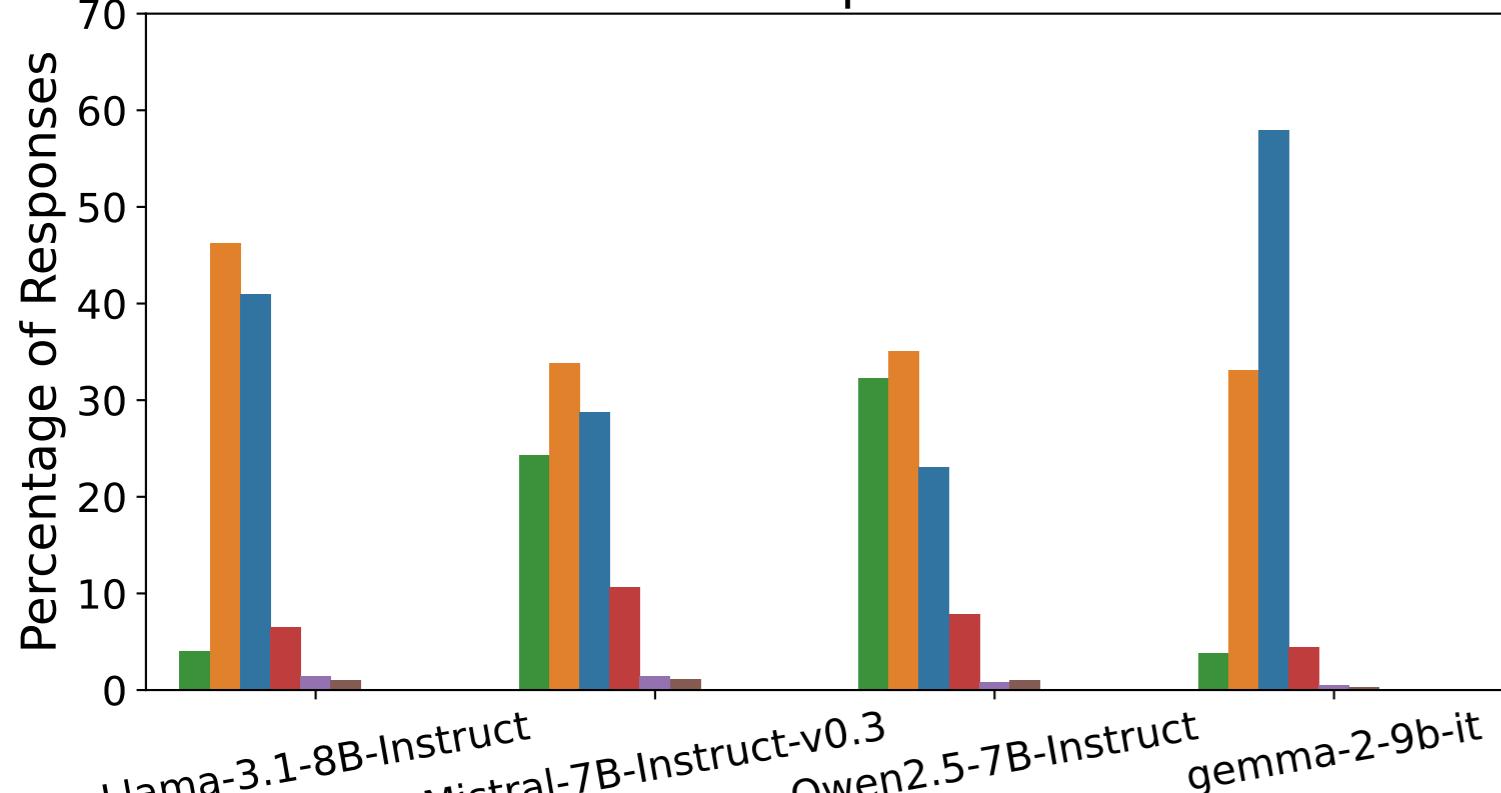
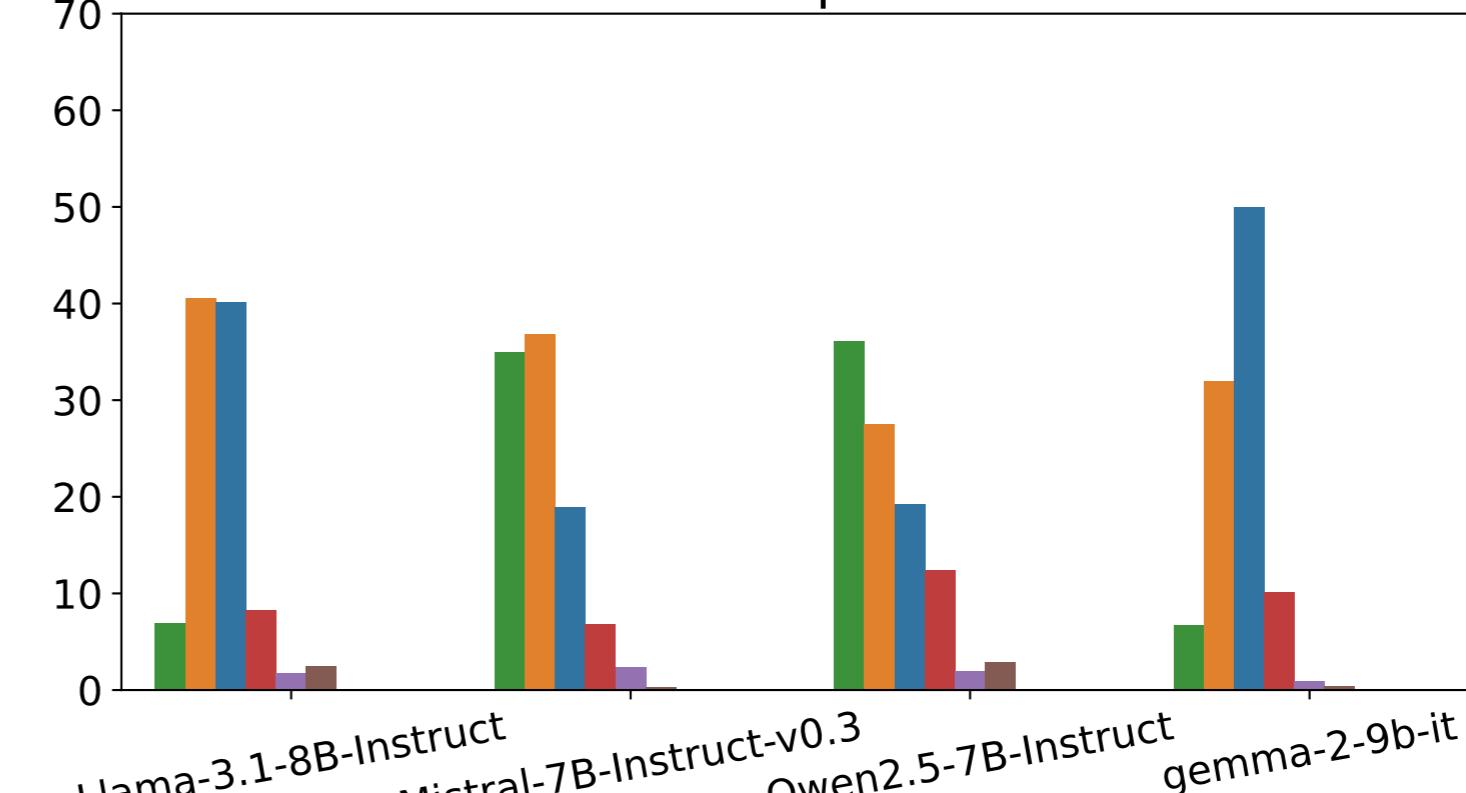


Distribution of Rewarded Responses (%) - Instruction-Tuned

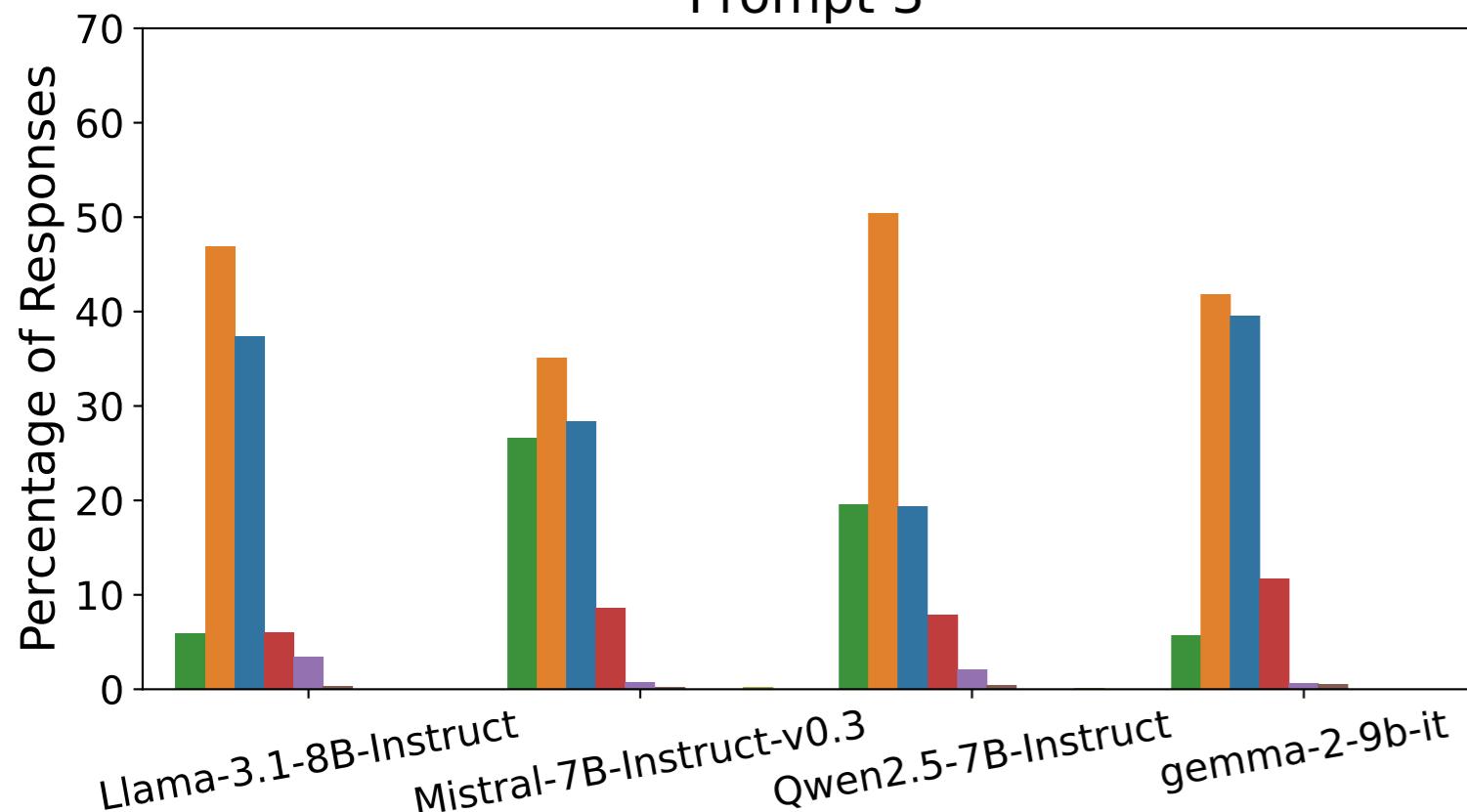
Prompt 1



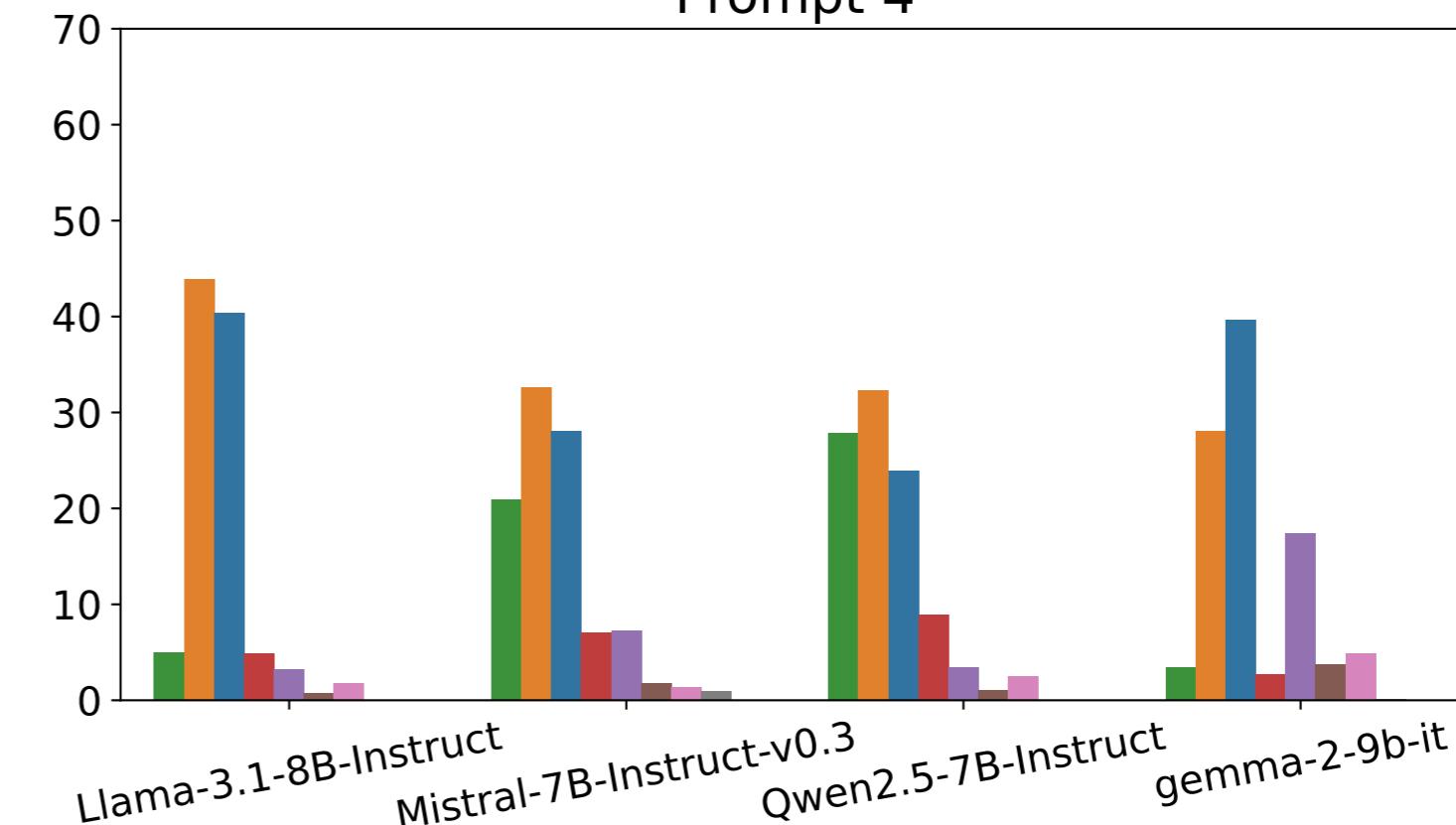
Prompt 2



Prompt 3



Prompt 4



Response Options

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- unusable