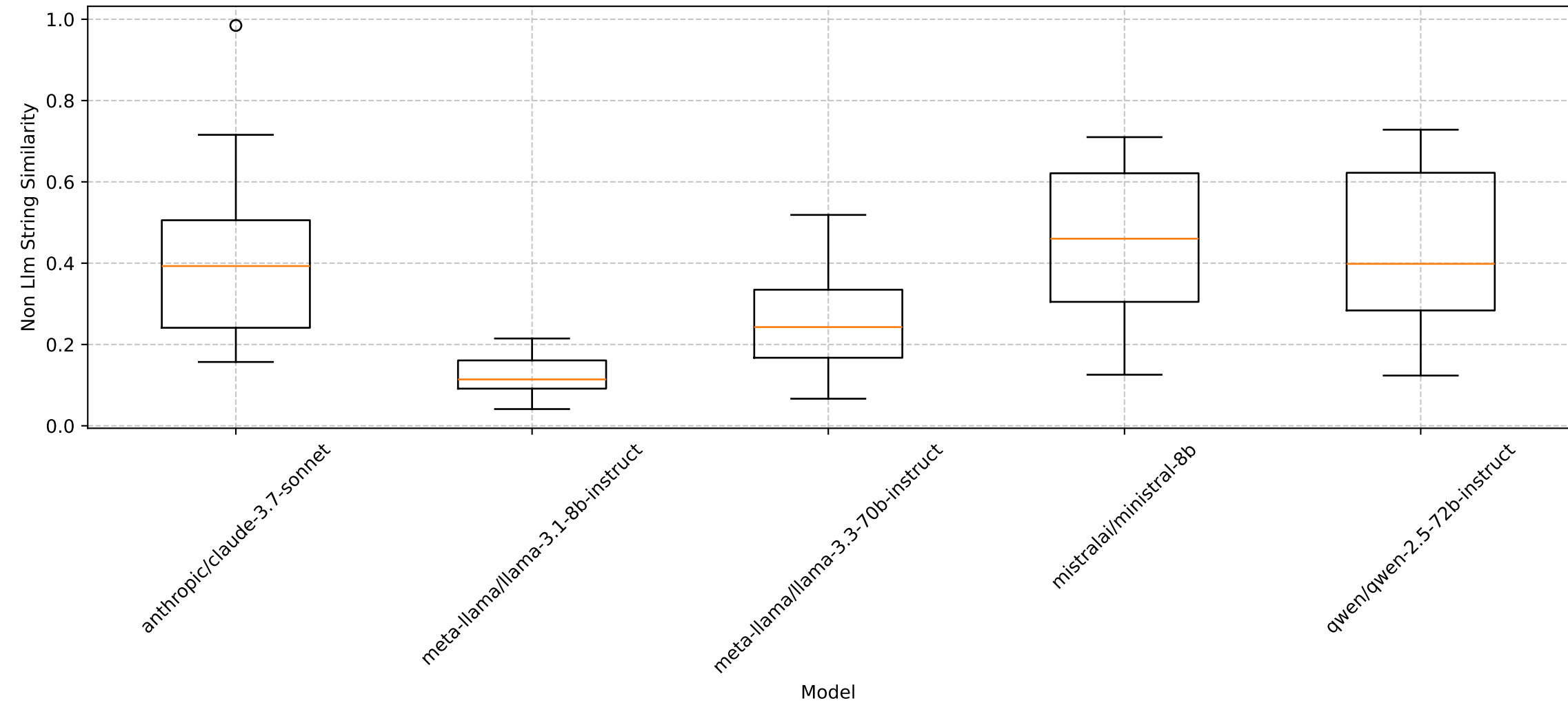
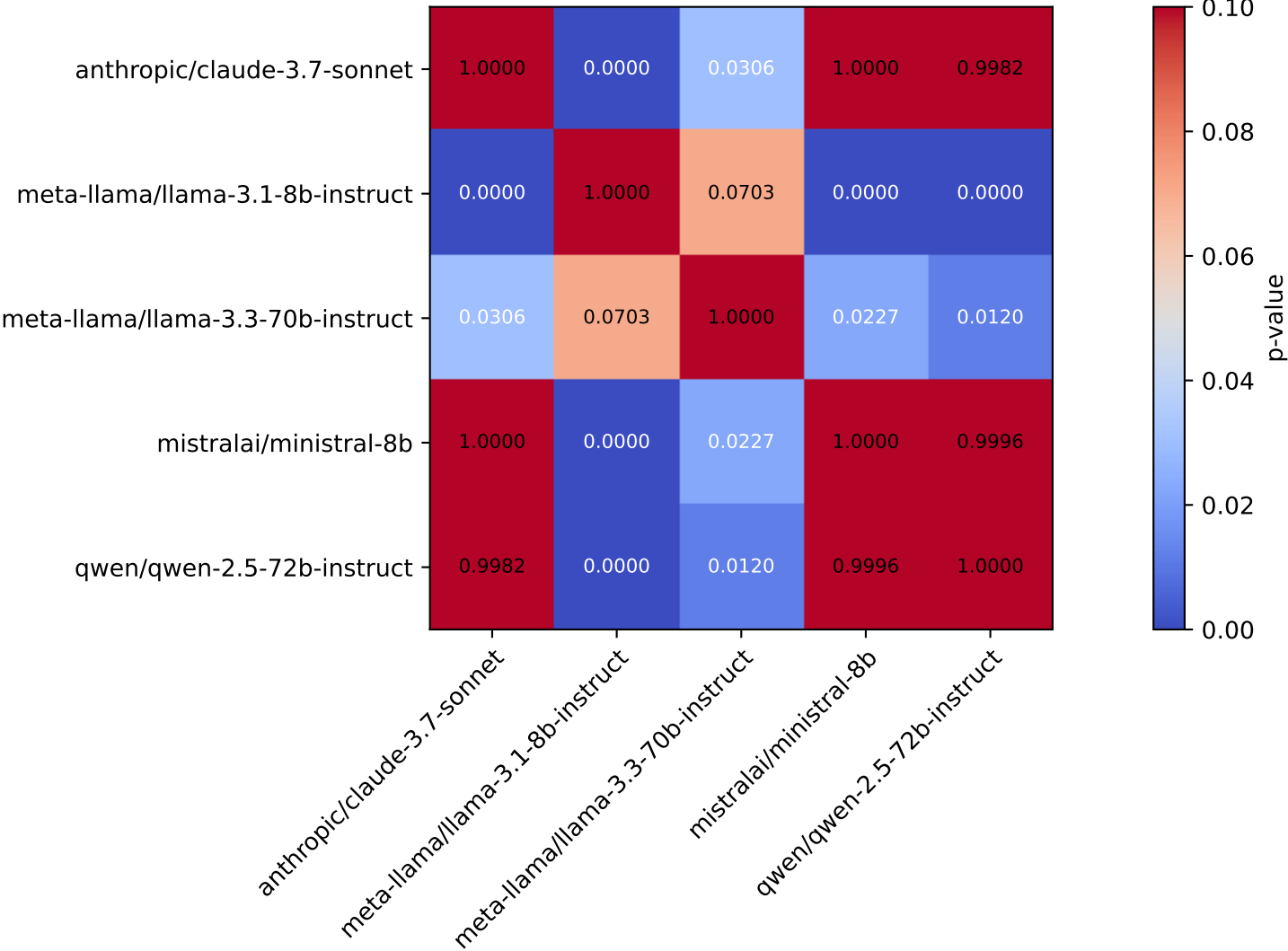


Comparison of non_llm_string_similarity across models
Friedman $\chi^2 = 51.9200$, $p = 0.0000$ (significant)



Post-hoc Nemenyi Test p-values



Pairwise Wilcoxon Signed-Rank Tests

Comparison	Statistic	p-value	Significant (p<0.05)
anthropic/claude-3.7-sonnet vs meta-llama/llama-3.1-8b-instruct	0.0000	0.0000	✓
anthropic/claude-3.7-sonnet vs meta-llama/llama-3.3-70b-instruct	12.0000	0.0001	✓
anthropic/claude-3.7-sonnet vs mistralai/ministral-8b	82.0000	0.4091	
anthropic/claude-3.7-sonnet vs qwen/qwen-2.5-72b-instruct	93.0000	0.6742	
meta-llama/llama-3.1-8b-instruct vs meta-llama/llama-3.3-70b-instruct	11.0000	0.0001	✓
meta-llama/llama-3.1-8b-instruct vs mistralai/ministral-8b	0.0000	0.0000	✓
meta-llama/llama-3.1-8b-instruct vs qwen/qwen-2.5-72b-instruct	0.0000	0.0000	✓
meta-llama/llama-3.3-70b-instruct vs mistralai/ministral-8b	3.0000	0.0000	✓
meta-llama/llama-3.3-70b-instruct vs qwen/qwen-2.5-72b-instruct	14.0000	0.0002	✓
mistralai/ministral-8b vs qwen/qwen-2.5-72b-instruct	96.0000	0.7562	