

Probe Recall@1FPR

- Harmful-related triggers
- Safe-related triggers
- Other triggers
- ★ Strongest red trigger
- ⋯ Probe baseline (0.909)
- Trend line

