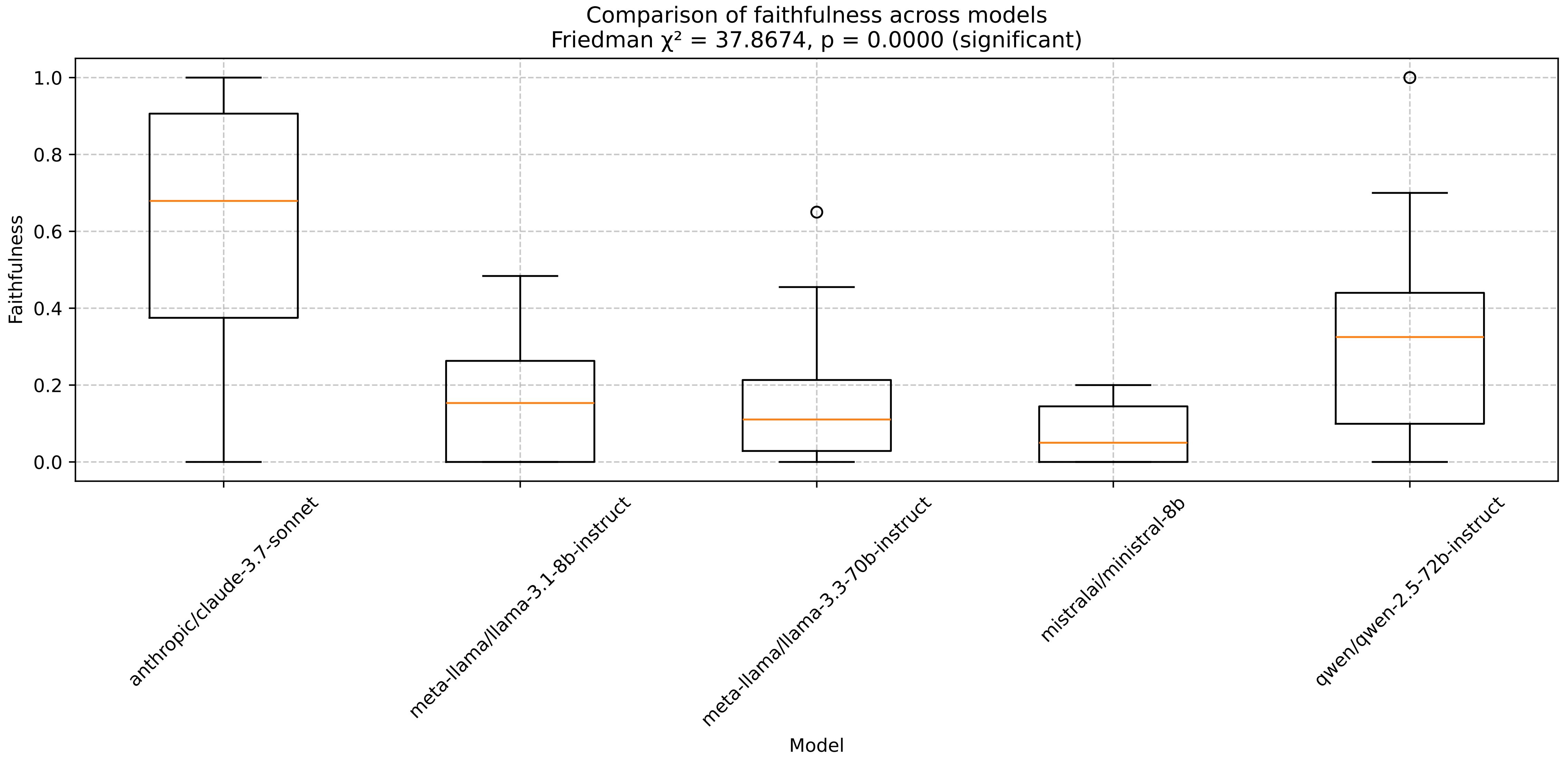
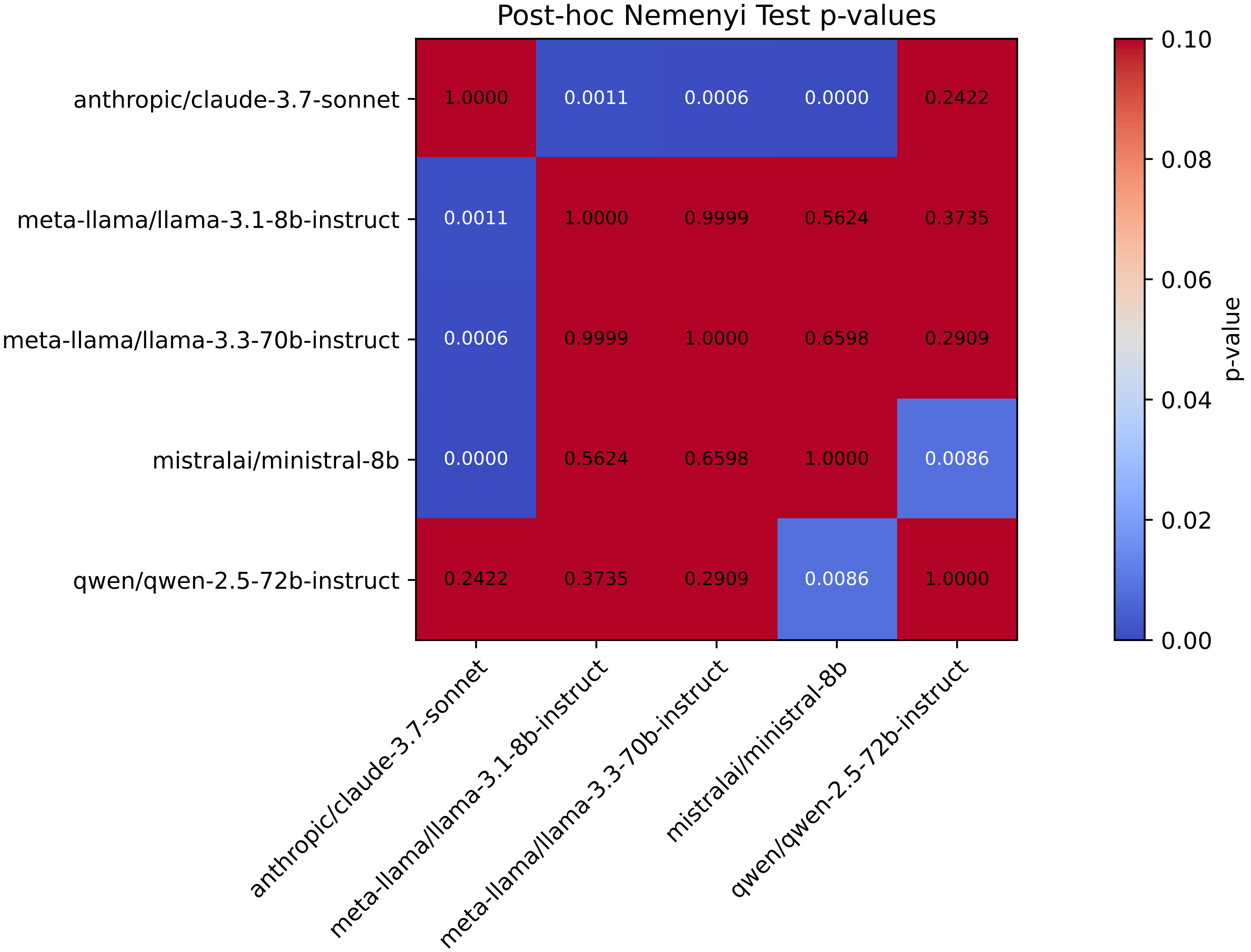

Comparison of faithfulness across models
Friedman χ² = 37.8674, p = 0.0000 (significant)
Faithfulness
1.0
0.8
0.6
0.4
0.2
0.0
anthropic/claude-3.7-sonnet
meta-llama/llama-3.1-8b-instruct
meta-llama/llama-3.3-70b-instruct
mistralai/ministral-8b
qwen/qwen-2.5-72b-instruct
Model



Post-hoc Nemenyi Test p-values

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	4.0000	0.0002	✓
pic/claude-3.7-sonnet vs meta-llama/llama-3.3-70	6.0000	0.0005	✓
anthropic/claude-3.7-sonnet vs mistralai/ministral	0.0000	0.0002	✓
ropic/claude-3.7-sonnet vs qwen/qwen-2.5-72b-ir	15.0000	0.0021	✓
na/llama-3.1-8b-instruct vs meta-llama/llama-3.3-	73.0000	0.5861	
ta-llama/llama-3.1-8b-instruct vs mistralai/minist	17.0000	0.0083	✓
lama/llama-3.1-8b-instruct vs qwen/qwen-2.5-72b	25.0000	0.0048	✓
ta-llama/llama-3.3-70b-instruct vs mistralai/minist	28.0000	0.0385	✓
ama/llama-3.3-70b-instruct vs qwen/qwen-2.5-72	11.0000	0.0019	✓
nistralai/ministral-8b vs qwen/qwen-2.5-72b-instr	8.0000	0.0007	✓