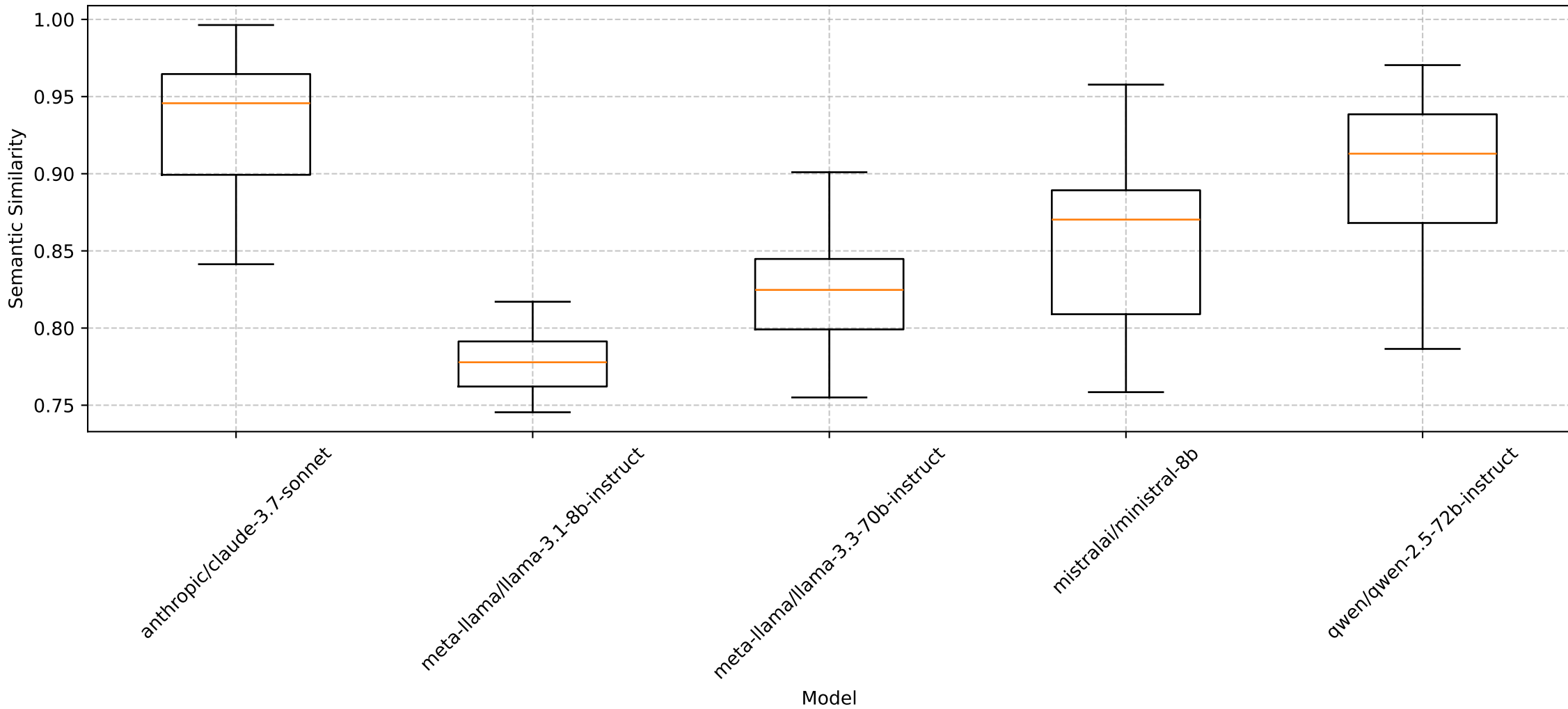
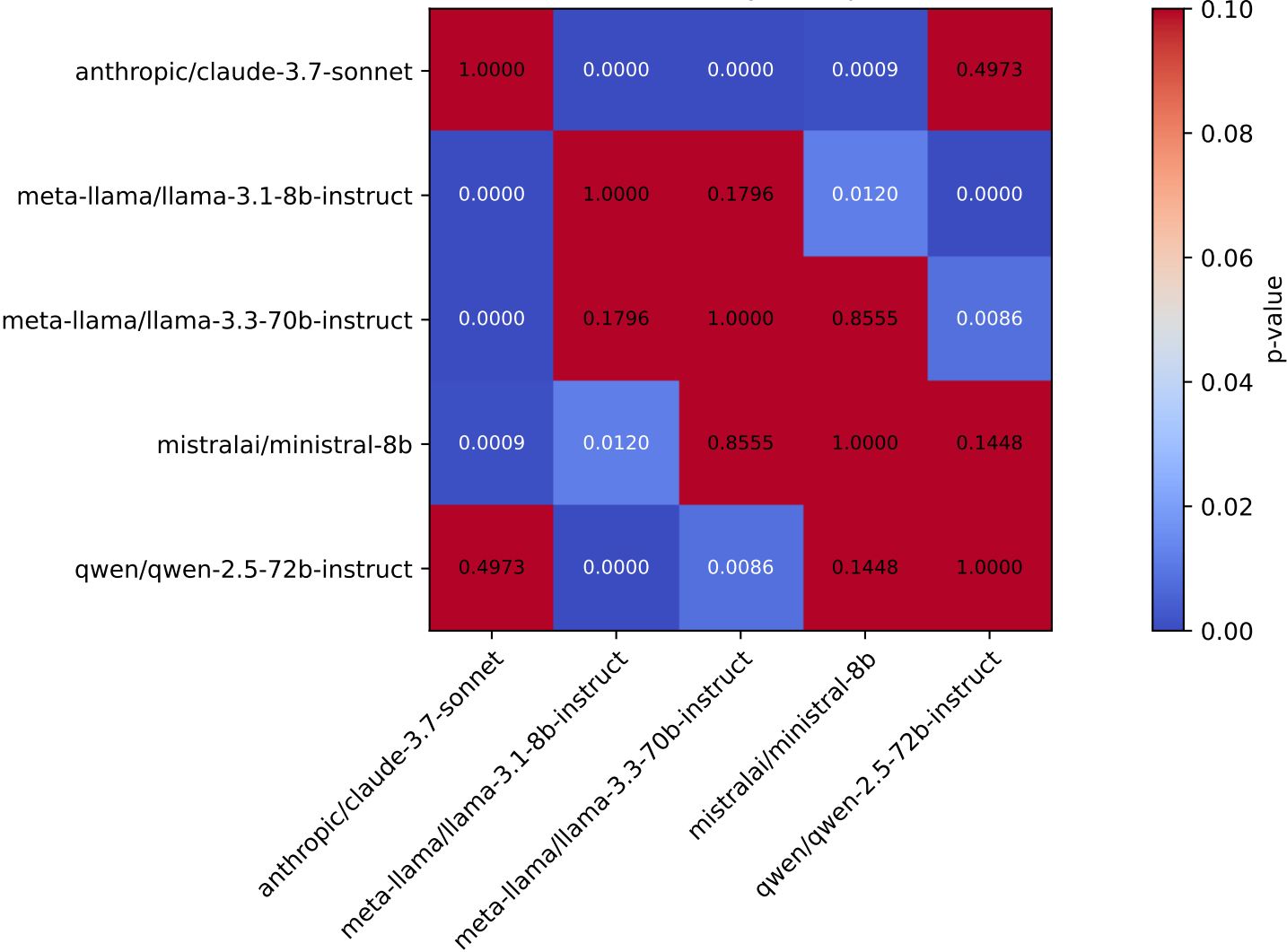


Comparison of semantic_similarity across models
Friedman $\chi^2 = 61.8800$, $p = 0.0000$ (significant)



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	0.0000	0.0000	✓
anthropic/claude-3.7-sonnet vs mistralai/ministral-8b	1.0000	0.0000	✓
anthropic/claude-3.7-sonnet vs qwen/qwen-2.5-72b-instruct	19.0000	0.0006	✓
meta-llama/llama-3.1-8b-instruct vs meta-llama/llama-3.3-70b-instruct	10.0000	0.0001	✓
meta-llama/llama-3.1-8b-instruct vs mistralai/ministral-8b	5.0000	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	60.0000	0.0973	
meta-llama/llama-3.3-70b-instruct vs qwen/qwen-2.5-72b-instruct	2.0000	0.0000	✓
mistralai/ministral-8b vs qwen/qwen-2.5-72b-instruct	33.0000	0.0056	✓