Project Proposal

<Aseel Meshal AlKhadaidi>

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	Helping doctors to identify children's x-ray images weather they have pneumonia or not. This provides more development to the health industry. Using ML helps the doctors to identify the disease fast so they can get to treatment, the sooner the better.
Choice of Data Labels What labels did you decide to add to your data? And why did you decide on these labels vs any other option?	I have developed three response options for annotators: Yes, No, and Not Sure. Annotators are instructed to select an option based on the relevance of the query to the provided X-ray image. 1. **Yes:** If the query is "Healthy" and the X-ray image is clear, with no detected cloudiness or opacity, the appropriate response is "Yes," indicating that the image corresponds to the query of healthy X-ray images. Conversely, if the query is "Pneumonia" and the X-ray image exhibits cloudiness or opacity, the response should also be "Yes," as the image is relevant to the query. 2. **No:** If the query and the X-ray image do not match—for example, if the query is "Healthy" but the X-ray image shows signs of pneumonia—the appropriate response is "No." 3. **Not Sure:** To address potential uncertainties or difficulties in annotation, a "Not Sure" option has been included. This option should be selected when the annotator is uncertain about the relevance of the image to the query.

Test Questions & Quality Assurance

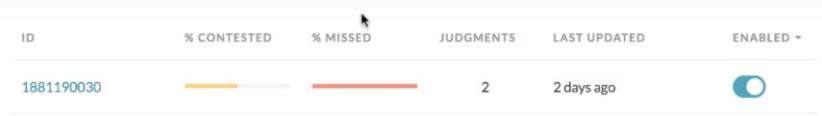
Number of Test Questions

Considering the size of this dataset, how many test questions did you develop to prepare for launching a data annotation job?

I decided to make it **14 test questions**. Made the answer distribution even by making the answers **50% yes** and **50% no** of multiple healthy and Pneumonia cases.

Improving a Test Question

Given the following test question which almost 100% of annotators missed, statistics, what steps might you take to improve or redesign this question?



1. Updating the instructions and providing more complex examples for the annotators to keep an eye on.
2. Updating the test question reason for the labeling to get the annotator more understanding.
3. Maybe adding more specific labels or updating the question to make it more clear.

Contributor Satisfaction

Say you've run a test launch and gotten back results from your annotators; the instructions and test questions are rated below 3.5, what areas of your Instruction document would you try to improve (Examples, Test Questions, etc.)

1. As I see here the instructions weren't clear enough for the annotators, so I'll change them to make it more clear and understandable, add more complex examples to make them understand more what cases have a pneumonia and what not.
2. Also since the test question is also rated badly I'll change it to make it more clear and understandable. Probably also remove ambiguous questions.
3. And finally improving the ease of job, like how the annotators access this job how they annotate it.

Limitations & Improvements

Data Source Consider the size and source of your data; what biases are built into the data and how might the data be improved?	The dataset is relatively small and partially labeled with binary classifications (Yes/No). To address potential biases, we could expand the dataset and ensure a well-balanced representation for each label: Yes, No, and Not Sure. By including a more diverse range of examples and ensuring that each option is adequately represented, we can mitigate biases that may arise from over-representation or under-representation of any particular label. This approach will improve the dataset's overall balance and help in achieving more accurate and reliable annotations.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	• **Balance Labels:** Actively manage and balance the frequency of each label to avoid over-representation of any category. • **Update Guidelines:** Regularly review and refine annotation guidelines to ensure consistency and clarity. • **Quality Control:** Implement checks for accuracy, such as cross-reviewing annotations or using multiple annotators.