

## Cross-Domain Robustness

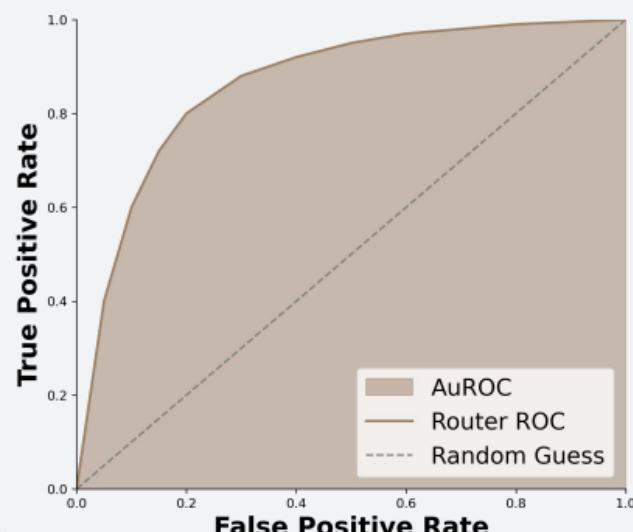
### In-Domain



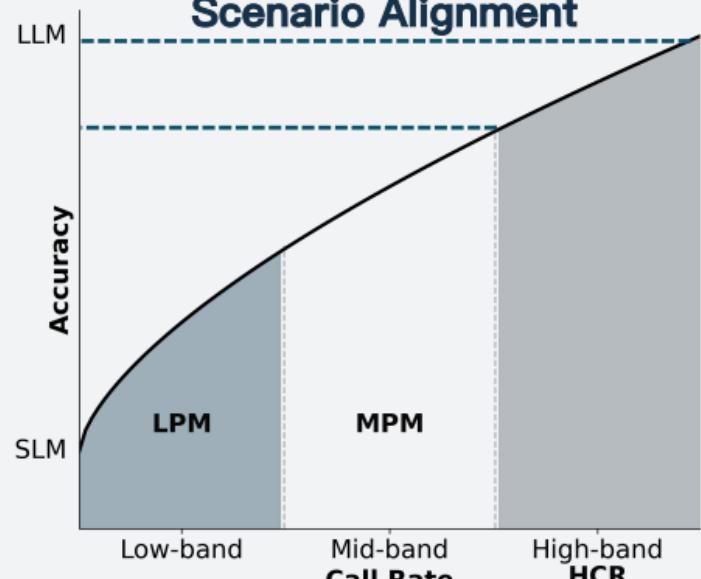
### Out-of-Domain



## Router Ability



## Scenario Alignment



## Evaluation Framework

prompt

Layer 1

Layer 2

Layer L

$Z^{(1)}$

$Z^{(2)}$

$Z^{(L)}$

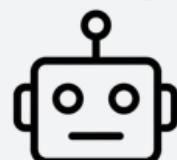
$$\widehat{Z}(x) = \sum_{l=1}^L \alpha_l z^{(l)}(x)$$

$$\alpha \sim Dir(\beta)$$

PROBE

Methodology

Use SLM  
(continue generating)



probe  
score



Call LLM