis did you perform? If you have a graph analyzing incentives, include a graph here.

We did not.

Question assigned to the following page: 1

What the crowd gives you

What does the crowd provide for you?

The crowd is the core of our app. They provide the information that makes our app useful. Essentially, they are the ones that provide dense captions for the images that we aggregate to create useful captions.

Is this something that could be automated?

The point of our app is to provide valuable training data for multimodal LLMs and ML models so that one day they will be able to perform the same tasks that these workers are performing on a high level.

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

It is difficult to automate this task because the most advanced machine learning models in the world still cannot caption images on the level that humans can. This is because object recognition is not something that comes naturally to machines. In the future, however, we can see how this would be possible to automate.

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

We did not.

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

N/A

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?

N/A

If you have a graph analyzing a machine learning component, include the graph here.

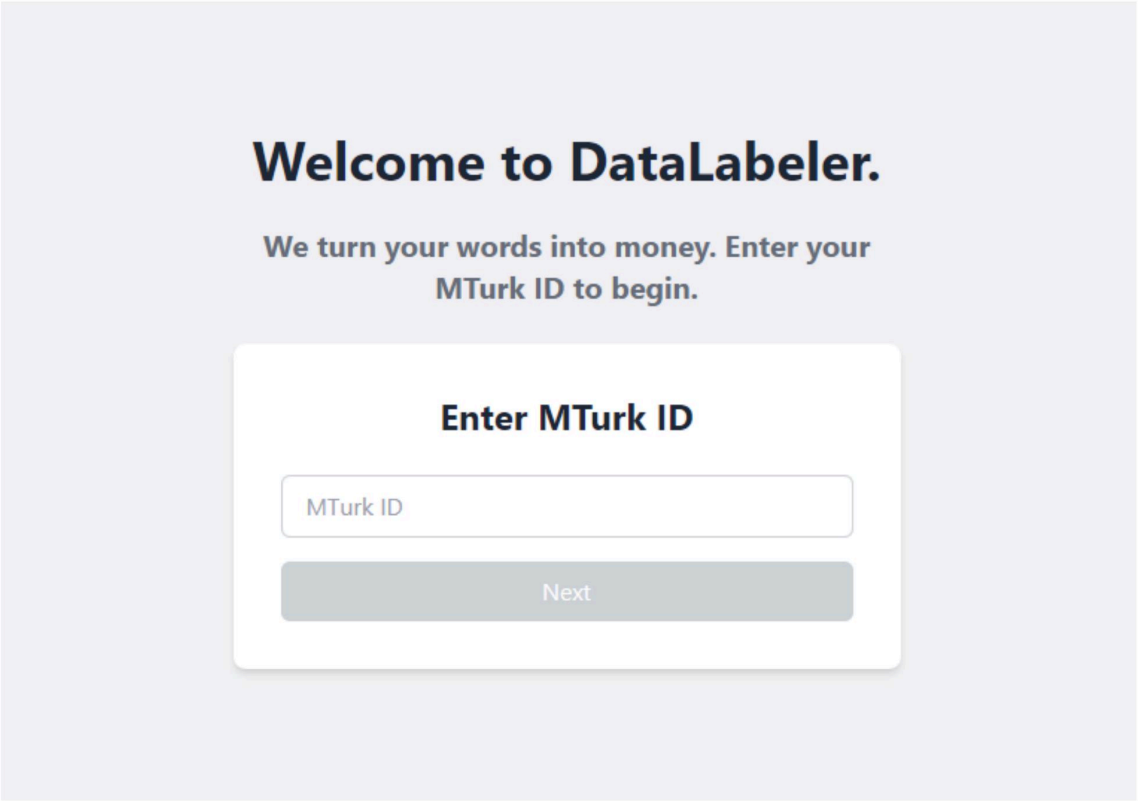
N/A

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

We created a whole frontend for the users to record their image captions and submit these captions. We also created a survey for the HIT that gives users access to our app.

Question assigned to the following page: 1

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

Welcome to the Task

Thank you for participating! In the next step, you will be presented with an image. Your task is to carefully observe the image and then describe what you see. Please speak freely, but ensure your recording lasts at least **60 seconds**.

At the end of the task, provided your recording passes our **quality control standards**, you will receive a confirmation code that you can use to claim your payment on MTurk.

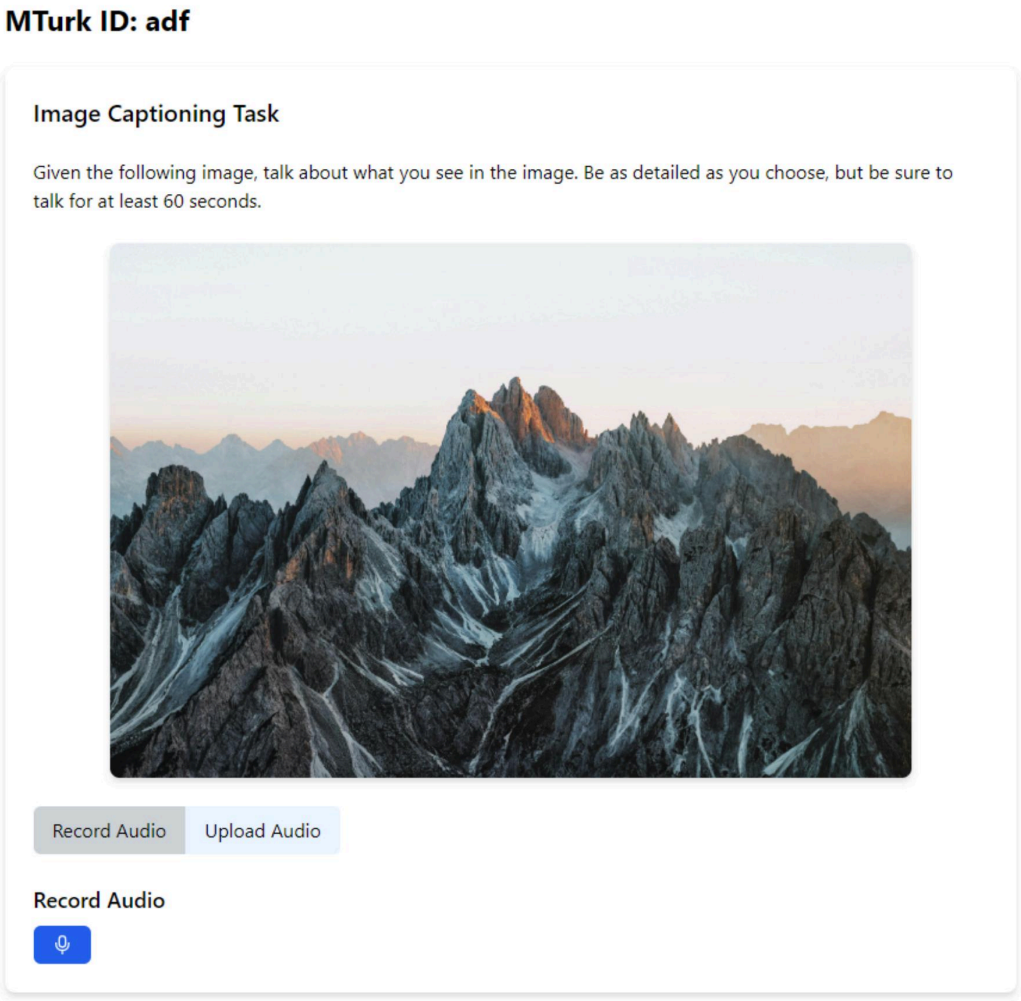
Important Guidelines:

- Do not muffle your microphone or stand too far away.
- Ensure your audio is clear and relevant to the task.
- Recording irrelevant audio or low-quality submissions may violate the terms and conditions and result in payment revocation.

By clicking 'Start Task', you are agreeing to our terms and conditions.

Start Task

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 focus was to provide a clean design for the app so as to not distract from the image much. We also wanted everything to be as simple and straightforward as possible to provide a smooth experience for the workers in order to get higher quality data.

Question assigned to the following page: 1

Ethics

Should my application exist at all?

We believe this application plays an essential role in advancing the future of large language models (LLMs) and machine learning as a whole. By contributing to the development of high-quality training data, this application helps improve the accuracy, relevance, and scalability of ML systems. As the demand for smarter and more adaptable AI systems grows, tools like ours can make a significant impact by addressing gaps in data quality and accessibility. While its existence is justified from a technological standpoint, we recognize the responsibility to ensure it operates ethically and with minimal unintended consequences. This means continuously evaluating its purpose, scope, and methods to align with broader societal needs and ethical guidelines.

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

There is potential for harm if harmful or inappropriate images are introduced into the image rotation. Exposure to such content could cause emotional distress or even trauma for workers tasked with labeling or moderating the data. However, we take proactive measures to reduce this risk by carefully moderating the database and ensuring that harmful images are excluded before deployment. While these safeguards minimize the likelihood of harm, it is important to acknowledge that no system is entirely foolproof. Should harmful content inadvertently slip through, it could have a lasting psychological impact on workers, emphasizing the need for ongoing vigilance, comprehensive moderation protocols, and the availability of support resources for workers who may encounter distressing material.

Are you fairly compensating the workers for their time?

Yes, we strive to fairly compensate workers for their time and efforts. Compared to many similar tasks on platforms like Mechanical Turk, our compensation rates are more competitive and considerate of the time spent. Workers are paid approximately 15 cents per minute, which equates to about $9 per hour—above the federal minimum wage in the United States. We recognize that fair pay is essential for maintaining worker morale, ensuring consistent participation, and fostering trust. However, we also acknowledge that compensation standards can vary globally, and we aim to regularly review our pay structure to ensure it remains fair and equitable across different regions. Beyond pay, we could also explore offering additional incentives such as performance bonuses or recognition to further support our workers.

Question assigned to the following page: 1

If you are creating a dataset for machine learning:

> Are you allowed to access the data that you're labeling (have you ensured that it isn't private data)?

We have ensured that all the data we use is open source and publicly available. Specifically, the images in our dataset are sourced from Unsplash, a platform that provides royalty-free, high-quality photographs. If the application were to be deployed, we would properly attribute the images to their creators as per the platform's guidelines. By exclusively using open-source data, we minimize the risk of infringing on privacy rights or accessing restricted information. That said, we are mindful of the ethical considerations involved in using publicly available data and aim to ensure that it is utilized responsibly and in line with both legal and ethical standards.

> Is there the potential for introducing bias in your machine learning classifier (unintended discrimination against a group because of underrepresentation)?

Yes, there is a potential for introducing bias in the classifier, particularly because our dataset and transcription service currently focus exclusively on English-speaking countries and English audio. This creates a risk of overrepresentation of Western ideologies and perspectives while underrepresenting other cultures and languages. Such bias could lead to models that are less inclusive and less effective for users from non-English-speaking backgrounds. To address this issue, expanding our transcription services to support a broader range of languages would be a critical improvement. Additionally, ensuring that the training image dataset includes diverse cultural, linguistic, and geographic representations would help mitigate unintended discrimination and make the application more globally applicable.

> ML algorithms sometimes amplify bias in the training data. What would be the consequences if this happens in your ML system?

If the ML system amplifies bias present in the training data, it could lead to unfair and exclusionary outcomes. For example, overrepresentation of Western perspectives might result in a model that struggles to understand or accurately respond to inputs from underrepresented cultures or languages. This could alienate users, reduce trust in the system, and perpetuate harmful stereotypes. In certain cases, such biases could have serious real-world consequences, particularly in applications where fairness and inclusivity are critical, such as hiring, education, or content moderation. Additionally, a biased model is less adaptable and scalable, as it may fail to perform well in diverse scenarios. To address this, it is vital to actively identify and mitigate bias during data collection and training, as well as to audit the system regularly for unintended consequences. By doing so, we can create a more robust, equitable, and widely applicable system.

Question assigned to the following page: 1

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

Our evaluation is thoughtfully conducted and considers key factors such as worker well-being, data ethics, and potential bias. However, there is room for improvement to ensure it stands up to rigorous scientific scrutiny. While we address potential risks and propose mitigation strategies, the evaluation could benefit from deeper analysis and more visual aids to support our conclusions. Additionally, if this were to be published or deployed on a larger scale, we would prioritize implementing measures such as peer review, formal bias audits, and comprehensive validation testing. These steps would strengthen the reliability and fairness of our system, ensuring that our conclusions are not only sound but also scientifically robust. Regular iteration and open collaboration with the research community would further enhance the credibility of our work.

Question assigned to the following page: 1

Skills

Do your crowd workers need specialized skills?

No, our crowd workers do not need any specialized skills or training.

What sort of skills do they need?

Our crowd workers need to be able to understand English, speak English with some level of proficiency, and recognize common objects in pictures.

Do the skills of individual workers vary widely?

Generally, the skills of individual workers do not vary much, since the task itself is a very simple one.

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

Generally, the skills do not vary widely, but someone who is hyper-observant and descriptive in their language will be a lot more useful than someone who is not.

Did you analyze the skills of the crowd?

We did not analyze the skills of the crowd.

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 did not analyze the skills of the crowd. We looked at the quality of their results, by comparing the word lengths and similarity scores of their singular responses compared with LLM-generated responses and aggregations. We found that aggregated and

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

We did not analyze the skills of the crowd.

Question assigned to the following page: 1

Quality Control

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

For our limited experiments, our crowd mainly consisted of classmates, family and friends. The quality of the crowd did not give us any concerns. But if we were to scale up and use strangers as a crowd, the concern would be that the strangers would potentially try to game the task and provide descriptions of the pictures that were too short or simply irrelevant.

How do you ensure the quality of the work the crowd provides?
The current QC involves making sure that the transcribed response of the worker has at least 2 out of 10 key-words mentioned with regards to the picture (which were generated by LLMs beforehand) and that the response is at least 80 words long. This is a standard speaking speed and ensures that we receive enough relevant information about the picture.

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).

Our system includes several design features that help reduce the likelihood of poor-quality contributions at the outset. These features aim to promote quality through clear guidance, task framing, and intrinsic incentives for good performance.

First, we provide clear, concise, and specific instructions to workers. Each worker is required to speak for at least one minute about the image, describing it in as much detail as possible. We also emphasize that the description should focus on key objects, relationships, and context within the image. This instruction alone helps minimize irrelevant content, as it gives workers a concrete goal to achieve.

Second, we require a minimum one-minute response ensures that workers engage meaningfully with the task. Short, careless responses are automatically disqualified since they fail to meet the 80-word threshold. This design forces contributors to provide detailed, thoughtful descriptions and encourages active participation.

Additionally, before the crowd begins working on an image, we use a Large Language Model (LLM) to pre-generate a list of 10 essential "key-words" relevant to the image. These key-words capture important objects, activities, and contextual features of the image. By requiring workers to reference at least 2 out of these 10 keywords in their responses, we ensure that their descriptions are contextually relevant and avoid generic, low-effort inputs.

Finally, rather than relying on a single crowd-sourced caption, we collect three separate responses per image. This aggregation strategy is important for two reasons. First, it provides redundancy — if one caption is low-quality or incomplete, the other two can compensate. Second, it allows us to merge or synthesize the responses later, either manually or using LLMs,

Question assigned to the following page: 1

to create a more complete and comprehensive description of the image, which ensures that the result fits the overall quality standards.

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?

We did not compare different quality control strategies, since we only had one.
When analyzing the quality of what we got back, we compared the word lengths of the individual worker responses, the aggregated responses, and sample LLM generated captions. We found that on average, the individual worker responses and aggregated responses had around the same length, but both were significantly longer than the LLM generated captions. However, this increase in word count did not correlate positively with quality. It makes sense, as word count should not be the sole indicator of quality.

What analysis did you perform on quality?

The analysis performed on quality involved word count comparisons of worker responses, LLM responses, and aggregated responses. We also analyzed each of these three responses in terms of similarity score, which involved seeing how many of the 10 keywords were referenced in each type of response.

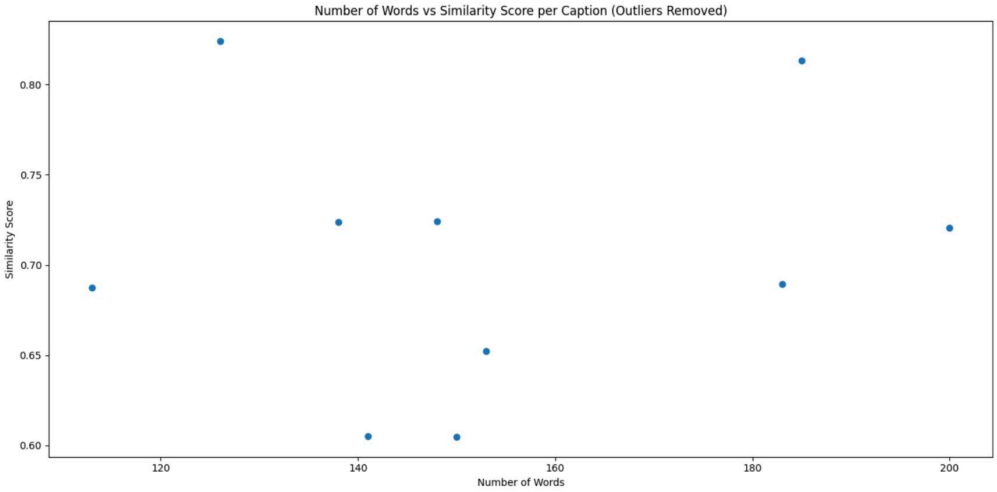
What questions did you investigate? What conclusions did you reach?

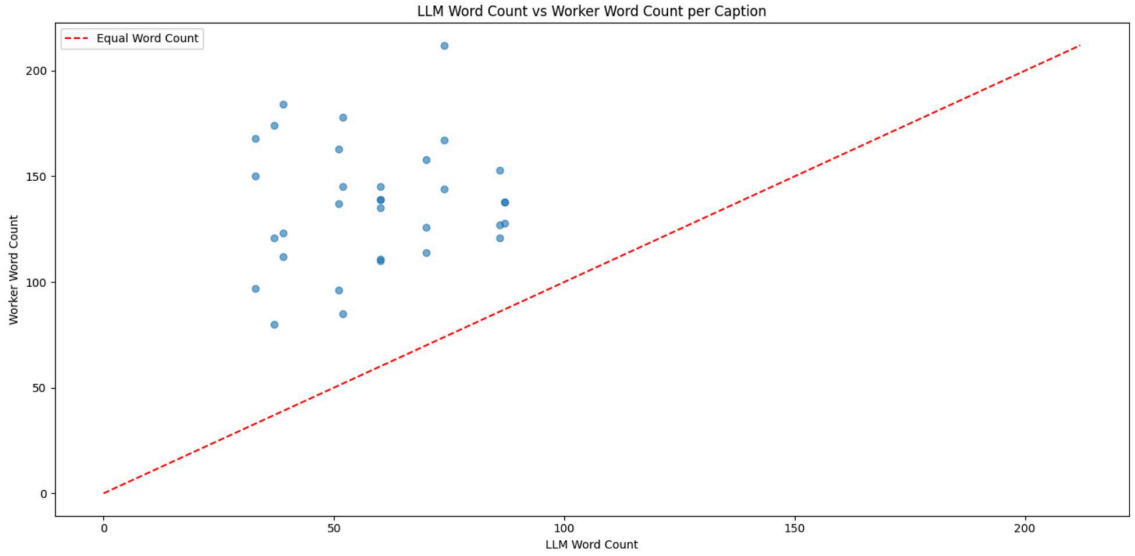
We investigated “How to make sure workers provide quality captions?” and “How does the quality of captions vary between workers, aggregated, and LLMs?”.
The conclusions we reached were that having a robust quality control metric as well as aggregation significantly improved the final caption quality, as it made sure we had long enough responses and it combined multiple perspectives and filled in gaps left by individual worker responses. We also came to the conclusion that aggregated captions consistently had higher quality and word counts compared to individual captions. This demonstrated that combining multiple worker perspectives created a more detailed and dense description of the image, which demonstrated the power of crowdsourcing.

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

These following graphs are also available in the Aggregation section following this.

Question assigned to the following page: 1


Number of Words vs Similarity Score per Caption (Outliers Removed)
Similarity Score
0.80
0.75
0.70
0.65
0.60
120
140
160
180
200
Number of Words


LLM Word Count vs Worker Word Count per Caption
Equal Word Count
Worker Word Count
200
150
100
50
0
0
50
100
150
200
LLM Word Count

Question assigned to the following page: 1

Aggregation

How do you aggregate the results from the crowd?

After quality control was applied with response-length and keyword filtering, we aggregated crowd worker inputs in groups of 3 for each image, with each image being assigned the aggregated caption for the first 3 accepted captions. No further responses are collected for this image beyond the 3 initial inputs. We then use Llama-3.1b to combine the captions with the intent of keeping the distinct information from each and corroborating each source with the others. We performed some prompt tuning for the task and settled on the following instructions for the LLM:

```
"""Your job is to combine the following {n} image captions into 1 unified
description capturing all of the information in each.
    Please do your best to keep as many of the details as possible while
maintaining consistency of the scene.
    Remove phrases and words that do not make sense in the context
provided by the responses. Please do not report anything else. Only return
the description.
    {joined}
    Please provide the resulting description in your following message:"""
```

In this prompt, n is replaced with the number of captions (3 generally, but it could be adapted for more or less). The “joined” variable is replaced with captions separated by newlines and altered by adding the text “Caption: ” in front. We prompt the model as the user, with no other system prompts or chat history. The maximum number of tokens in the output is set to 8192, though this does not seem to be reached by any of the captions.

This model was hosted on Cerebras, and they provide an API and a Python SDK to make requests. Our aggregation implementation was a single function that accepted the list of captions to aggregate, so we could call it anywhere.

Did you analyze the aggregated results?

Yes. See Project Analysis and the next question.

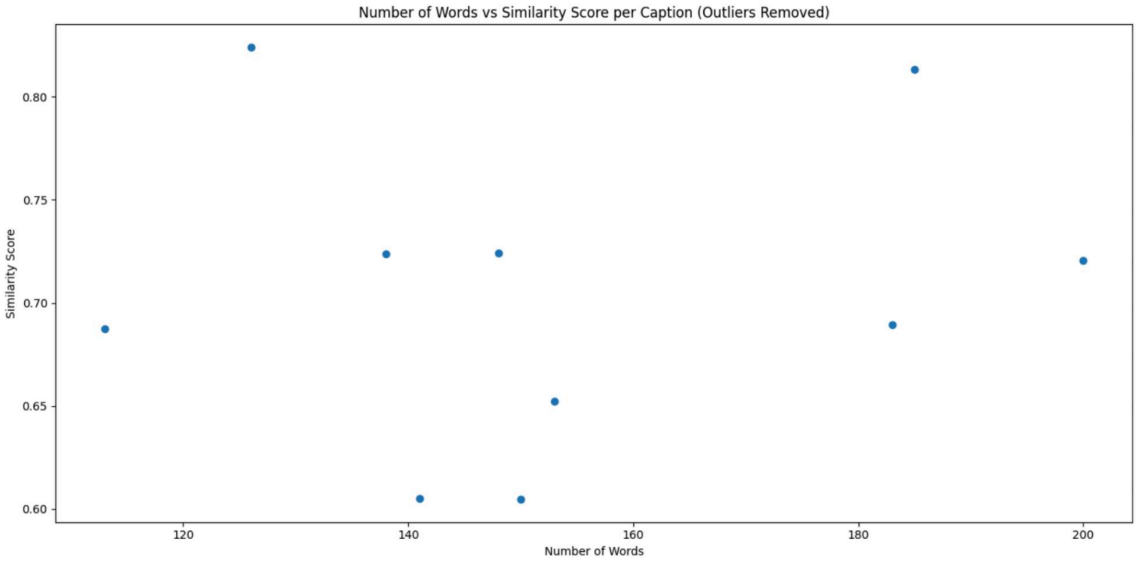
What analysis did you perform on the aggregated results? What questions did you investigate? Did you compare aggregated responses against individual responses? What conclusions did you reach?

We focused on comparing the performance of individual workers, aggregated worker responses, and the LLM-generated captions across key metrics such as word count and quality. The primary objective was to investigate whether aggregation improves the quality of captions and how these aggregated results relate to the contributions of individual workers and the LLM.

Question assigned to the following page: 1

Initially, we saw that the aggregated captions consistently outperformed individual worker captions in quality, as measured by textual similarity to a reference LLM caption (first graph in the next section). This suggests that combining multiple perspectives through aggregation enriches the coherence and relevance of the captions as we saw multiple times throughout this report (with more in the Project Analysis section).

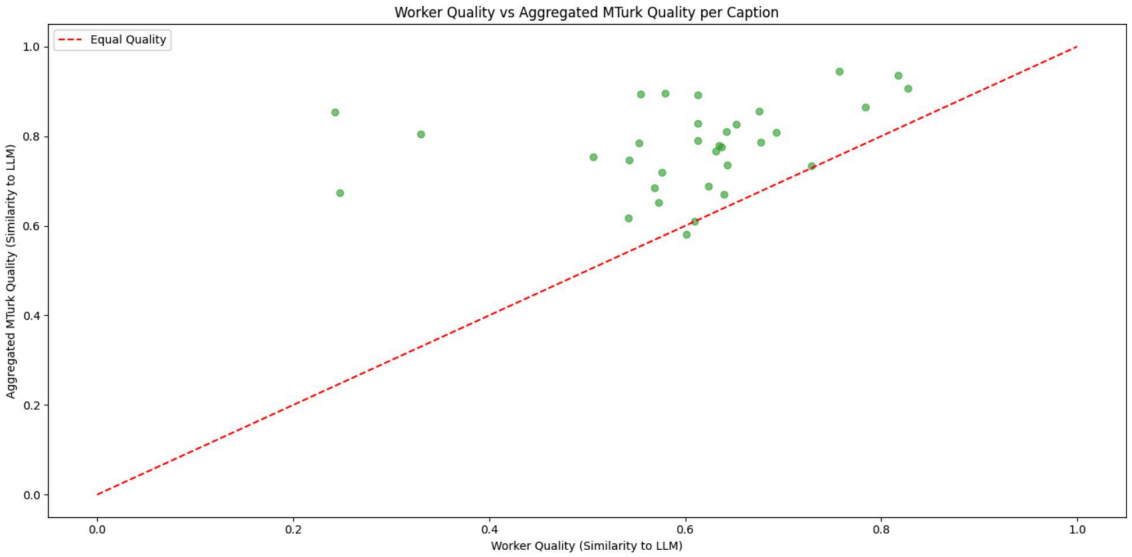
Individual workers tended to produce captions with a higher word count compared to the LLM outputs (second graph in the next section), but in aggregation both individual workers and their aggregated captions had around the same overall word count (third graph in the next section). However, this increase in word count did not correlate positively with quality. This reinforces our conclusion that we made on the last milestone: word count alone is not a reliable indicator of caption quality.

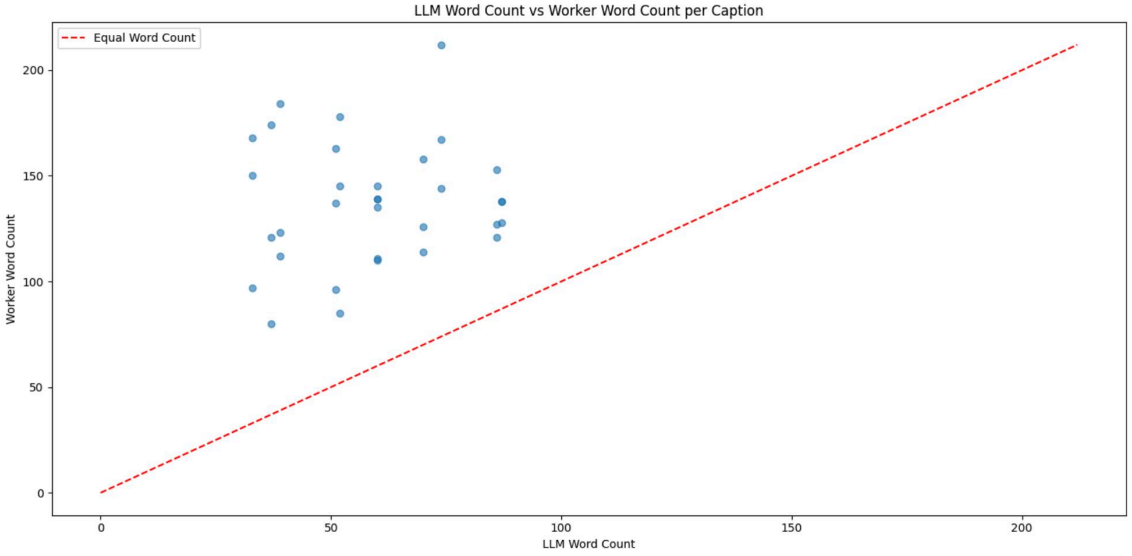


Comparing these findings against the LLM, it became evident that while the LLM captions were concise and not as dense: they lacked the contextual richness and adaptability shown by the aggregated worker responses.

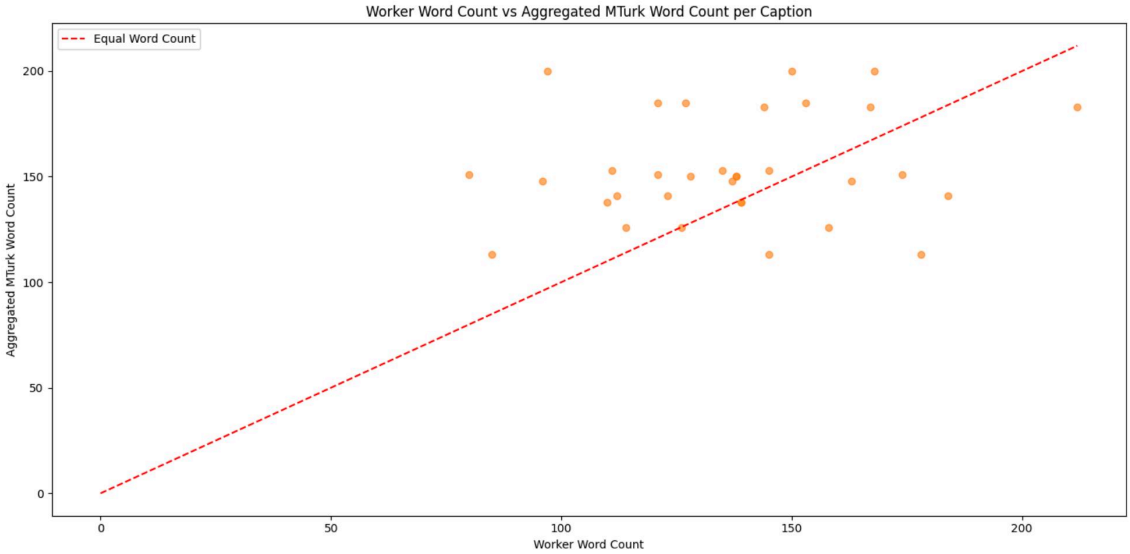
Question assigned to the following page: 1

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.

No, our aggregated results will be saved for other use and not displayed to the user.

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.

N/A

Question assigned to the following page: 1

Scaling Up

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

The problem of creating dense image captions for training multimodal LLMs and other machine learning models is vast. These captions require high levels of detail and contextual understanding, which are not easily achieved through automation at present. The demand for such data spans industries like computer vision, autonomous vehicles, healthcare imaging, and AI-driven content generation. Hence, the scale is enormous, as it involves creating millions of high-quality, contextually relevant captions.

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

Yes, the project would greatly benefit from contributions by thousands of people.

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

Increasing the number of contributions would allow us to aggregate over more diverse perspectives, increase redundancy, generate more data, and in general speed up data collection. All of these factors would benefit our project and increase the quality and quantity of dense captions that we are able to produce. Additionally, having a large diverse dataset is essential for training useful models.

What challenges would scaling to a large crowd introduce?

The first would be quality control. Ensuring that captions from a large, diverse crowd meet quality standards becomes more complex. Workers may try to game the system for payment, providing low-effort or irrelevant descriptions. Second, larger crowds require robust backend systems to handle increased traffic, data processing, storage, and transcription needs. Scaling up also increases costs related to paying workers, hosting infrastructure, and using APIs for keyword generation and transcription.Finally, if the crowd is not diverse enough, scaling up could inadvertently amplify biases already present in the data or in the keyword generation process.

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

We did not perform an analysis about scaling up our project. We did have a brief conversation with Professor Lumbroso about the costs and business model of creating a larger scale project. He talked to us about finding the cost per person/caption for operating our system and that we would charge this amount if we would to create a service using our product.kashkau

What analysis did you perform on the scaling up?

N/A

Question assigned to the following page: 1

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

We did not consider scaling our project.

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

We did not consider scaling our project.

Question assigned to the following page: 1

Project Analysis

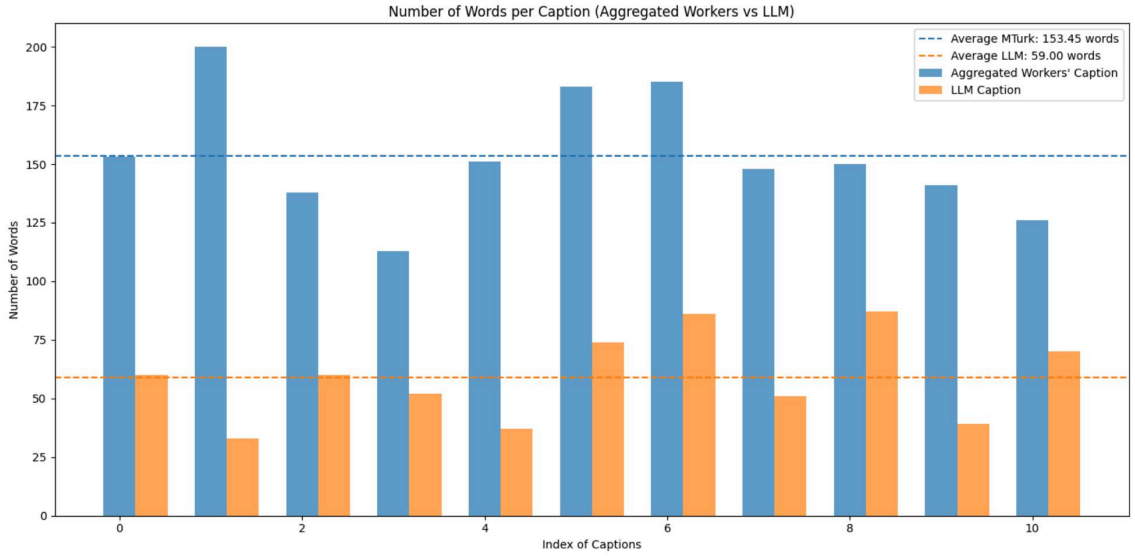
Did your project work? How do you know? Analyze some results, discuss some positive outcomes of your project.

Our project worked even better than we expected. While looking at the aggregated transcriptions, we noticed that they contained details that we didn't initially see in our first look. Here's an example, which highlights some obvious and not-so-obvious details the workers mentioned:

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

This graph shows how our model compares to a standard LLM caption given the image and a small prompt. Our prompt to the LLM was "`Create a dense caption for this image. Use only concrete nouns (i.e. building, tree) not abstract nouns (i.e. love, care)`" to ensure fairness.

Question assigned to the following page: 1



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

- Collecting valid data after quality control has been applied: a lot of our samples did not fit the minimum requirements (captions were less than 80 words or did not meet the keyword criteria).
- Deploying Whisper. We had to switch to another service (lemonfox.ai) since it fit our RAM requirements and deployed on Render.
- Getting the MTurk API to work. The documentation for this was sparse and finding and setting up environment keys was difficult at first. Additionally, the API was all new to us making it a steep learning curve overall.

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

No.

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

No major changes were made. However, we would like to continue collecting data at some point and continue working on the project to scale it further!

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.

One limitation of the project is the challenge of maintaining QC, especially when transitioning from a simulated crowd to real-world workers. In the simulation, workers did not attempt to game

Question assigned to the following page: 1

the system, as their motivation was not tied to financial incentives (i.e. they were students needing the grade, so they had to make meaningful contributions). However, in practice, workers who are paid per task may try to meet quality thresholds in the quickest way possible, often at the expense of genuine effort. This could result in submissions that superficially pass the criteria but fail to provide meaningful training data. Furthermore, we suspect our data will be of less quality once we actually give it to workers instead of students, given that our sample was mostly Penn students who are more likely to generate higher quality data.

Scalability is another significant challenge. As the volume of data grows, so do the costs of compensating workers, processing submissions, storing data, paying for APIs that generate keywords and transcribe audio. Managing this growth while maintaining profitability requires balancing fair incentives for workers with efficient use of resources. Furthermore, the infrastructure needed to handle large-scale data collection and aggregation must be robust, and we certainly will need to change some technical components in our architecture (switch from using MEGA for storing pictures). which can become a bottleneck without significant investment in scalable cloud solutions and optimized pipelines like EC2 and S3.

Bias is also an inherent risk in the method. Relying on LLM-generated keywords as part of the quality control process could introduce biases inherent in the LLM itself, reinforcing patterns that already exist rather than challenging or expanding them. Another limitation lies in the aggregation of captions into dense descriptions. While combining captions for multiple images can result in richer data, it also risks losing nuance or context, especially if the images are not closely related. This simplification will eventually diminish the usefulness of the captions for training other multimodal models.

From an engineering perspective, scaling this product would require addressing these challenges through enhanced quality control mechanisms, optimized cost structures, and efficient infrastructure. For example, implementing multi-layered validation processes that combine automated checks with human oversight could balance fairness with stringency. Providing feedback loops for workers could help improve submissions while maintaining their engagement. On the technical side, adopting scalable cloud infrastructure and streamlining workflows could handle the increased data load effectively. In terms of sources of error, several issues arise from the methodology itself. The simulated crowd may not accurately reflect real-world worker behaviors, which would lead to unrealistic expectations when transitioning to live operations. Similarly, reliance on pre-existing LLM keywords for validation might bias the dataset, excluding descriptions that are valid but less common (also, generating data to train an LLM with an LLM defeats the entire purpose!) Variability in audio quality (due to different microphones, background noise, or accents) could also impact transcription accuracy and the quality of aggregated captions.

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

It deviated but in a good way! We expected results to be ~90 words average per aggregated transcription, but we got 153.45 words on average.

Question assigned to the following page: 1

If your results deviated, why might that be?

It might be dependent on the prompt we use to aggregate the results and having very high quality data (post-QC).

Question assigned to the following page: 1

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?

Our project required relatively intense technical implementation.

Our frontend was a single-page application built on JavaScript and React. We used Vite for our development environment. For styling, we used some component libraries and Tailwind CSS. We needed to handle recorded and uploaded audio on the frontend side, as well as MTurk id input and displaying the confirmation code.

Our backend was centered around a Python server using FastAPI and Uvicorn. This handled our application logic and routing of requests. We also used asyncio to schedule a background task to poll the MTurk service for completed tasks so that we could approve or reject them, as well as a self-ping to prevent the hosting server from spinning down due to inactivity.

The client first makes a request to the backend for a random image from one of the images without captions in the Supabase Postgres table (the first 5 images by UUID were selected from to encourage complete captions). The image was retrieved from the image store (Mega.io) and sent to the user. The backend then accepted audio and Mturk ID from the user and made requests to the Lemonfox automatic speech recognition (ASR) API to transcribe user speech. Then, it would pass these through quality control, making requests to GPT-4o if needed to generate keywords for an image and updating these in the database. If the caption passed, the backend would add the entry in the database, performing aggregation on the captions if 3 captions were created for that image. For aggregation, we would prompt the Cerebras Llama-3.1b API client and receive a response to persist in the database. On successful user submission, a confirmation code would be generated and the user would submit this to MTurk. Our backend would then get this information from MTurk and accept the task. The API also allows for admin work such as acknowledging and removing HITs to poll for submissions, adding new images, and querying aggregated captions.

We created a task-generation script that would generate a certain number of tasks in MTurk and would update the app via an endpoint we created so the app would monitor those tasks.

The APIs we used in this project were all relatively new to us, especially the MTurk client (Akash is the GOAT) and Mega client. Learning how to use them, debugging, and ultimately integrating them with the application took some time. Additionally, managing the asynchronous nature of the application, especially with web requests, httpx, and MTurk made this project an interesting challenge

To deploy the application, we used Vercel for the frontend and Render for the backend, both with free tier hosting options. We ran into trouble earlier due to the Render free tier not offering

Question assigned to the following page: 1

enough memory to host a small OpenAI Whisper model instance, so this is why we switched to a third-party ASR provider.

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

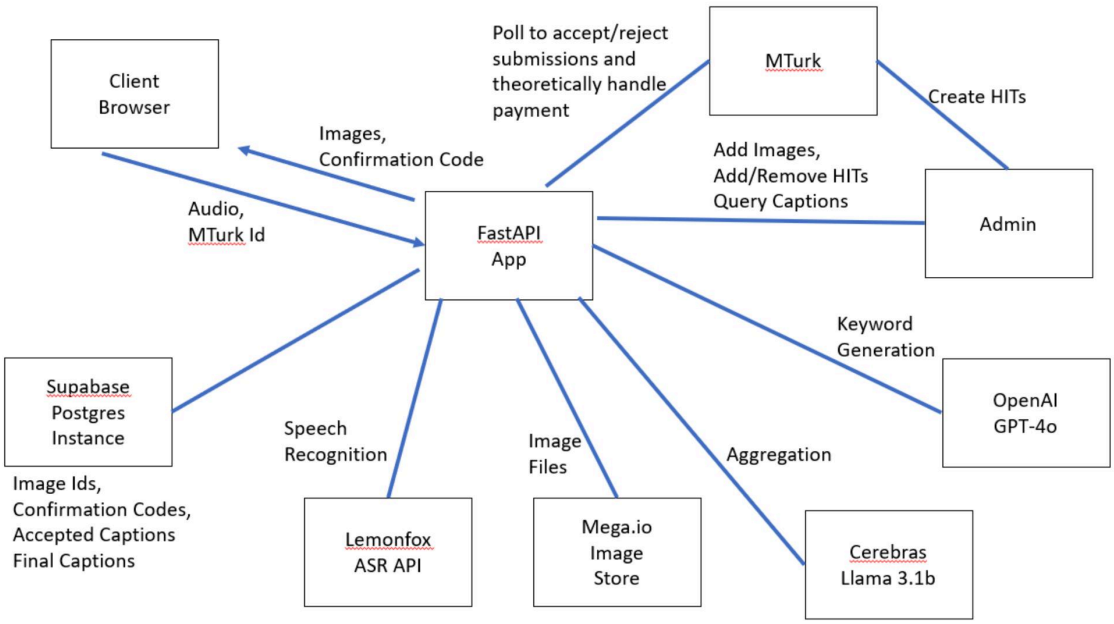
We had a lot of difficulty handling the image and audio file component of the application due to the new language features and APIs used. In the beginning, we were having trouble correctly sending and receiving the files in all parts of our application.

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.

In the frontend, storing the audio and image files required handling object URLs for blobs stored in the browser. We also had to understand and work with the MediaRecorder web API to allow the user to record audio on our site.

In the backend, we needed to handle file requests and responses, generating temporary files and handling image types correctly. We convert all images to JPEG before being added to the image store and their filename is their image code in the database. Reading the images asynchronously and returning them to the client from the store was hard to get working. The audio is in the webm format. We turn the user response into a temporary file before sending it to the ASR API.

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



Question assigned to the following page: 1

Other info

Is there anything else you'd like to say about your project?

If the project were to scale, we would hope to eventually collect enough data to make this a tool that provides data to open-source models like Molmo and contribute to development of projects of that kind.

If you have additional information about your project that didn't fit into the above questions, put it here.

N/A