Per-Task Accuracy (SWE-bench Evaluated)

Legend:
- Claude (29/50)
- GPT-5 (16/50)
- Llama (2/50)

Y-axis: Resolved Runs (/5)

X-axis: 12907, 13033, 13236, 13398, 13453, 13579, 13977, 14096, 14182, 14309