Model	AggreFact-XSUM FTSOTA					
	3 Prompts		5 Prompts		9 Prompts	
	Uncal. (%)	Platt (%)	Uncal. (%)	Platt (%)	Uncal. (%)	Platt (%)
AdaBoost	5.1	2.1	12.0	5.7	18.7	4.3
BernoulliNB	9.2	5.1	15.9	5.0	21.5	7.2
CatBoost	8.2	2.5	8.1	5.2	7.7	6.7
DecisionTree	8.9	6.4	7.2	5.3	9.3	4.9
GradientBoosting	7.5	4.6	6.1	5.2	8.6	5.4
KNeighbors	11.0	4.9	10.4	4.7	9.8	4.5
LabelModel	14.7	4.1	23.8	4.7	22.8	5.3
LDA	6.9	4.5	7.4	6.1	7.1	4.1
LGBM	15.5	4.9	20.0	7.7	15.9	6.1
LogisticRegression	13.4	4.9	6.5	5.8	6.6	3.9
MultinomialNB	8.3	8.1	4.2	**2.6**	4.4	**2.6**
RandomForest	9.8	6.3	6.0	4.1	9.8	4.5
SVC	9.3	**0.9**	8.4	6.5	9.5	5.7
XGB	7.2	6.4	6.9	4.8	16.0	6.9

Table 4: Comparison of the Expected Calibration Error (ECE) for ensembling models before (uncalibrated) and after calibration using Platt Scaling, across various models and prompt pools. The calibration model was trained only on the three non-test datasets.