

Score (harsh) vs Confidence Threshold: train, winogrande

