Few-shot CoT Performance: In-Domain prompt v.s. Commonsense prompt

Legend:
- GPT-3.5 ID-Prompt
- GPT-3.5 OOD-Prompt
- Llama-70B-chat ID-Prompt
- Llama-70B-chat OOD-Prompt

Y-axis: # of correct answers in the applied exercise subset

X-axis: # of shots