

Figure 1: A line graph showing the performance of various probes over 1000 steps. The y-axis represents performance (0.0 to 1.0) and the x-axis represents steps (0 to 1000). The probes are: average-token-prob (magenta), verbalization-1s (teal), verbalization-2s (orange), p(true) (red), trained-probe (blue), perplexity (purple), jaccard-degree (brown), and ood-probe (grey). The trained-probe (blue) shows the highest performance, reaching near 1.0 by step 1000. The ood-probe (grey) and p(true) (red) show the lowest performance, remaining below 0.5 throughout the training process.

