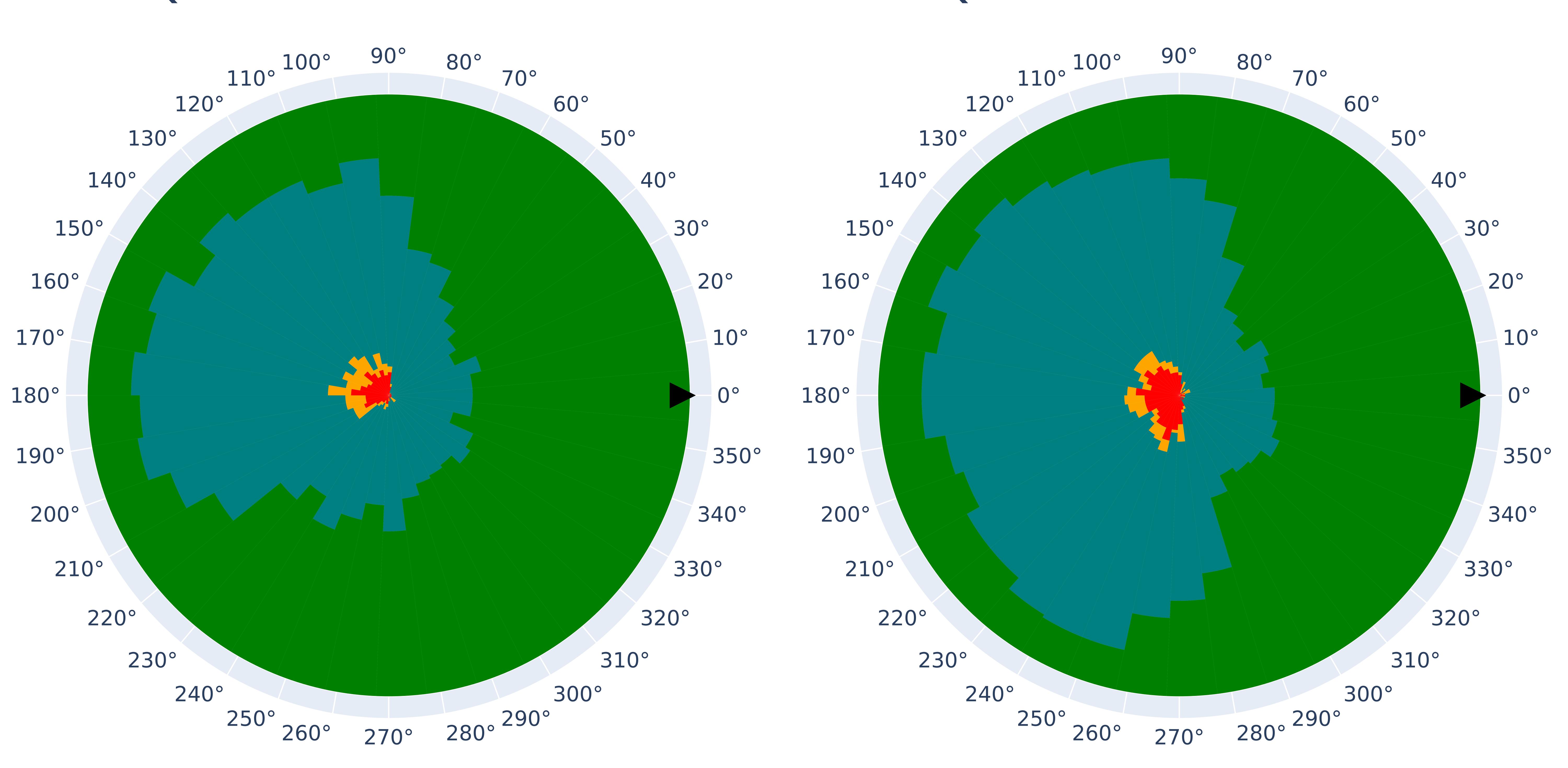
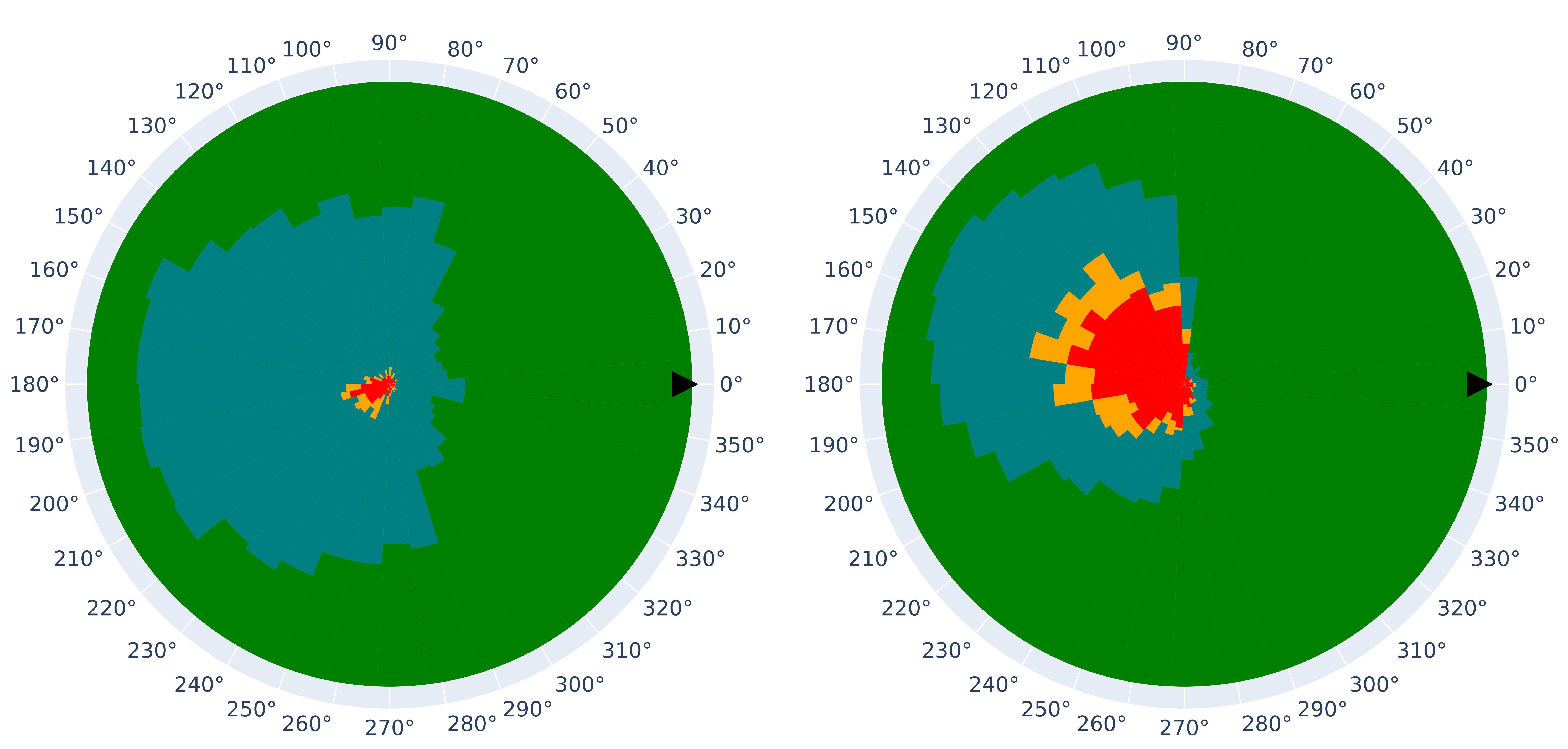
### Qwen2.5-3B-Instruct

#### Qwen2.5-7B-Instruct



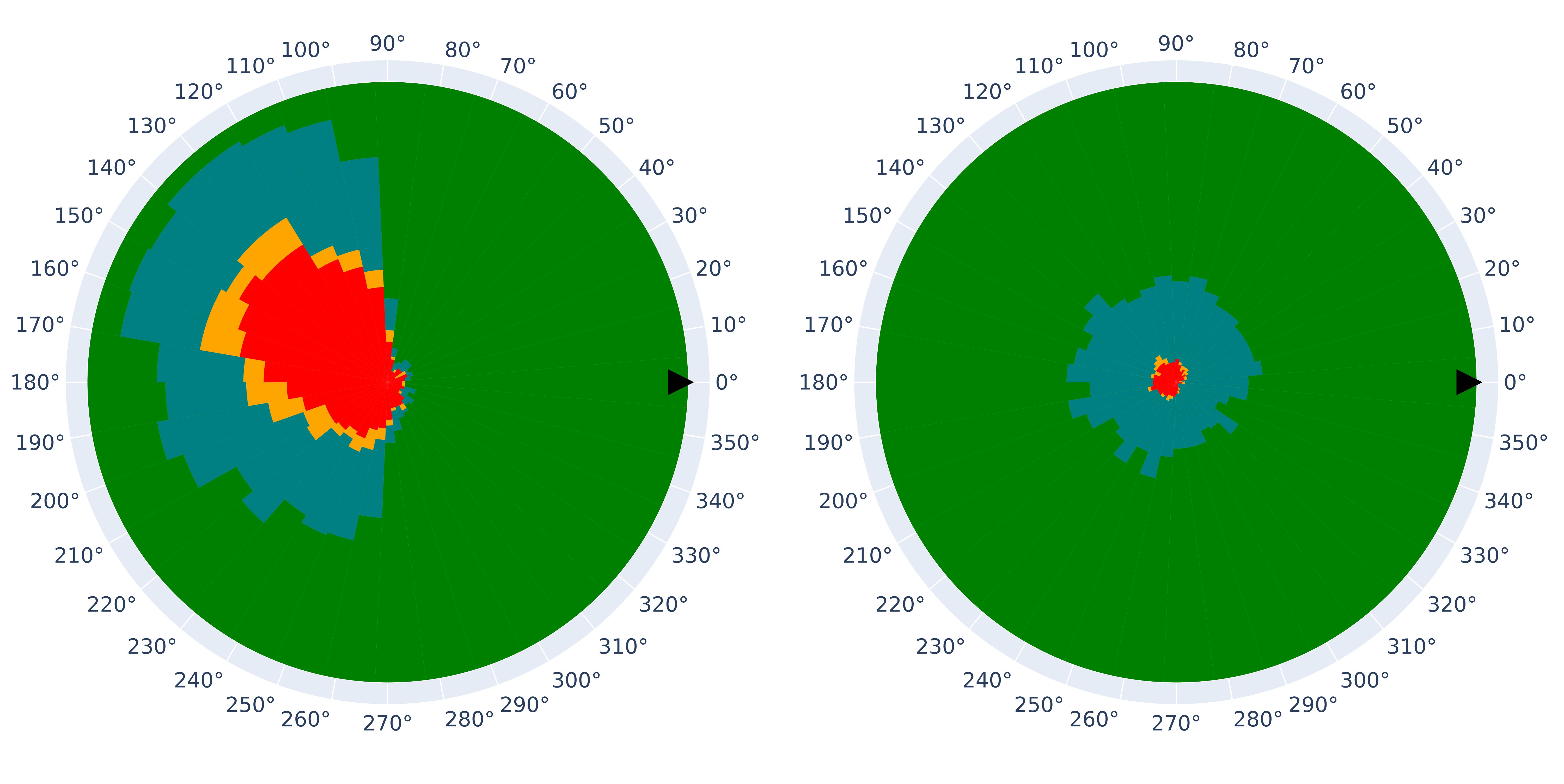
## Qwen2.5-14B-Instruct

### Llama-3.2-3B-Instruct



## Llama-3.1-8B-Instruct

# gemma-2-9b-it



■ direct ■ indirect ■ redirect ■ refusal ► feature direction