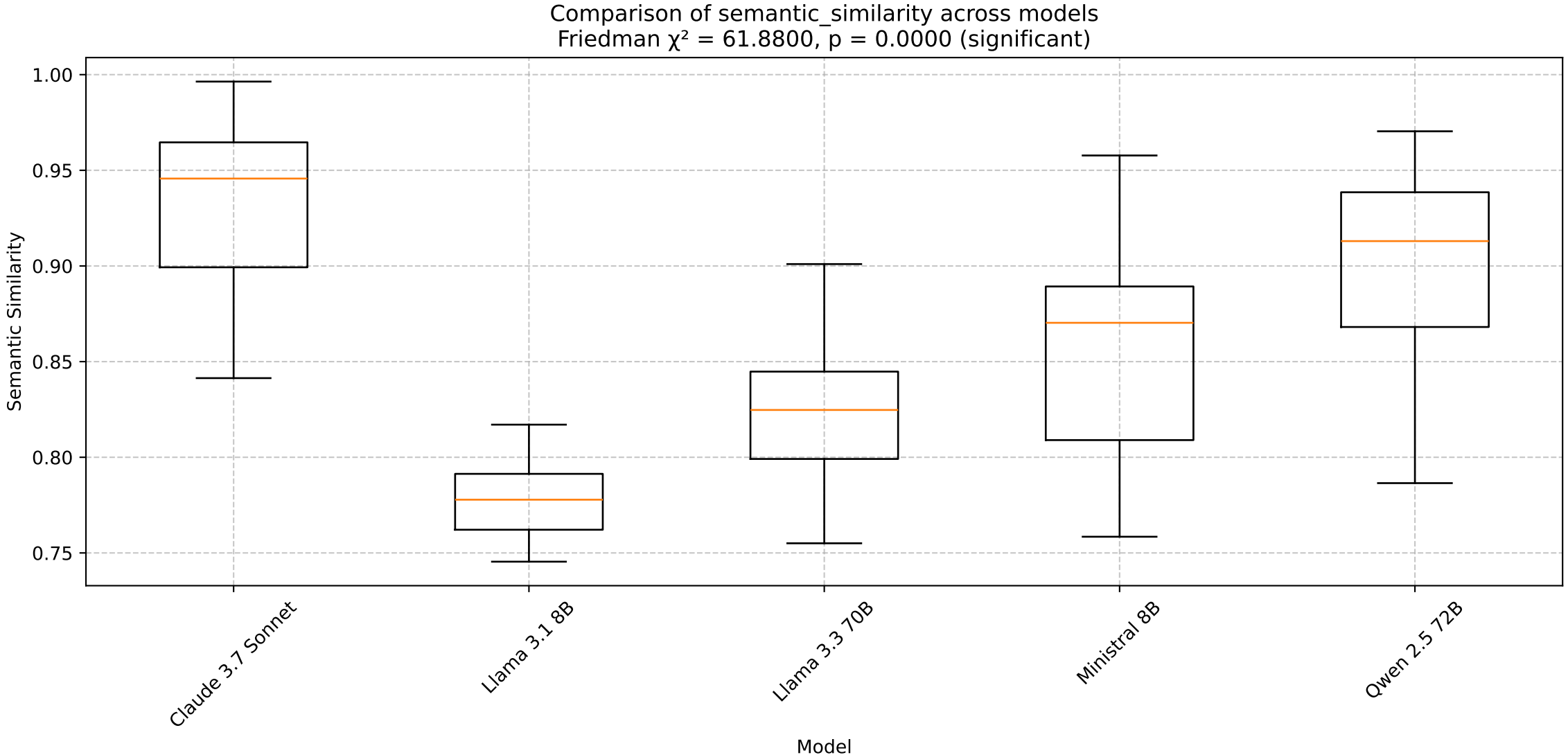
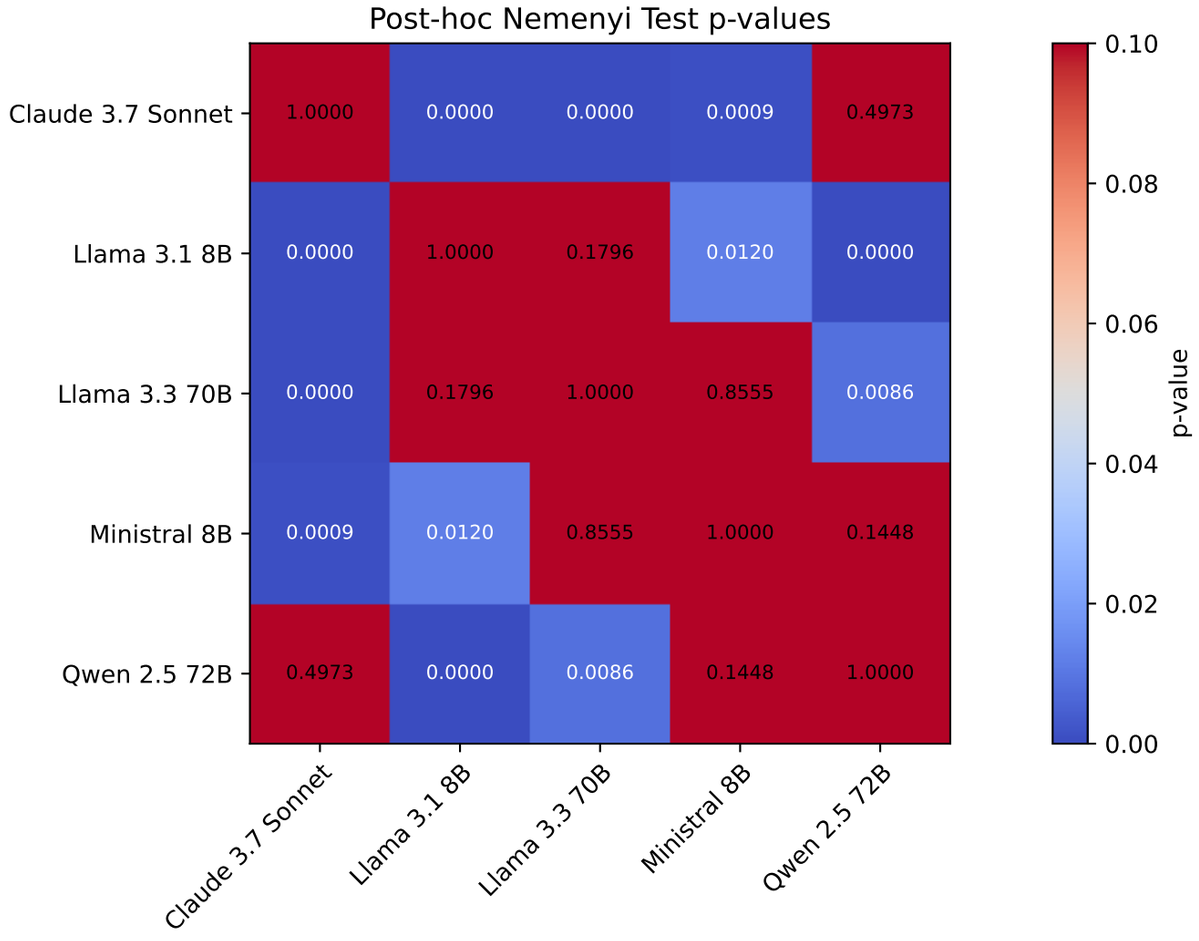

Comparison of semantic_similarity across models
Friedman χ² = 61.8800, p = 0.0000 (significant)
Semantic Similarity
1.00
0.95
0.90
0.85
0.80
0.75
Claude 3.7 Sonnet
Llama 3.1 8B
Llama 3.3 70B
Ministral 8B
Qwen 2.5 72B
Model



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	0.0000	0.0000	✓
Claude 3.7 Sonnet vs Ministral 8B	1.0000	0.0000	✓
Claude 3.7 Sonnet vs Qwen 2.5 72B	19.0000	0.0006	✓
Llama 3.1 8B vs Llama 3.3 70B	10.0000	0.0001	✓
Llama 3.1 8B vs Ministral 8B	5.0000	0.0000	✓
Llama 3.1 8B vs Qwen 2.5 72B	0.0000	0.0000	✓
Llama 3.3 70B vs Ministral 8B	60.0000	0.0973	
Llama 3.3 70B vs Qwen 2.5 72B	2.0000	0.0000	✓
Ministral 8B vs Qwen 2.5 72B	33.0000	0.0056	✓