

The graph illustrates the performance of different probes as a function of the Routing Ratio (0.0 to 1.0). The probes 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 (light brown line)
- ood-probe (dark grey line)

The x-axis is labeled "Routing Ratio" and ranges from 0.0 to 1.0. The y-axis represents performance, with a grid line at 0.5. The ood-probe and p(true) lines show the highest performance, starting around 0.45 and reaching 1.0. The trained-probe and perplexity lines start around 0.4 and reach 1.0. The jaccard-degree line starts around 0.35 and reaches 1.0. The verbalization-2s line starts around 0.3 and reaches 1.0. The verbalization-1s line starts around 0.1 and reaches 1.0. The average-token-prob line starts around 0.35 and reaches 1.0.

