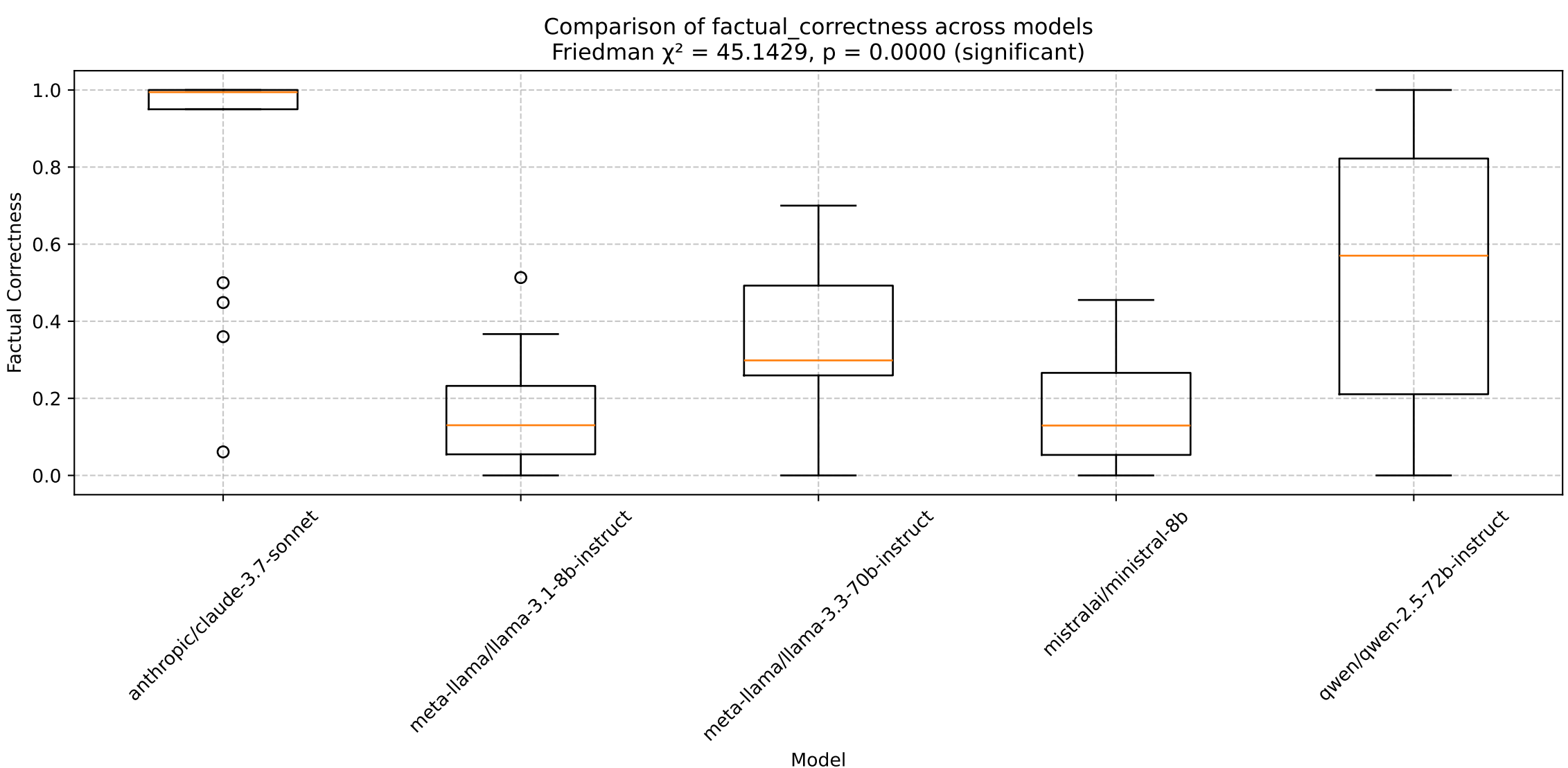
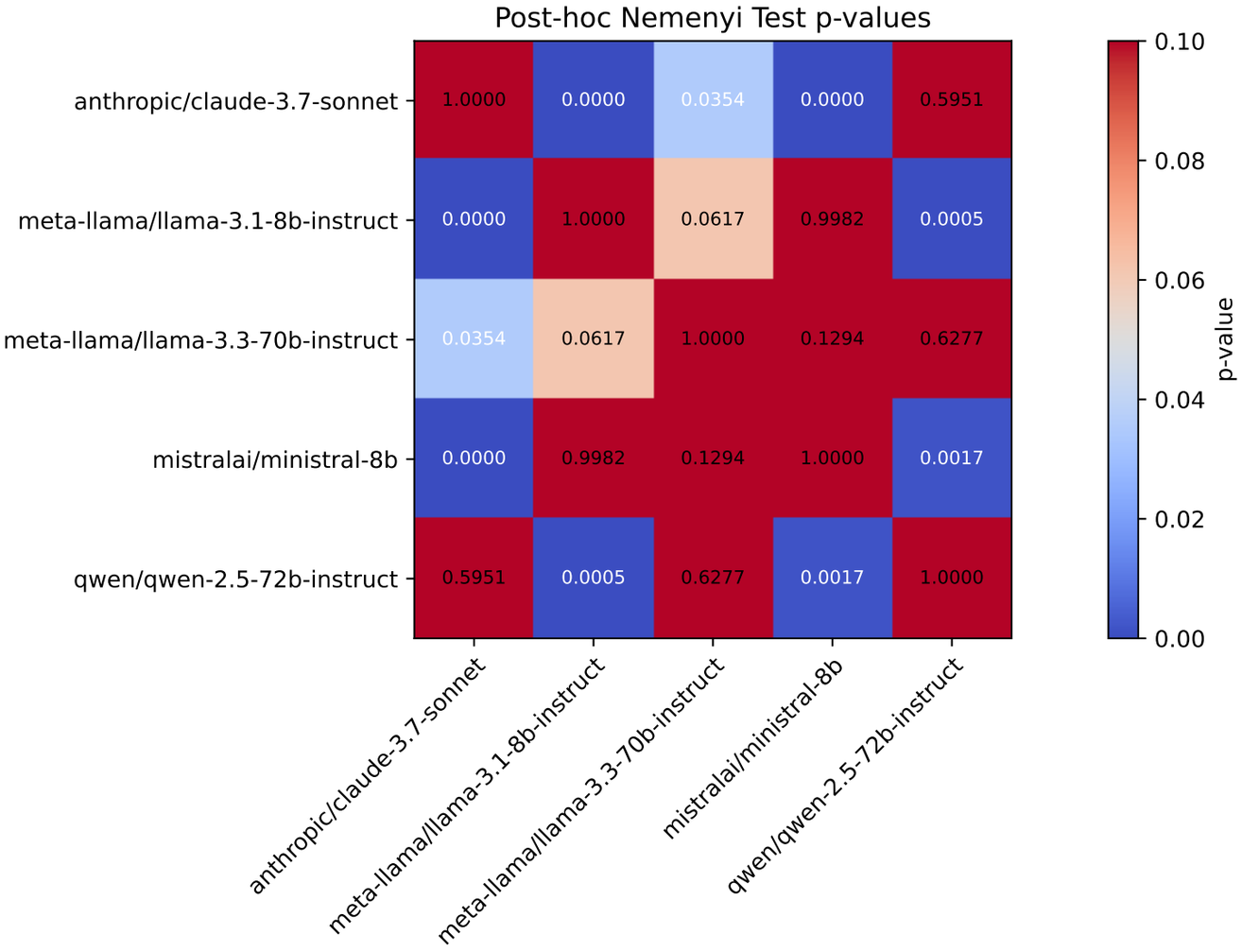

Comparison of factual_correctness across models
Friedman χ² = 45.1429, p = 0.0000 (significant)
Factual Correctness
1.0
0.8
0.6
0.4
0.2
0.0
anthropic/claude-3.7-sonnet
meta-llama/llama-3.1-8b-instruct
meta-llama/llama-3.3-70b-instruct
mistralai/ministral-8b
qwen/qwen-2.5-72b-instruct
Model



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	2.0000	0.0000	✓
pic/claude-3.7-sonnet vs meta-llama/llama-3.3-70	6.0000	0.0000	✓
anthropic/claude-3.7-sonnet vs mistralai/ministral	2.0000	0.0000	✓
ropic/claude-3.7-sonnet vs qwen/qwen-2.5-72b-i	25.0000	0.0017	✓
a/llama-3.1-8b-instruct vs meta-llama/llama-3.3-	22.0000	0.0010	✓
ta-llama/llama-3.1-8b-instruct vs mistralai/minist	98.0000	0.8124	
lama/llama-3.1-8b-instruct vs qwen/qwen-2.5-72b	13.0000	0.0002	✓
a-llama/llama-3.3-70b-instruct vs mistralai/minist	22.0000	0.0010	✓
ama/llama-3.3-70b-instruct vs qwen/qwen-2.5-72	37.0000	0.0196	✓
mistralai/ministral-8b vs qwen/qwen-2.5-72b-instr	13.0000	0.0002	✓