

You need to convert images to text (base64 encoded images)

The trained model will work best when there are at most 100 times more images for the most common label than for the least common label. => Remove very low occuring labels

Note: Perfect is the enemy of good. Perfect or very-high-average-precision scores could indicate that something is wrong in the model. The data is too easy and not varied enough.