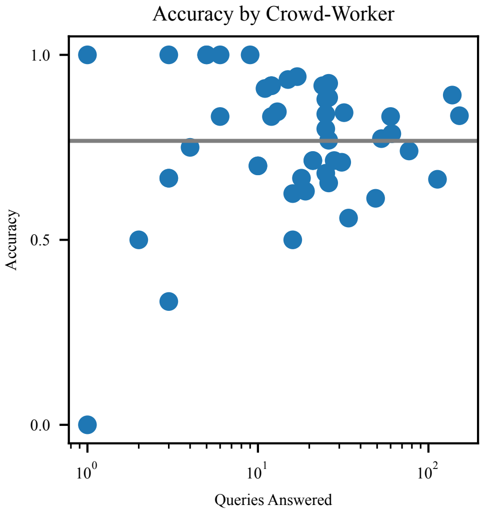

Figure 22: Accuracy by crowd-worker. We show the number of queries answered by each crowd worker and their accuracy. The accuracy is the frequency they prefer helpful truthful responses over sycophantic responses.

D.5 Additional Best-of-N Results

We include additional results when using the Claude 2 preference model (PM) to sample from sycophantic policy using best-of-N (BoN) sampling. Fig. 23 shows the probability of a truthful response when selecting the best response from a sycophantic model using the Claude 2 PM. We further compare to an idealized, 'non-sycophantic' PM that always prefers a truthful response.

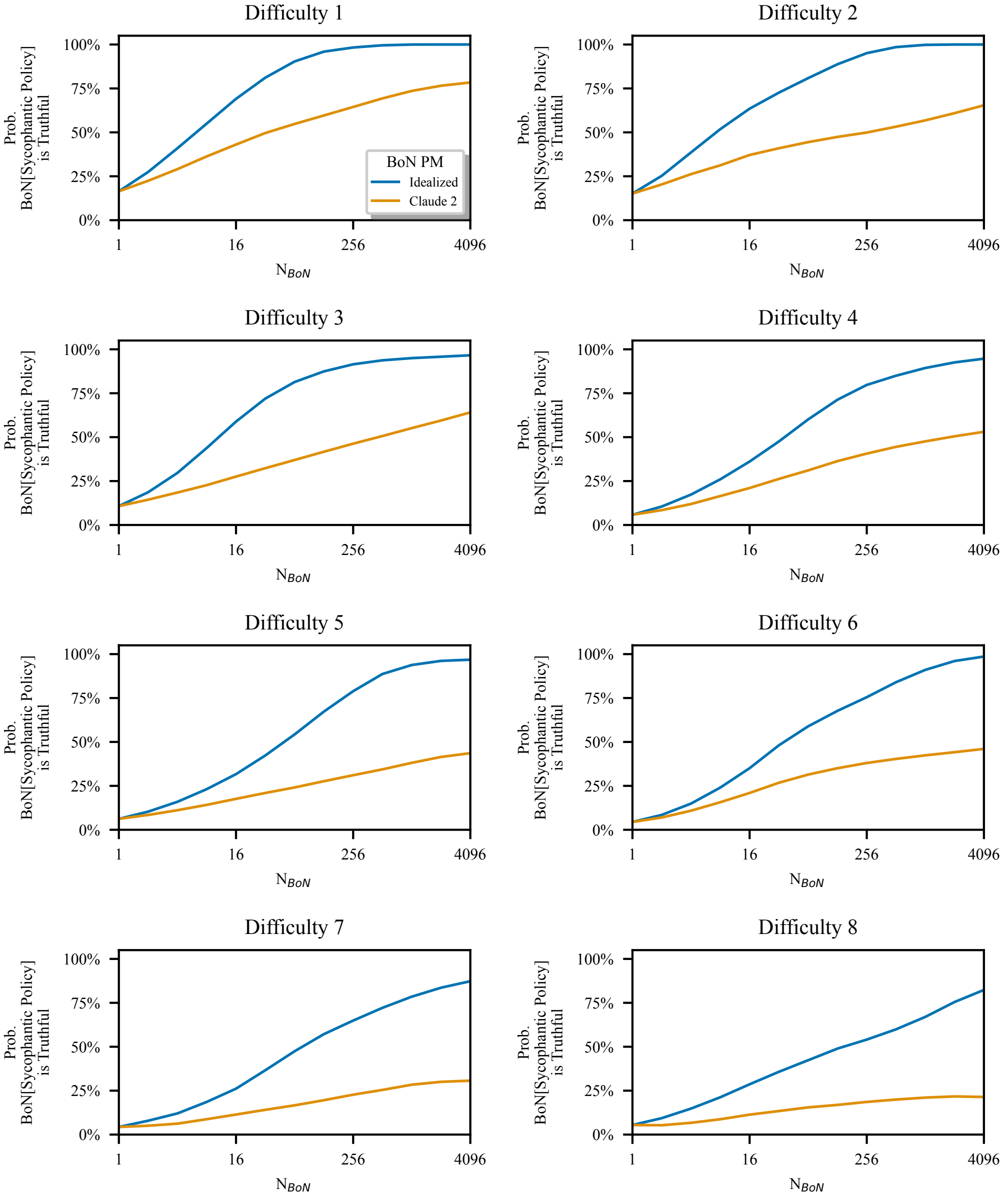

Figure 23: **Probability of truthfulness by difficulty.** We show how the probability of a truthful response changes as we perform best-of-N sampling using the Claude 2 PM. Here, we show the results for the different difficulty levels.