K-ClassicBench: Model Performance Heatmap

Model	classification	nli	punctuation	retrieval	translation
Claude-3-Opus	0.000	0.300	0.647	0.900	0.172
Claude-3.5-Sonnet	0.100	0.000	0.387	0.900	0.142
EXAONE-3.0-7.8B	0.000	0.000	0.358	0.900	0.195
GPT-3.5-Turbo	0.000	0.700	0.973	0.800	0.227
GPT-4-Turbo	0.000	0.600	0.808	0.900	0.274
Llama-3.1-8B	0.200	0.000	0.922	0.400	0.162
Qwen2.5-7B	0.000	0.000	0.869	1.000	0.102
gpt-4	0.000	0.500	0.910	1.000	0.368