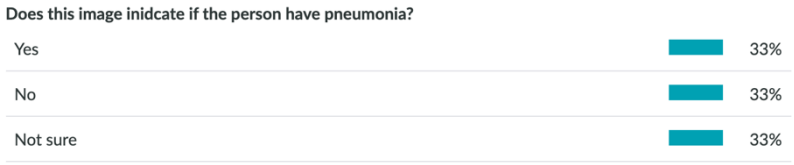
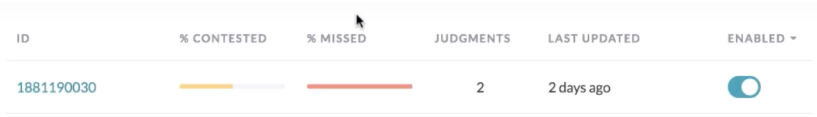
Project Proposal

Nourah Alhassan

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	Help medical professionals in identifying possible cases for pneumonia disease in x-ray images of patient lungs faster and eliminate cases that do not have any apparent symptoms. Using ML will help doctors to consider their decisions more carefully if the model show different results than what they assumed.
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 decided to use three labels “Yes”, “No”, and “Not Sure”. “Yes” & “No” is to determine if the x-rays image is of pneumonia symptoms. “Not sure” is there to leave room for uncertainty because it’s a sensitive case. I chose those three labels because they’re simpler than say “Healthy lungs” , “Pneumonia symptoms”, and “Unknown”.

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 choose to do 18 questions
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- Remove any ambiguities that night cause confusion. 2- Explain the steps more carefully 3- Add tips to help annotators identify the correct answers 4- Make sure the rules are clear and easy to understand
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- Provide more detailed description 2- Give more examples for each table

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?	Improvement: 1- Have a larger set of data large enough for a ML model to learn patterns. 2- We need the ML to deal with all possible scenarios 3- There's no bias as of now only 24 labels are given, 8 for each label. So in the case that we have some bias we will need to throw away data from the class that have more data 4- The Data source could use some improvement such as: different lightening conditions, illumination, angels and other considerations 5- Keep the data regularly updates
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	Consider more situations and incident, change labels accordingly if needed.