

Compute
Steering Vectors

```
graph LR; A[Compute Steering Vectors] --> B[Generate Training Data]; B --> C[LoRA Fine-tune]; C --> D[Evaluate Detection];
```

The diagram illustrates a four-step process for LoRA fine-tuning. Step 1: 'Compute Steering Vectors' (blue box). Step 2: 'Generate Training Data' (orange box). Step 3: 'LoRA Fine-tune' (green box). Step 4: 'Evaluate Detection' (purple box). Arrows connect the steps in sequence. Below each step, specific parameters or results are listed.

CAA method
 $\text{mean}(h^+) - \text{mean}(h^-)$

Generate
Training Data

50% steered
50% replay

LoRA
Fine-tune

$r=32, \alpha=64$
4 epochs

Evaluate
Detection

95.3% detection
0% FPR