Project Proposal

Nouv B. Al-Qahtani

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	Pneumonia is a serious lung infection that can be seen in chest X-rays. **The goal** is to build AI models that can quickly and accurately diagnose pneumonia by analyzing these X-ray images. The models will be trained on datasets of healthy and pneumonia X-rays. This will teach the AI to recognize the patterns and features of pneumonia. This automated approach can be more consistent and efficient than manual review by doctors. The aim is to create AI tools that can help doctors quickly detect pneumonia. This can lead to earlier treatment and better outcomes, especially for high-risk groups like children and the elderly. Machine learning is well-suited for this medical imaging task because it can generalize and adapt to new data.
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?	The labels are: 1. **Healthy** - Normal X-rays with no pneumonia signs 2. **Pneumonia** - X-rays clearly showing pneumonia 3. **Uncertain** - Ambiguous cases, not clearly healthy or pneumonia These labels let the annotators consistently categorize the X-ray images during data preparation. The "uncertain" label is important to capture cases that are unclear, rather than forcing a decision.

Test Questions & Quality Assurance

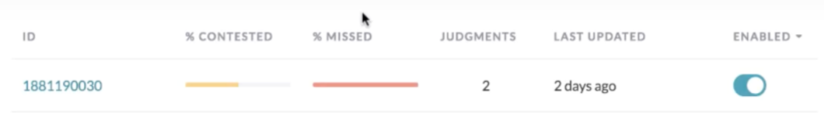
Number of Test Questions

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

I created **8 test questions**, with 2-3 questions for each of the 3 labels - Healthy, Pneumonia, and Uncertain.
This balanced set of test questions helps ensure the human annotators are assessed equally on their ability to identify the different types of X-ray findings, avoiding bias towards any particular label.

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?

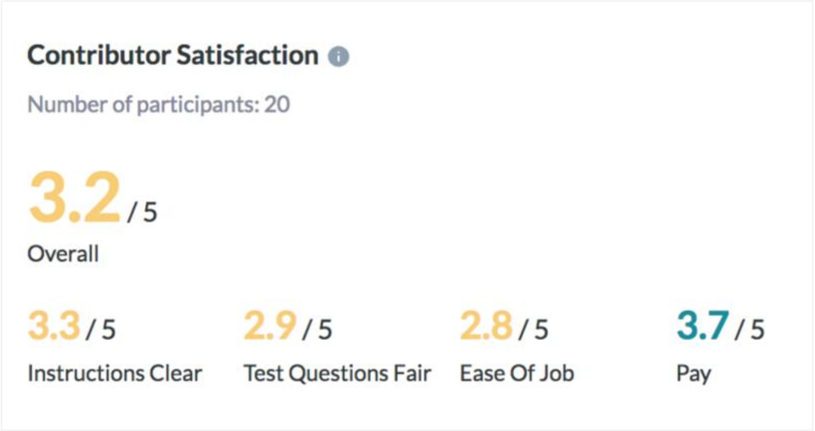


ID	% CONTESTED	% MISSED	JUDGMENTS	LAST UPDATED	ENABLED ▾
1881190030			2	2 days ago	

- I will make sure the question is unambiguous and uses plain language.
- I will avoid technical terms that may confuse the annotators.
- I will ensure the answer choices are reasonable and distinct.
- I can remove any misleading or confusing options.
- I will provide some background information to help annotators understand the topic.

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.)



Contributor Satisfaction

Number of participants: 20

3.2 / 5
Overall

3.3 / 5	2.9 / 5	2.8 / 5	3.7 / 5
Instructions Clear	Test Questions Fair	Ease Of Job	Pay

The main problems are the test questions and instructions are not clear enough.
For the test questions:

- I will look at them and make them better.
- I can get feedback from the people taking the tests to understand the issues.

For the instructions:

- I shall find parts that need to be simpler and more straightforward.
- I will add more examples and explanations.
- I can organize the instructions better.

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 data source is a small CSV file with 117 x-ray images. The small size of the data set may limit its representativeness. Since the data came from an unknown source, there could be biases in the types of patients or medical conditions included. Additionally, the brightness of some images may be misleading and introduce further biases. **To improve the data**, we should look to expand the size of the data set and collect more metadata about the images, such as patient demographics, medical history, and details about the imaging process. This would help identify and address any biases in the data, including those introduced by variations in image brightness. We could also try data augmentation techniques to artificially increase the diversity of the training data..
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	**To improve the longevity** of my solution, I'll focus on a few key areas. First, I'll regularly review and update my test questions as I encounter new data and edge cases. Second, I'll keep the rules and tips current by revising them to reflect changes in the data and problem space. Finally, I'll expand my dataset with more x-ray images, paying close attention to instances of pneumonia. This will help me develop a more accurate and adaptable model over time.