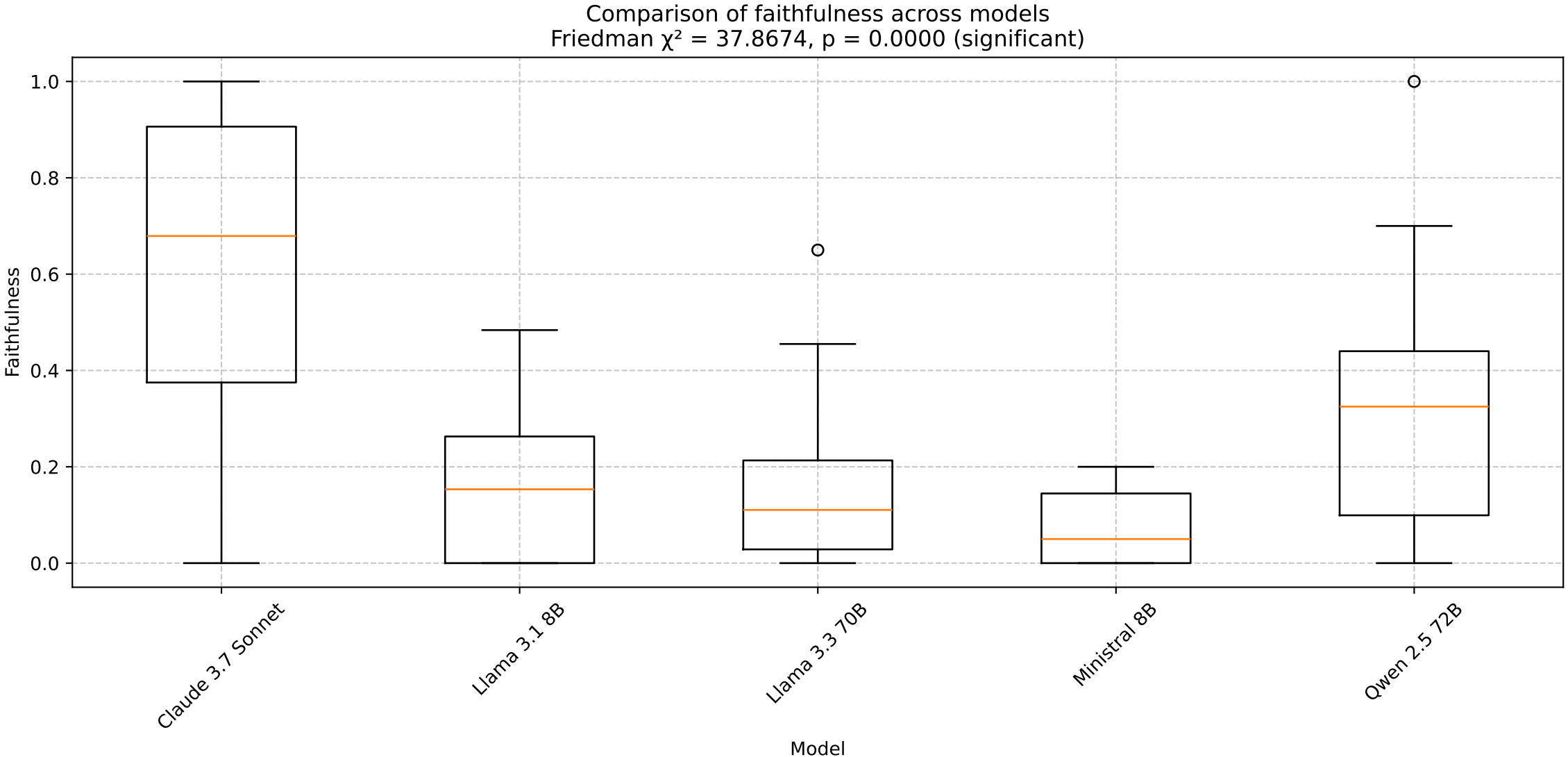
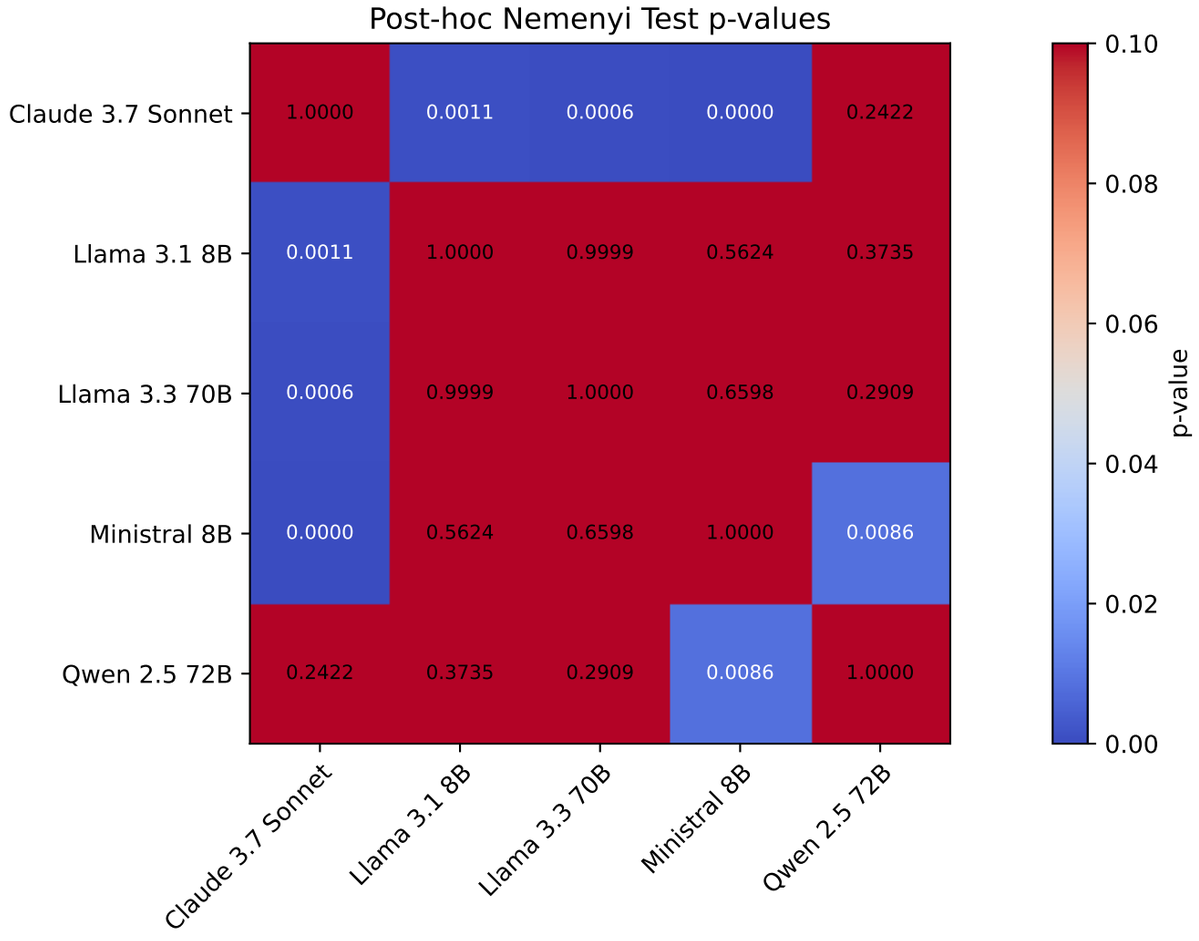

Comparison of faithfulness across models
Friedman χ² = 37.8674, p = 0.0000 (significant)
Faithfulness
1.0
0.8
0.6
0.4
0.2
0.0
Claude 3.7 Sonnet
Llama 3.1 8B
Llama 3.3 70B
Ministral 8B
Qwen 2.5 72B
Model



Pairwise Wilcoxon Signed-Rank Tests

Comparison	Statistic	p-value	Significant (p<0.05)
Claude 3.7 Sonnet vs Llama 3.1 8B	4.0000	0.0002	✓
Claude 3.7 Sonnet vs Llama 3.3 70B	6.0000	0.0005	✓
Claude 3.7 Sonnet vs Ministral 8B	0.0000	0.0002	✓
Claude 3.7 Sonnet vs Qwen 2.5 72B	15.0000	0.0021	✓
Llama 3.1 8B vs Llama 3.3 70B	73.0000	0.5861	
Llama 3.1 8B vs Ministral 8B	17.0000	0.0083	✓
Llama 3.1 8B vs Qwen 2.5 72B	25.0000	0.0048	✓
Llama 3.3 70B vs Ministral 8B	28.0000	0.0385	✓
Llama 3.3 70B vs Qwen 2.5 72B	11.0000	0.0019	✓
Ministral 8B vs Qwen 2.5 72B	8.0000	0.0007	✓