Steering Detection: Base vs Trained Models

Detection Rate (%)

- Base Model
- Steering-Aware

Model	Value
Qwen 2.5 32B	95.3%
Qwen 2.5 7B	85.5%
Gemma 2 9B	43.0%