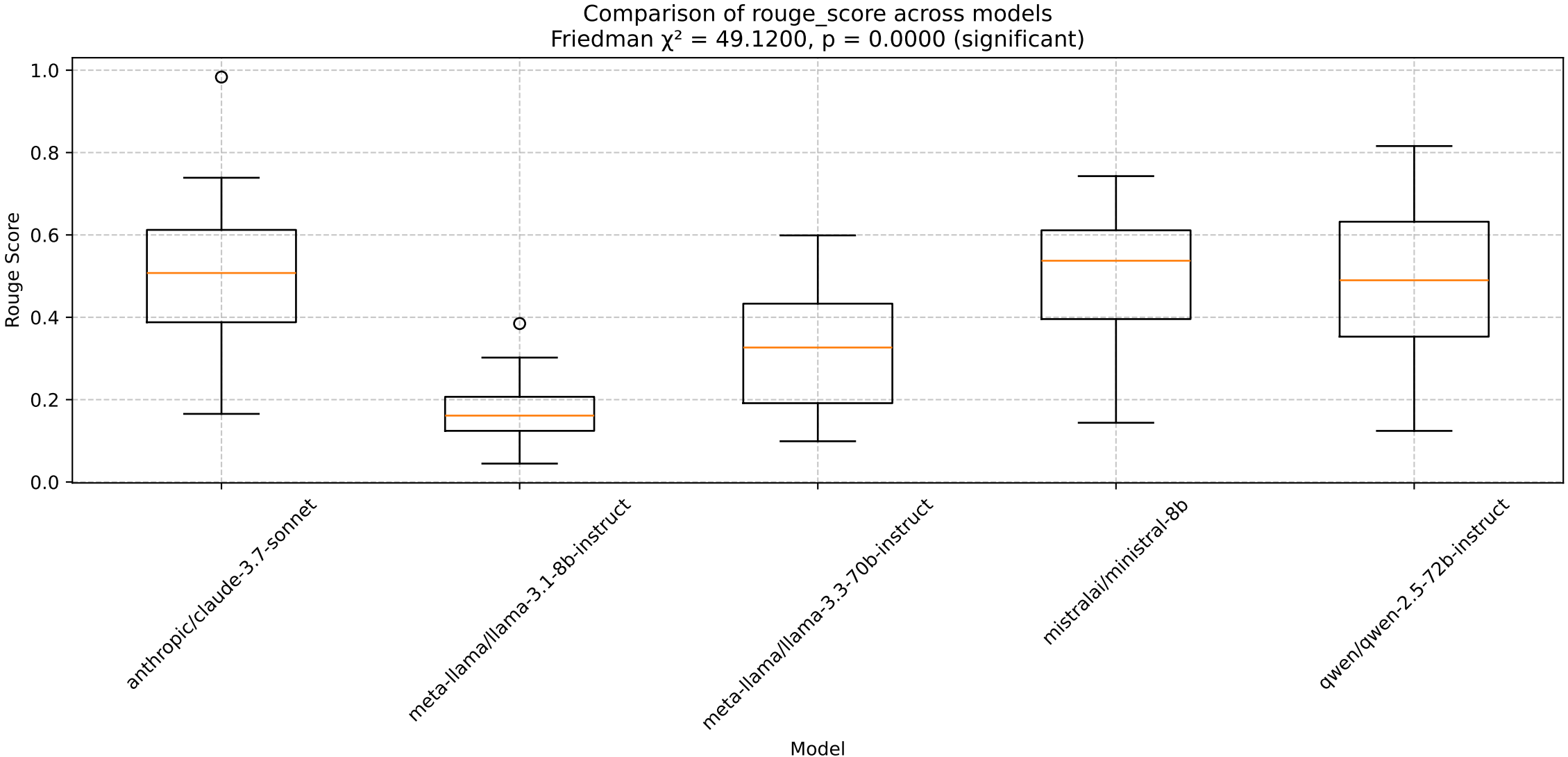
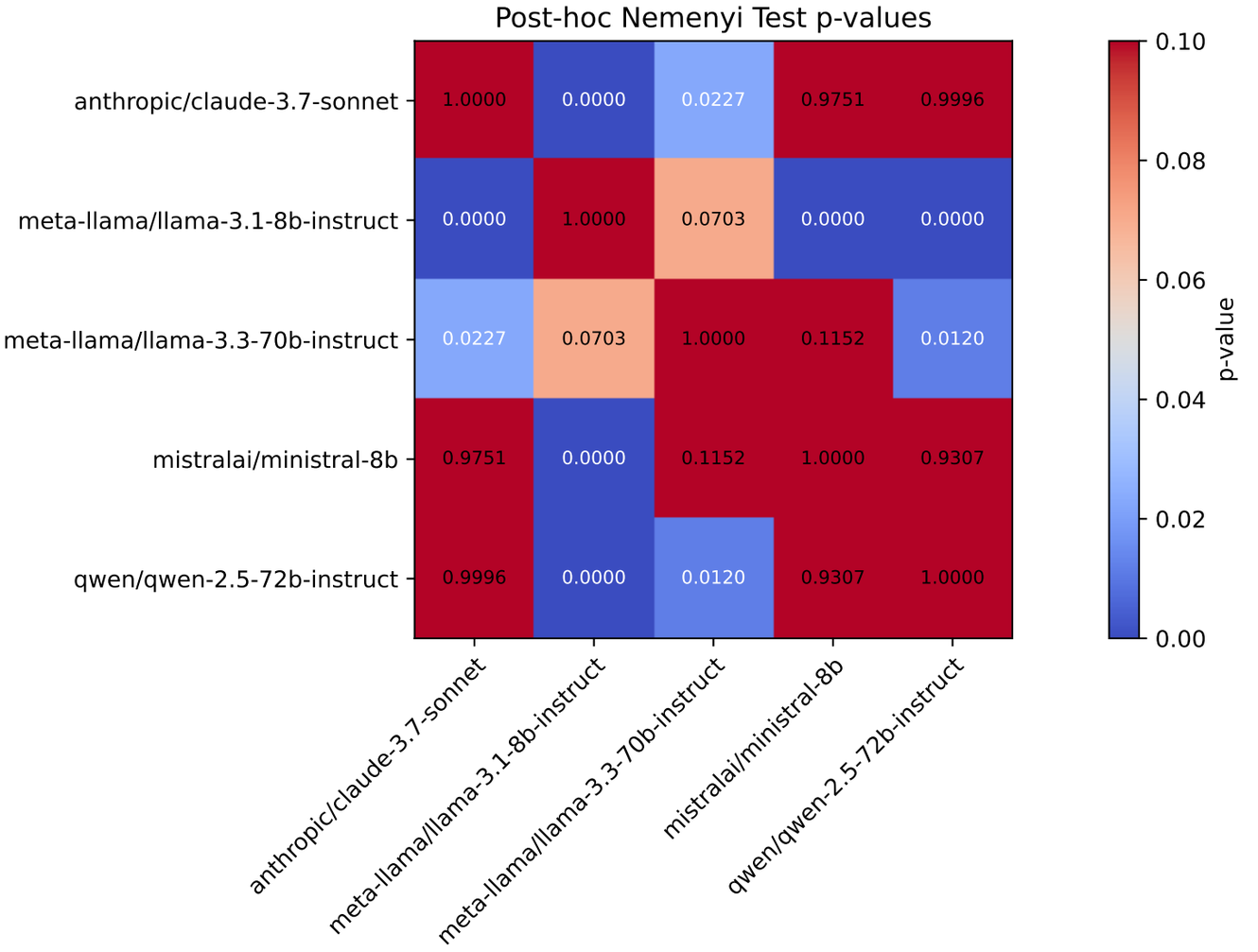

Comparison of rouge_score across models
Friedman χ² = 49.1200, p = 0.0000 (significant)
Rouge Score
1.0
0.8
0.6
0.4
0.2
0.0
anthropic/claude-3.7-sonnet
meta-llama/llama-3.1-8b-instruct
meta-llama/llama-3.3-70b-instruct
mistralai/ministral-8b
qwen/qwen-2.5-72b-instruct
Model



Pairwise Wilcoxon Signed-Rank Tests

Comparison	Statistic	p-value	Significant (p<0.05)
pic/claude-3.7-sonnet vs meta-llama/llama-3.1-8b	0.0000	0.0000	✓
pic/claude-3.7-sonnet vs meta-llama/llama-3.3-70	14.0000	0.0002	✓
anthropic/claude-3.7-sonnet vs mistralai/ministral	92.0000	0.6477	
ropic/claude-3.7-sonnet vs qwen/qwen-2.5-72b-i	92.0000	0.6477	
a/llama-3.1-8b-instruct vs meta-llama/llama-3.3-	18.0000	0.0005	✓
ta-llama/llama-3.1-8b-instruct vs mistralai/minist	0.0000	0.0000	✓
lama/llama-3.1-8b-instruct vs qwen/qwen-2.5-72b	0.0000	0.0000	✓
a-llama/llama-3.3-70b-instruct vs mistralai/minist	11.0000	0.0001	✓
ama/llama-3.3-70b-instruct vs qwen/qwen-2.5-72	20.0000	0.0007	✓
mistralai/ministral-8b vs qwen/qwen-2.5-72b-instr	85.0000	0.4749	