

Figure 1 is a line graph illustrating the performance of various probes as a function of the Routing Ratio (x-axis, ranging from 0.0 to 1.0). The y-axis represents the performance metric, which is not explicitly labeled but is implied to be a score or probability. The probes shown are:

- average-token-prob (pink line)
- verbalization-1s (teal line)
- verbalization-2s (orange line)
- p(true) (red line)
- trained-probe (blue line)
- perplexity (purple line)
- jaccard-degree (tan line)
- ood-probe (grey line)

The graph shows that the performance of most probes increases as the Routing Ratio increases, with jaccard-degree generally performing best and perplexity performing worst. The trained-probe and ood-probe show a sharp increase in performance as the Routing Ratio approaches 1.0.

