

Table 4: R^2 of linear regression models fit using different stratification features. The symbol “*” indicates statistical significance at confidence level $\alpha = 0.05$ with respect to the null model, according to the likelihood ratio test.

	TREC robust 2004			TREC DL 2019			TREC DL 2020		
	Mistral 7B	Llama 8B	Qwen 30B	Mistral 7B	Llama 8B	Qwen 30B	Mistral 7B	Llama 8B	Qwen 30B
A	0.211 *	0.622 *	0.538 *	0.249 *	0.313 *	0.310 *	0.389 *	0.458 *	0.440 *
ΔP	0.097 *	0.031 *	0.0 (9.551e-01)	0.023 *	0.013 *	0.003 *	0.063 *	0.010 *	0.026 *
ΔP_2	0.092 *	0.113 *	0.008 *	0.022 *	0.047 *	0.017 *	0.065 *	0.065 *	0.071 *
P	0.145 *	0.173 *	0.256 *	0.084 *	0.027 *	0.027 *	0.147 *	0.032 *	0.104 *
PPL	0.040 *	0.026 *	0.015 *	0.096 *	0.017 *	0.028 *	0.213 *	0.007 *	0.043 *
$A + P$	0.238 *	0.634 *	0.569 *	0.253 *	0.331 *	0.311 *	0.390 *	0.474 *	0.450 *
$A + PPL$	0.225 *	0.623 *	0.541 *	0.250 *	0.313 *	0.311 *	0.390 *	0.458 *	0.440 *
$A + \Delta P$	0.238 *	0.632 *	0.555 *	0.249 *	0.338 *	0.315 *	0.389 *	0.479 *	0.458 *
$A + \Delta P_2$	0.238 *	0.631 *	0.563 *	0.249 *	0.328 *	0.314 *	0.389 *	0.474 *	0.458 *
$P + PPL$	0.168 *	0.182 *	0.257 *	0.113 *	0.042 *	0.041 *	0.224 *	0.037 *	0.113 *
$P + \Delta P$	0.156 *	0.197 *	0.427 *	0.088 *	0.027 *	0.038 *	0.169 *	0.034 *	0.117 *
$P + \Delta P_2$	0.155 *	0.173 *	0.428 *	0.088 *	0.047 *	0.027 *	0.170 *	0.065 *	0.104 *
$PPL + \Delta P$	0.119 *	0.052 *	0.016 *	0.103 *	0.029 *	0.029 *	0.233 *	0.017 *	0.060 *
$PPL + \Delta P_2$	0.115 *	0.126 *	0.018 *	0.102 *	0.059 *	0.038 *	0.234 *	0.069 *	0.096 *
$\Delta P + \Delta P_2$	0.101 *	0.194 *	0.121 *	0.023 *	0.067 *	0.077 *	0.066 *	0.115 *	0.182 *
$A + P + PPL$	0.250 *	0.634 *	0.569 *	0.253 *	0.331 *	0.312 *	0.391 *	0.475 *	0.450 *