Score (Balanced) vs Confidence Threshold: MSP, hellaswag