Score vs Confidence Threshold: test, winogrande

Legend:
- Falcon 7B
- Falcon 40B
- Llama2 7B
- Llama2 13B
- Llama2 70B
- Mistral 7B
- Mixtral 8x7B
- SOLAR 10.7B
- Yi 6B
- Yi 34B

X-axis: Confidence Threshold
Y-axis: Score