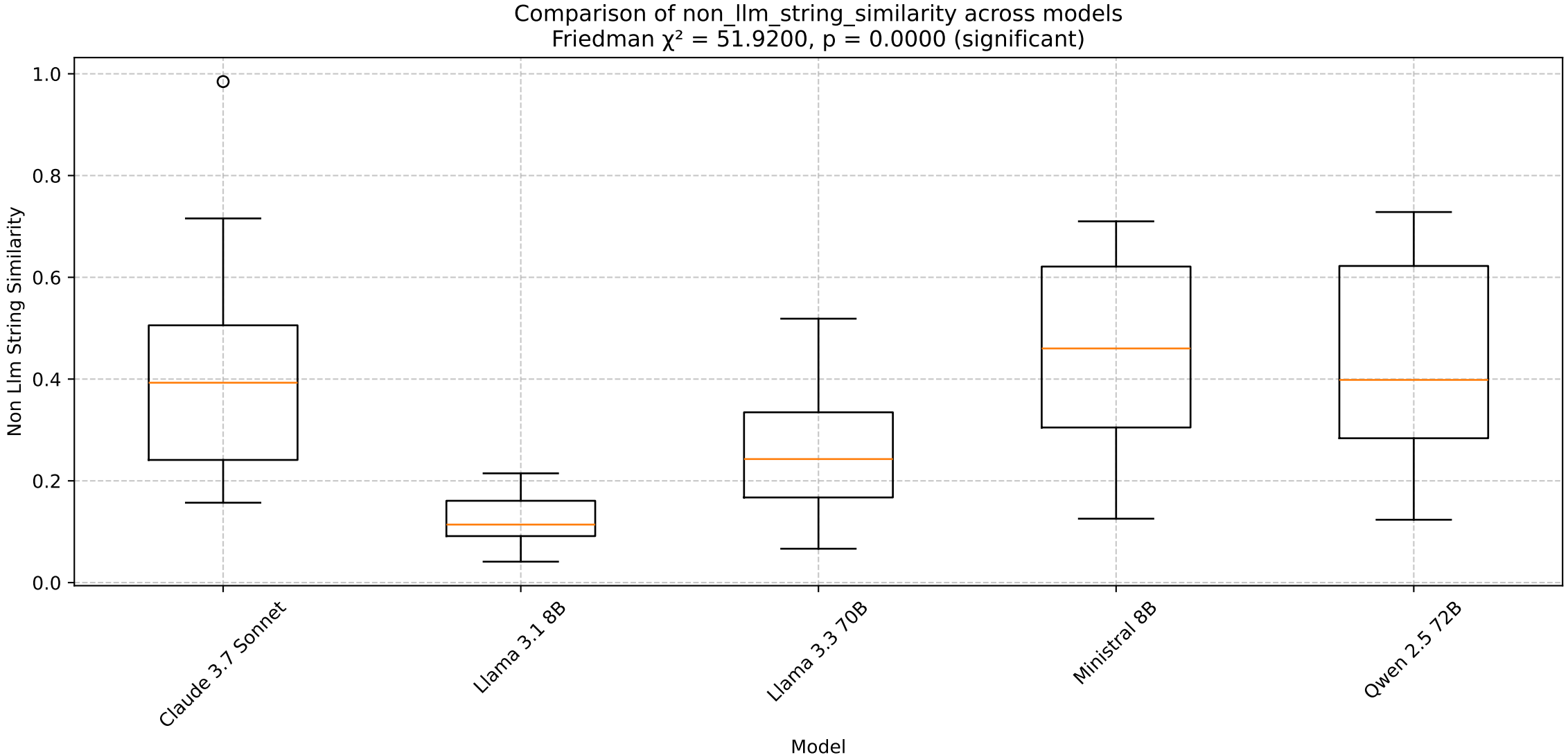
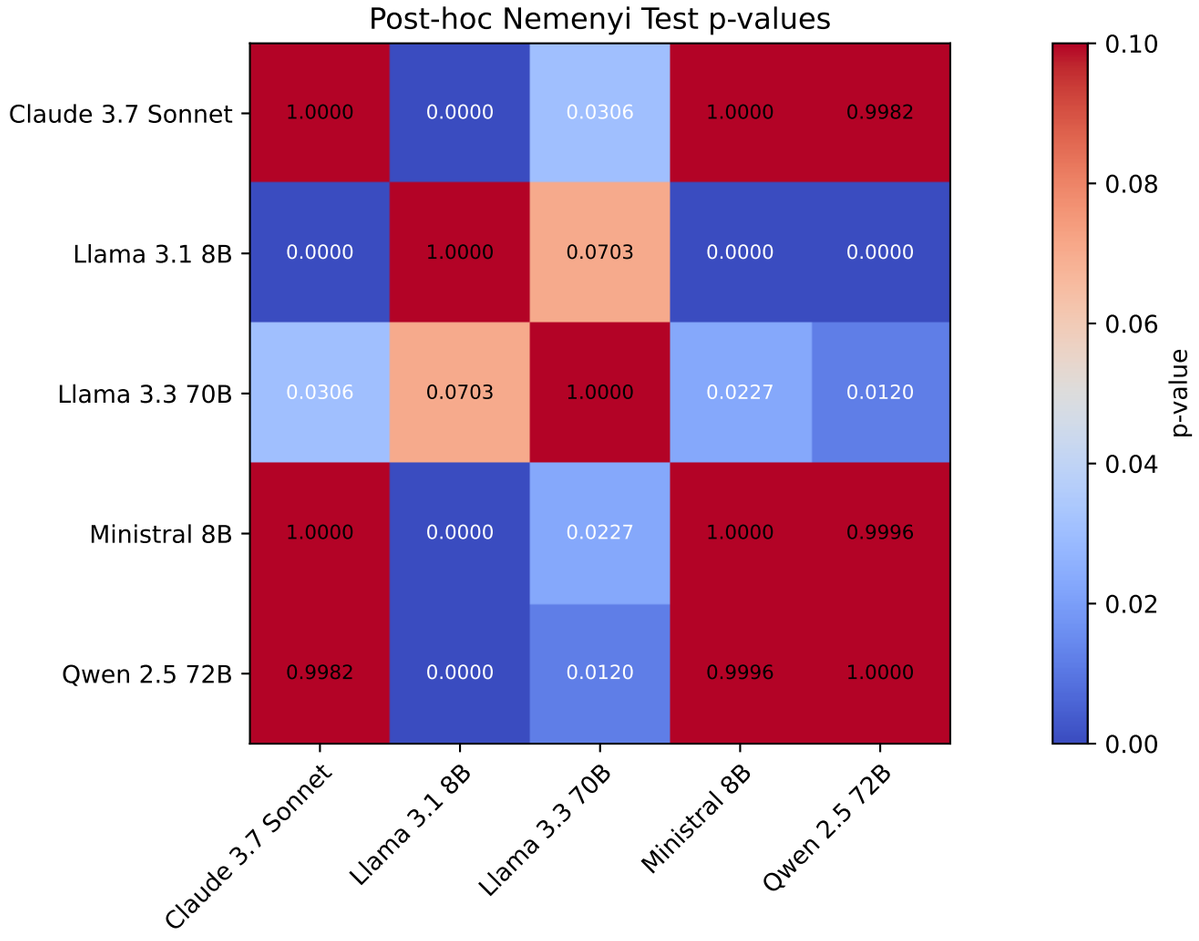

Comparison of non_llm_string_similarity across models
Friedman χ² = 51.9200, p = 0.0000 (significant)
Non Llm String Similarity
1.0
0.8
0.6
0.4
0.2
0.0
Claude 3.7 Sonnet
Llama 3.1 8B
Llama 3.3 70B
Ministral 8B
Qwen 2.5 72B
Model



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	12.0000	0.0001	✓
Claude 3.7 Sonnet vs Ministral 8B	82.0000	0.4091	
Claude 3.7 Sonnet vs Qwen 2.5 72B	93.0000	0.6742	
Llama 3.1 8B vs Llama 3.3 70B	11.0000	0.0001	✓
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	3.0000	0.0000	✓
Llama 3.3 70B vs Qwen 2.5 72B	14.0000	0.0002	✓
Ministral 8B vs Qwen 2.5 72B	96.0000	0.7562	