hellaswag_10k 16 models Logistic Regression 1.00 SVM XGBoost baseline 0.98 0.96 accuracy 0.94 0.92 0.90 0.4 0.6 0.5 0.7 0.3 0.8 0.9 1.0 uncertainty threshold