Evaluation of Bicycle Detection by Crowd Annotators: An Analysis of Performance and Quality.

Problem statement:

We were given a dataset to evaluate the quality and performance of a new crowd for a specific task - identifying bicycles in street images.

Dataset:

The dataset consists of two files - the anonymized annotator responses and the reference dataset. The annotator responses include task input, task output, and metadata about the vendor and annotator. The reference dataset is used as a basis to evaluate the performance of the annotators.

Tasks:

We performed the following tasks on the dataset:

1 Gathered insights about the annotators

2 Analyzed annotator responses for corrupt data and can't solveresponses

3 Evaluated the balance of the reference set

4 Identified good and bad annotators using the reference set

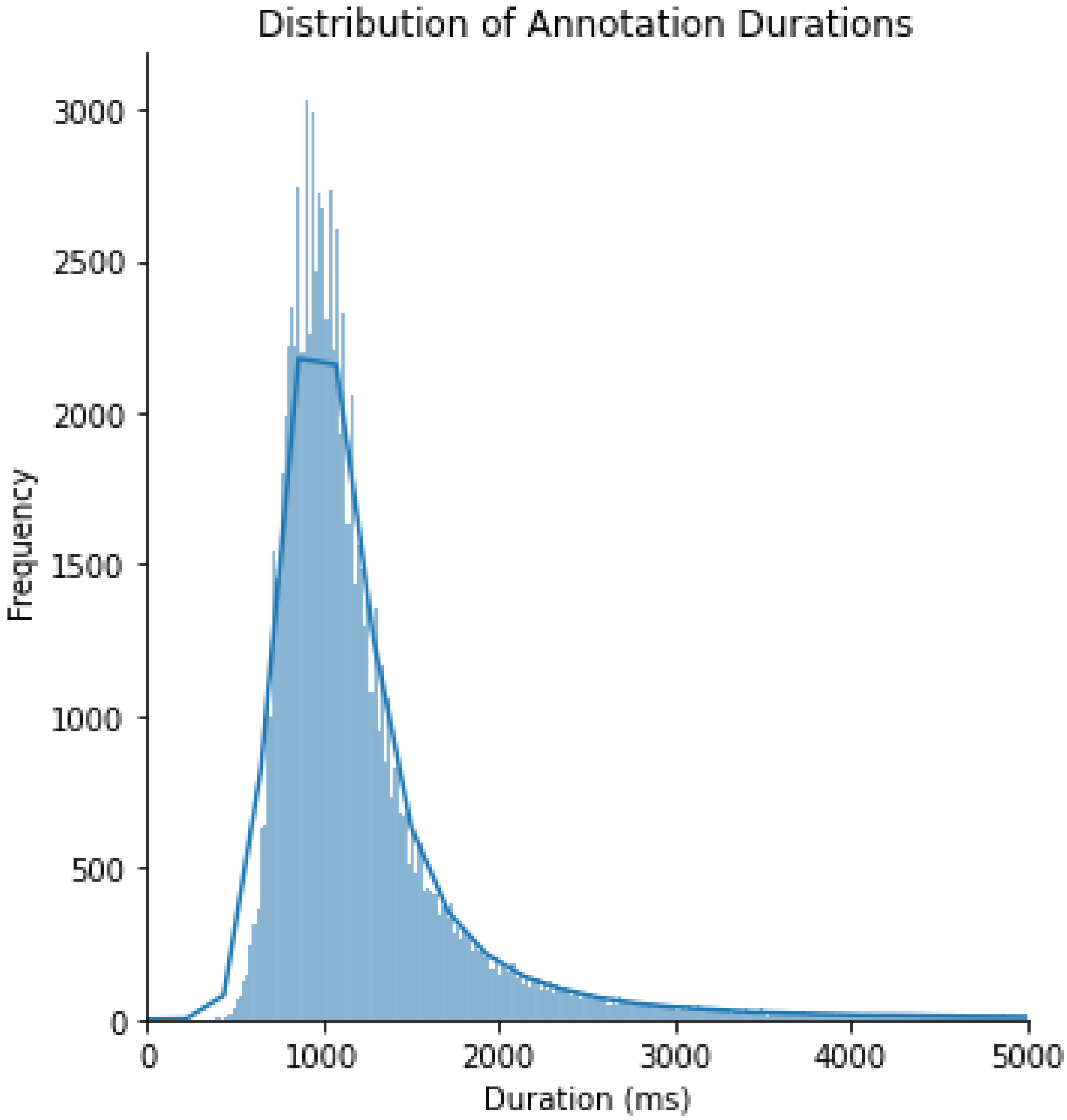
5 Prepared a presentation to summarize our findings

Findings:

Our analysis revealed the following key findings:

1 A total of 22 annotators contributed to the dataset.

2 The average, min and max annotation times were 1289.92, 10 and 42398 ms respectively.
The below chart shows the distribution of annotation durations :



3 There were differences in the number of results produced by the annotators.

In the following table we see the number of results produced by each annotator:

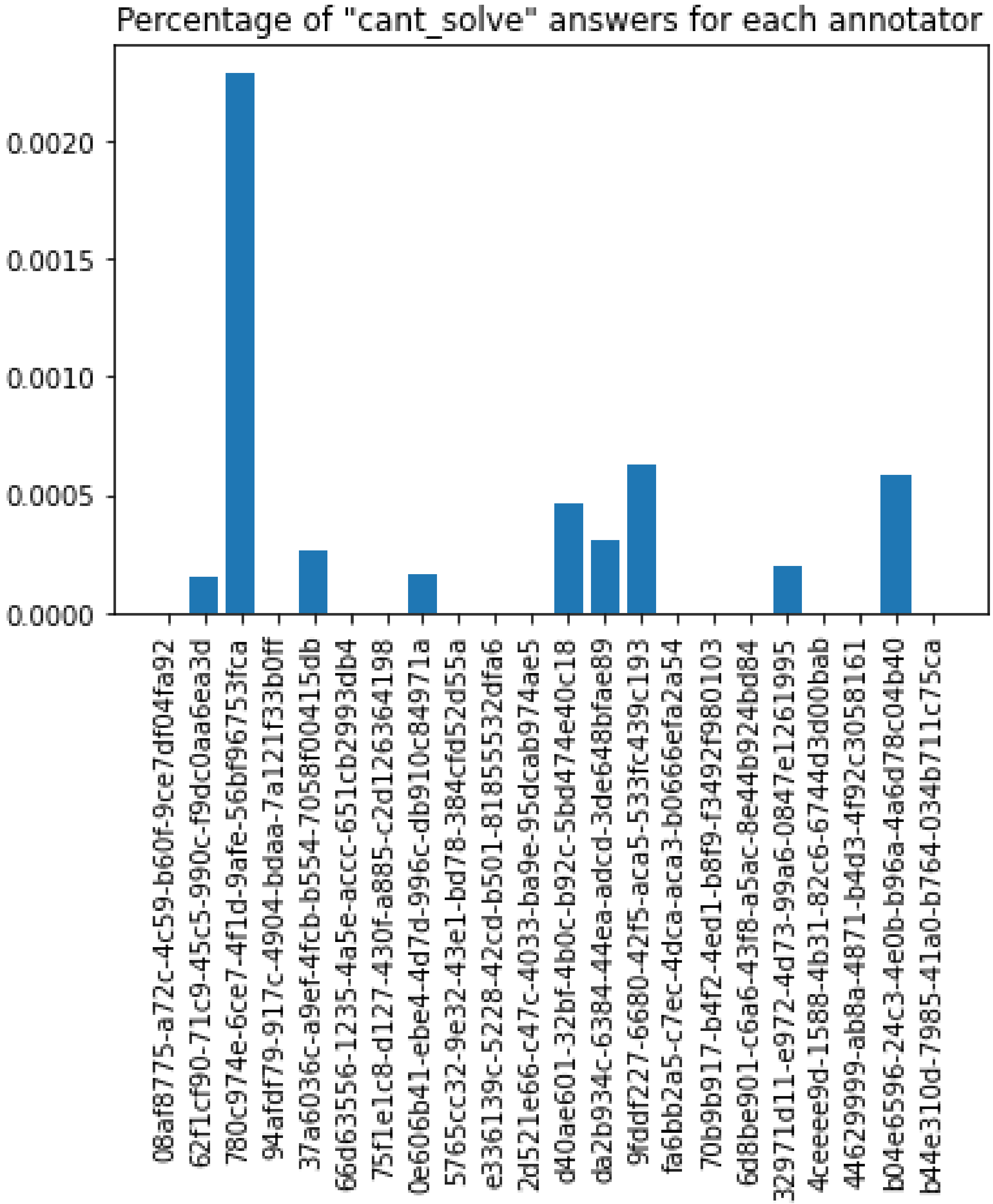
Annotator ID	Number of results
08af8775-a72c-4c59-b60f-9ce7df04fa92	6210
0e606b41-ebe4-4d7d-996c-db910c84971a	6126
2d521e66-c47c-4033-ba9e-95dcab974ae5	630
32971d11-e972-4d73-99a6-0847e1261995	5170
37a6036c-a9ef-4fcb-b554-7058f00415db	7596
44629999-ab8a-4871-b4d3-4f92c3058161	1280
4ceeee9d-1588-4b31-82c6-6744d3d00bab	315
5765cc32-9e32-43e1-bd78-384cfd52d55a	5337
62f1cf90-71c9-45c5-990c-f9dc0aa6ea3d	6436
66d63556-1235-4a5e-accc-651cb2993db4	5061
6d8be901-c6a6-43f8-a5ac-8e44b924bd84	4860
70b9b917-b4f2-4ed1-b8f9-f3492f980103	2950
75f1e1c8-d127-430f-a885-c2d126364198	6088
780c974e-6ce7-4f1d-9afe-56bf96753fca	1745
94afdf79-917c-4904-bdaa-7a121f33b0ff	3485
9fddf227-6680-42f5-aca5-533fc439c193	6421
b04e6596-24c3-4e0b-b96a-4a6d78c04b40	1725
b44e310d-7985-41a0-b764-034b711c75ca	170
d40ae601-32bf-4b0c-b92c-5bd474e40c18	2175
da2b934c-6384-44ea-adcd-3de648bfae89	6537
e336139c-5228-42cd-b501-81855532dfa6	3475
fa6bb2a5-c7ec-4dca-aca3-b0666efa2a54	7078

4 Annotators highly disagreed on some questions, as shown by the distribution of the number of matches (i.e., the number of annotators who agreed on the answer) for the questions in the dataset.

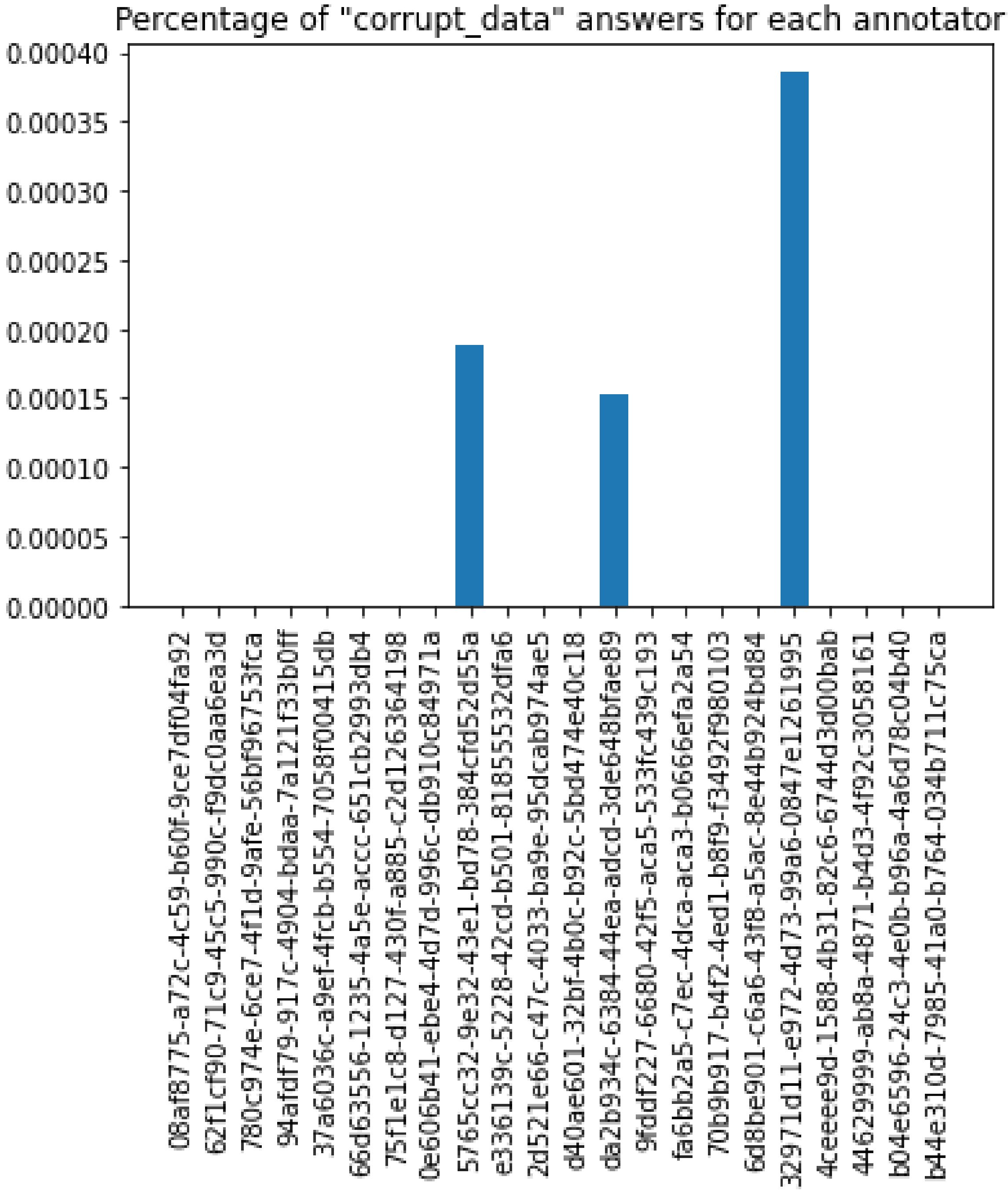
The average number of matches per question is 11.91, with a standard deviation of 9.99, indicating a high degree of variability in the number of matches across questions. While some questions had all 22 annotators agreeing on the answer, others had as few as two annotators agreeing. This highlights the variability in the annotation process and the need for careful quality control measures.

5 Corrupt data and can't solveresponses occurred 4 and 17 times respectively

To further investigate the behavior of annotators, we calculated the percentages of times can't solveänd corrupt data"were used for each annotator. The below charts show the percentage of can't solveör corrupt dataöptions, which represent the probability that a particular annotator used either the can't solveör corrupt dataöptions.



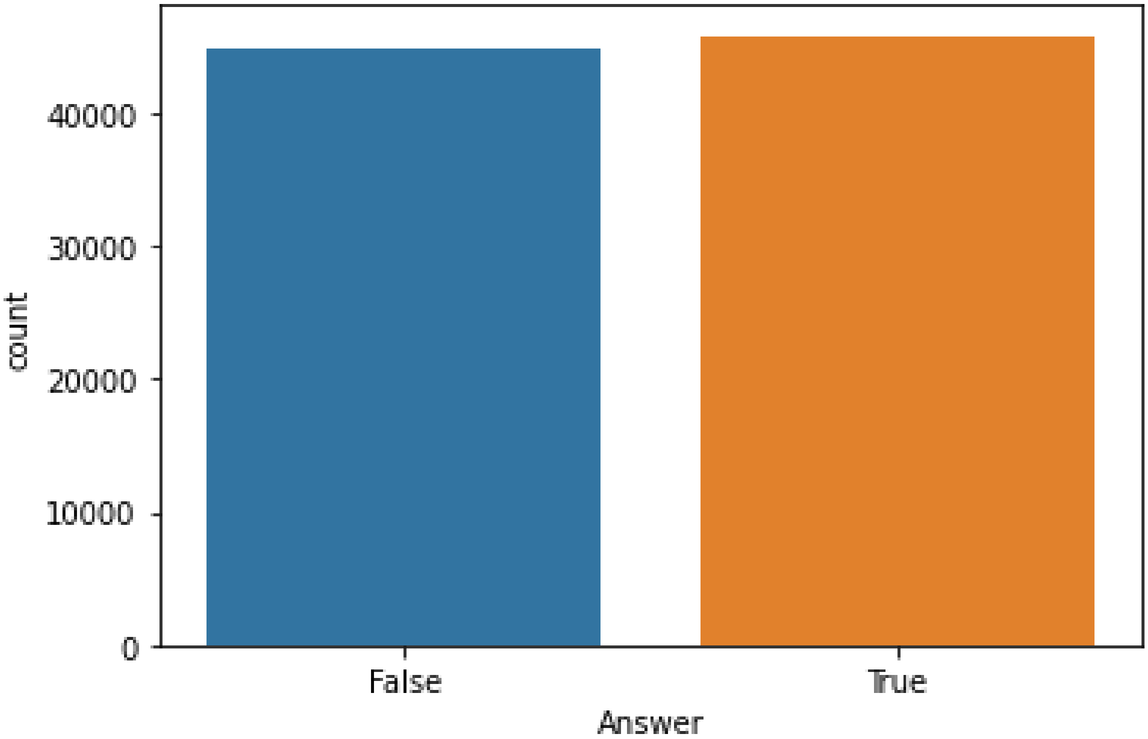
In this chart we see the percentage of questions that could not be solved for each annotator. For example, the annnotator with ID ”780c974e-6ce7-4f1d-9afe-56bf96753fca”has a value of 0.002287021154945683, which suggests that 0.228% of questions for this annotator could not be solved.



In this chart we see the percentage of corrupt data for each annotator. For example, the annotator with ID ”62f1cf90-71c9-45c5-990c-f9dc0aa6ea3d”has a value of 0.00015535187199005747, which suggests that 0.015% of the data for this annotator is corrupt.

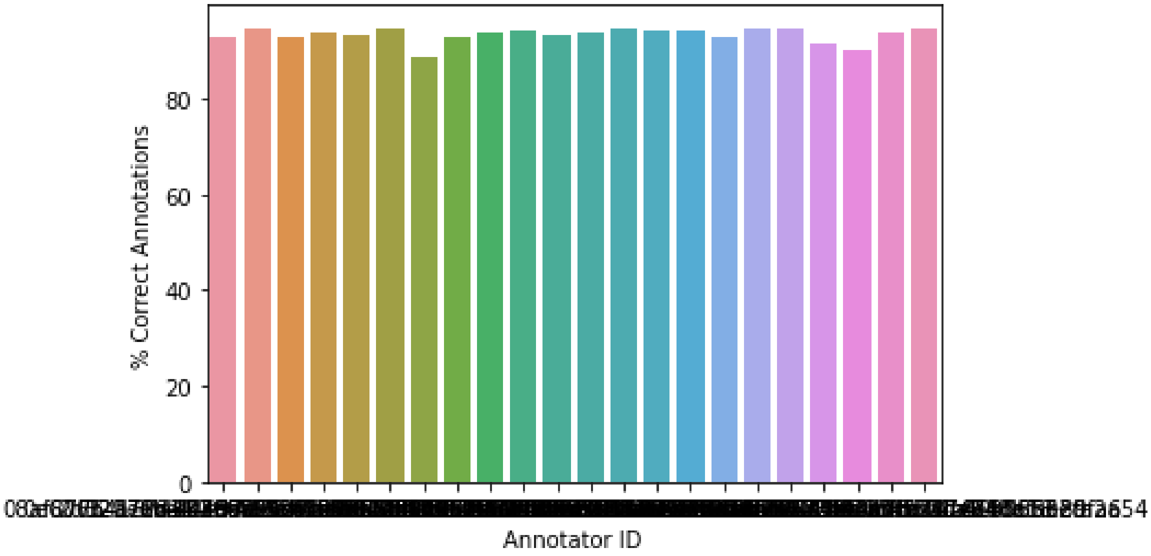
6 It’s important to note that the reference set is slightly unbalanced, but the difference is relatively small, so it’s still reasonable to consider the reference set balanced.

In the below chart we see the distribution of answers in the reference set:



7 Annotator performance varied with some performing better than others

The below chart represents the percentage of correct annotations by annotator ID:



Based on the percentages, it seems that most of the annotators had a high level of accuracy in their annotations, with values ranging from 88.9% to 94.8%. The average percentage of correct annotations across all annotators is approximately 93.6%. However, it's important to note that there are some variations in the accuracy levels between the different annotators, with some having higher accuracy than others. This may suggest that some annotators were more experienced or more attentive than others when reviewing the examples. Overall, the high accuracy levels indicate that the annotators did a good job in completing the task.

In the below table we can see also the Cohen's Kappa score for each annotator :

Annotator ID	Cohen's Kappa Score
08af8775-a72c-4c59-b60f-9ce7df04fa92	0.8591838052950709
62f1cf90-71c9-45c5-990c-f9dc0aa6ea3d	0.8726244950554839
780c974e-6ce7-4f1d-9afe-56bf96753fca	0.8845063777223684
94afdf79-917c-4904-bdaa-7a121f33b0ff	0.8841310362649101
37a6036c-a9ef-4fcb-b554-7058f00415db	0.8696987332912907
66d63556-1235-4a5e-accc-651cb2993db4	0.8846380287985731
75f1e1c8-d127-430f-a885-c2d126364198	0.8968389403060609
0e606b41-ebe4-4d7d-996c-db910c84971a	0.8965200662483188
5765cc32-9e32-43e1-bd78-384cfd52d55a	0.8528661818337308
e336139c-5228-42cd-b501-81855532dfa6	0.8739504567034042
2d521e66-c47c-4033-ba9e-95dcab974ae5	0.8571140567511718
d40ae601-32bf-4b0c-b92c-5bd474e40c18	0.8298240828263477
da2b934c-6384-44ea-adcd-3de648bfae89	0.7983585245359592
9fddf227-6680-42f5-aca5-533fc439c193	0.8562598296647896
fa6bb2a5-c7ec-4dca-aca3-b0666efa2a54	0.8937460339336775
70b9b917-b4f2-4ed1-b8f9-f3492f980103	0.8799840024639883
6d8be901-c6a6-43f8-a5ac-8e44b924bd84	0.8666804616907834
32971d11-e972-4d73-99a6-0847e1261995	0.8727918074851753
4ceeee9d-1588-4b31-82c6-6744d3d00bab	0.7748023775966665
44629999-ab8a-4871-b4d3-4f92c3058161	0.8966835517964197
b04e6596-24c3-4e0b-b96a-4a6d78c04b40	0.8933951332560834
b44e310d-7985-41a0-b764-034b711c75ca	0.8939414945237765

Cohen's kappa score is a measure of inter-annotator agreement that takes into account the agreement that could occur by chance between two annotators. It ranges from -1 to 1, where values closer to 1 indicate a high degree of agreement between annotators, values closer to 0 indicate agreement no better than chance, and values closer to -1 indicate disagreement between annotators.

The scores that we have above suggest a high level of agreement between the annotators, with most of the scores above 0.8, which is considered a substantial agreement. A score above 0.9 is considered almost perfect agreement. It is important to note that Cohen's kappa score is dependent on the number of categories being annotated and the prevalence of those categories in the data.