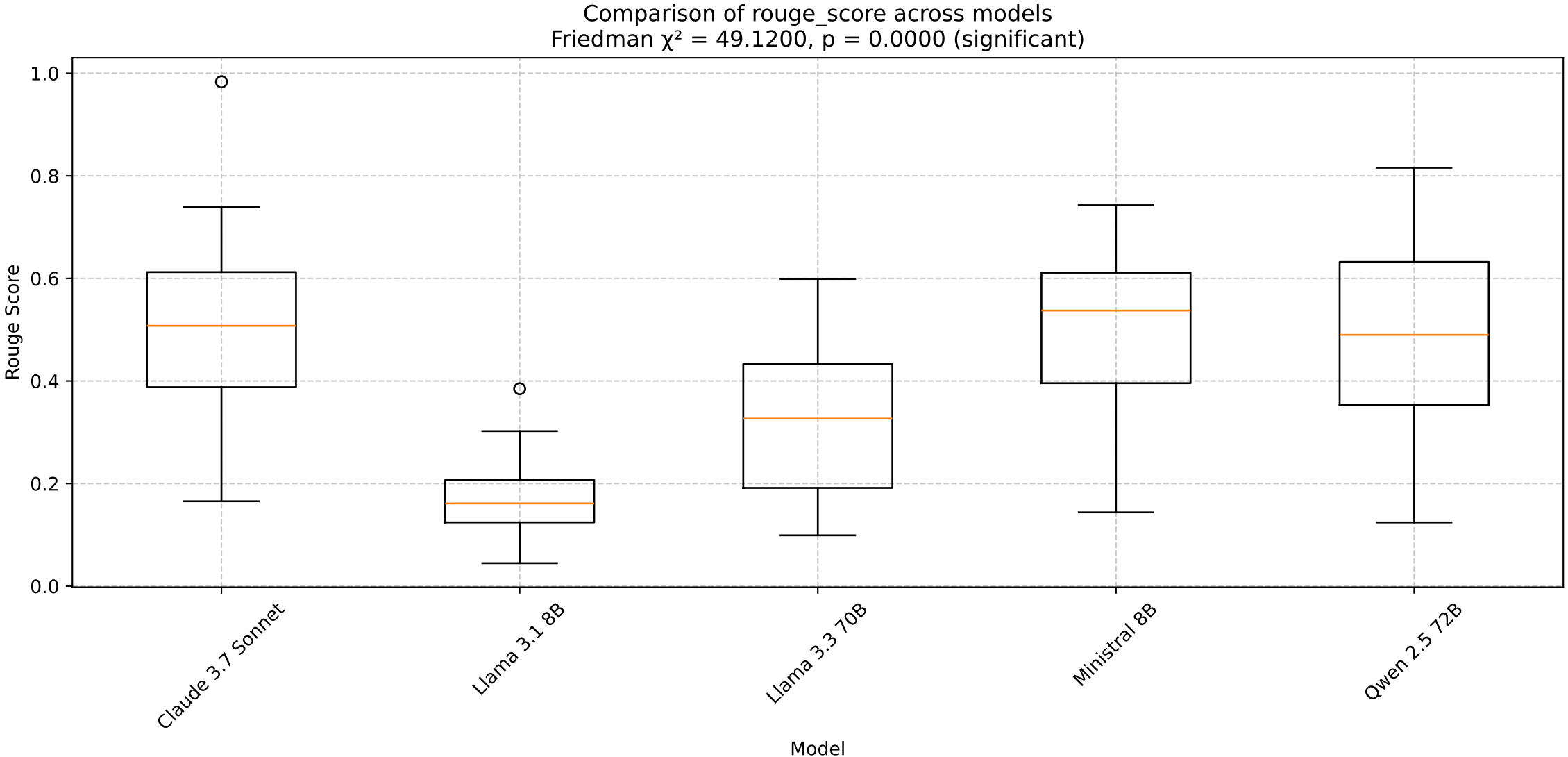
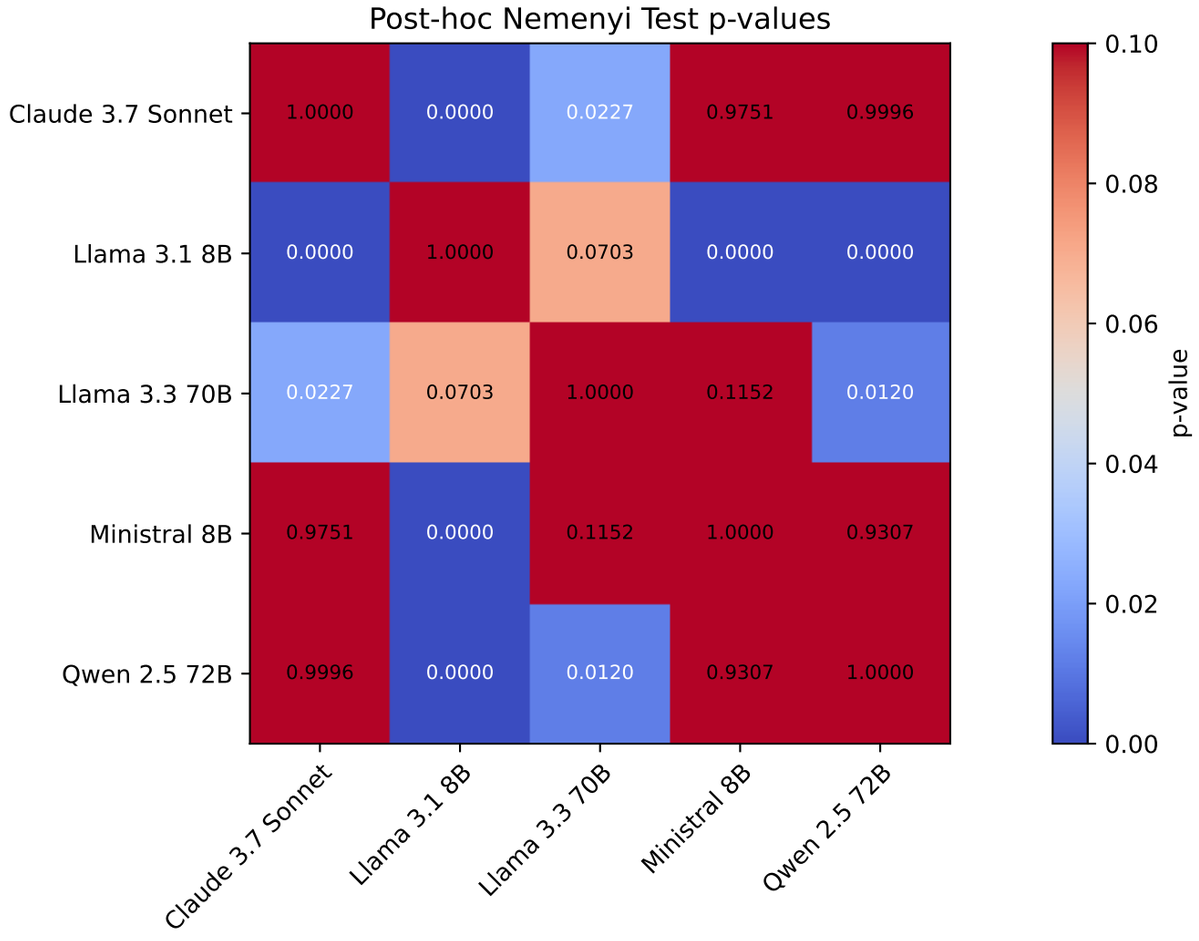

Comparison of rouge_score across models
Friedman χ² = 49.1200, p = 0.0000 (significant)
Rouge Score
1.0
0.8
0.6
0.4
0.2
0.0
Claude 3.7 Sonnet
Llama 3.1 8B
Llama 3.3 70B
Ministral 8B
Qwen 2.5 72B
Model



Pairwise Wilcoxon Signed-Rank Tests

Comparison	Statistic	p-value	Significant (p<0.05)
Claude 3.7 Sonnet vs Llama 3.1 8B	0.0000	0.0000	✓
Claude 3.7 Sonnet vs Llama 3.3 70B	14.0000	0.0002	✓
Claude 3.7 Sonnet vs Ministral 8B	92.0000	0.6477	
Claude 3.7 Sonnet vs Qwen 2.5 72B	92.0000	0.6477	
Llama 3.1 8B vs Llama 3.3 70B	18.0000	0.0005	✓
Llama 3.1 8B vs Ministral 8B	0.0000	0.0000	✓
Llama 3.1 8B vs Qwen 2.5 72B	0.0000	0.0000	✓
Llama 3.3 70B vs Ministral 8B	11.0000	0.0001	✓
Llama 3.3 70B vs Qwen 2.5 72B	20.0000	0.0007	✓
Ministral 8B vs Qwen 2.5 72B	85.0000	0.4749	