

Score vs Confidence Threshold: train, winogrande

