Model	ECE ↓	Brier score ↓	AUC ↑	Accuracy ↑ $\tau = 0.5$	Accuracy ↑ $\tau = t^*$	Score σ st. dev.
Llama 3 70B (it)	0.15	0.24	0.70	0.60	0.61	0.12
Llama 3 70B	0.09	0.24	0.67	0.55	0.61	0.05
Llama 3 8B (it)	0.19	0.28	0.60	0.57	0.57	0.11
Llama 3 8B	0.08	0.25	0.53	0.56	0.54	0.03
Mixtral 8x22B (it)	0.32	0.33	0.66	0.59	0.61	0.15
Mixtral 8x22B	0.20	0.28	0.63	0.44	0.54	0.04
Mixtral 8x7B (it)	0.45	0.45	0.66	0.52	0.55	0.29
Mixtral 8x7B	0.28	0.32	0.60	0.44	0.56	0.03
Mistral 7B (it)	0.41	0.42	0.59	0.57	0.53	0.19
Mistral 7B	0.05	0.25	0.57	0.56	0.55	0.02
Yi 34B (chat)	0.35	0.36	0.65	0.56	0.52	0.02
Yi 34B	0.08	0.24	0.62	0.56	0.52	0.06
Gemma 7B (it)	0.42	0.43	0.53	0.56	0.50	0.02
Gemma 7B	0.04	0.24	0.61	0.58	0.57	0.02
Gemma 2B (it)	0.34	0.36	0.49	0.56	0.54	0.03
Gemma 2B	0.09	0.26	0.48	0.44	0.45	0.01
LR	0.04	0.24	0.58	0.56	-	-
XGBoost	0.02	0.19	0.77	0.70	-	-

Table 1: Zero-shot LLM results on the **ACSTravelTime** benchmark task, together with supervised learning baselines fitted on 1.3M samples.

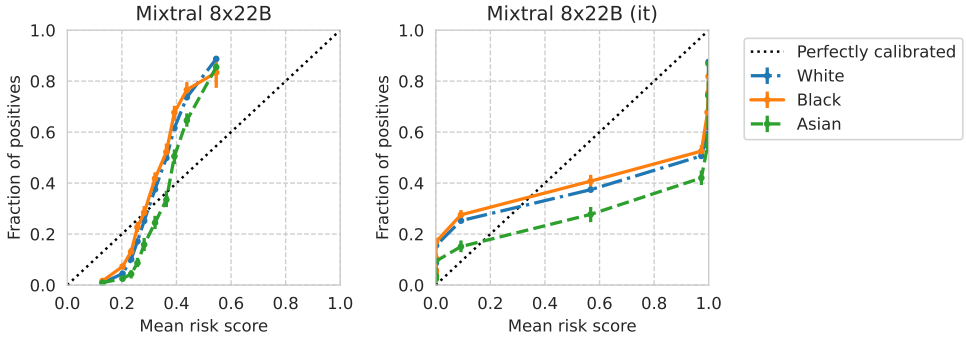

Figure 1: Calibration curves per sub-group for the Mixtral 8x22B models on the ACSIncome task.

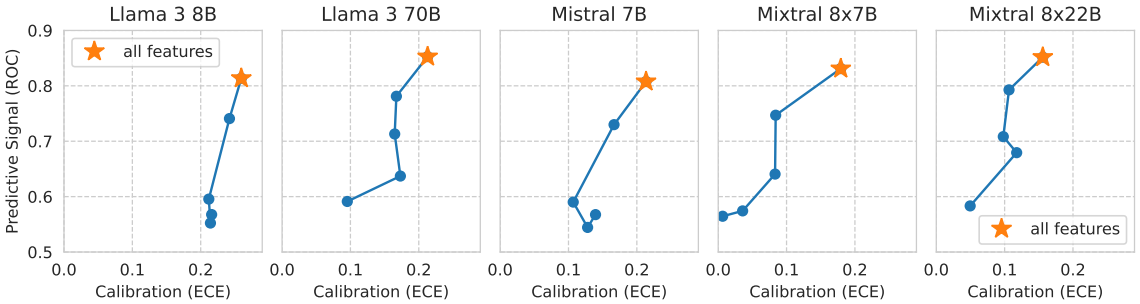

Figure 2: Evaluation of calibration (ECE) and predictive performance (AUC) on base Llama and Mistral models, with an increasing number of features provided on the ACSIncome task. For each dot along the line we add two features, up to all 10 features being used in the point marked with a star. Models can successfully use each extra feature to increase predictive signal, but calibration trends worse the more features are added.