

Supplementary Material

Figure 1 illustrates the annotation accuracy of turkers on the ComVE dataset, when filtering by the maximum allowable labels for each turker. Unlike the *minimum label* filter reported in the main text, accuracy estimates continue to show high variability for different values of this filter, and a pattern is not evident. Therefore, not all filters (even those that seem reasonably similar or symmetric to one another) are equally viable or effective for obtaining more precise accuracy estimates in the presence of noise.

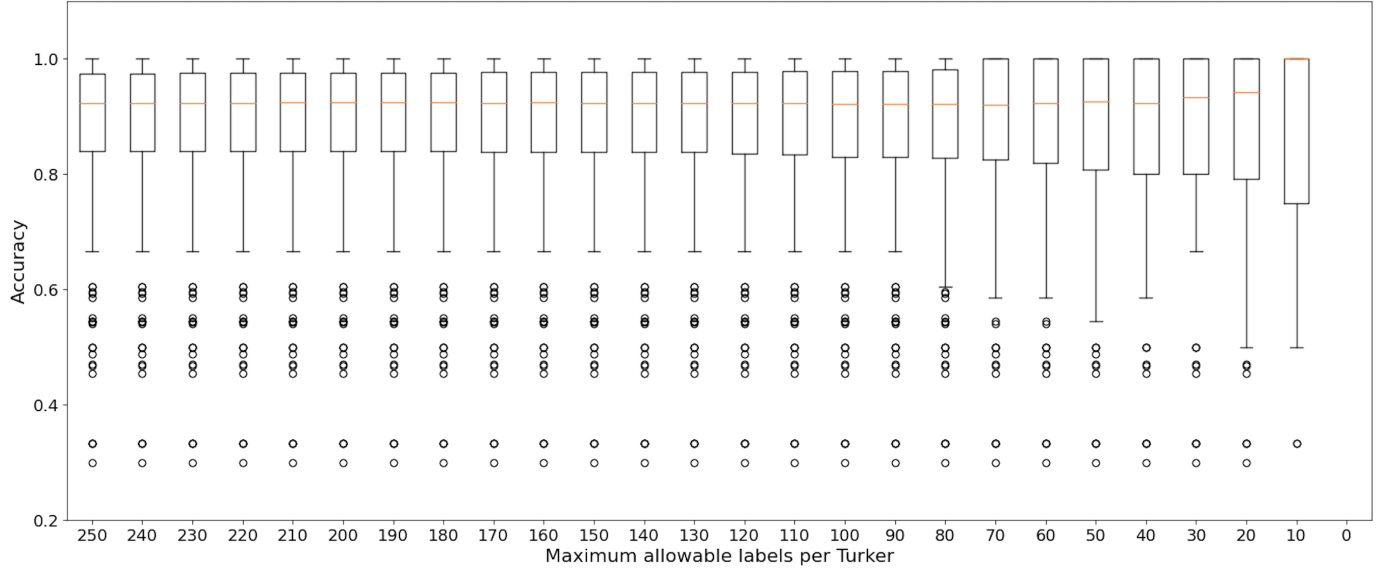


Figure 1. The boxplot of the annotation accuracy of turkers on the ComVE dataset when controlling the maximum allowable labels for each turker, with the ground-truth label for a prompt computed using the mode of all labels acquired from the (four) remaining annotators who labeled that prompt, as described in *Noise Audit and Analysis*.