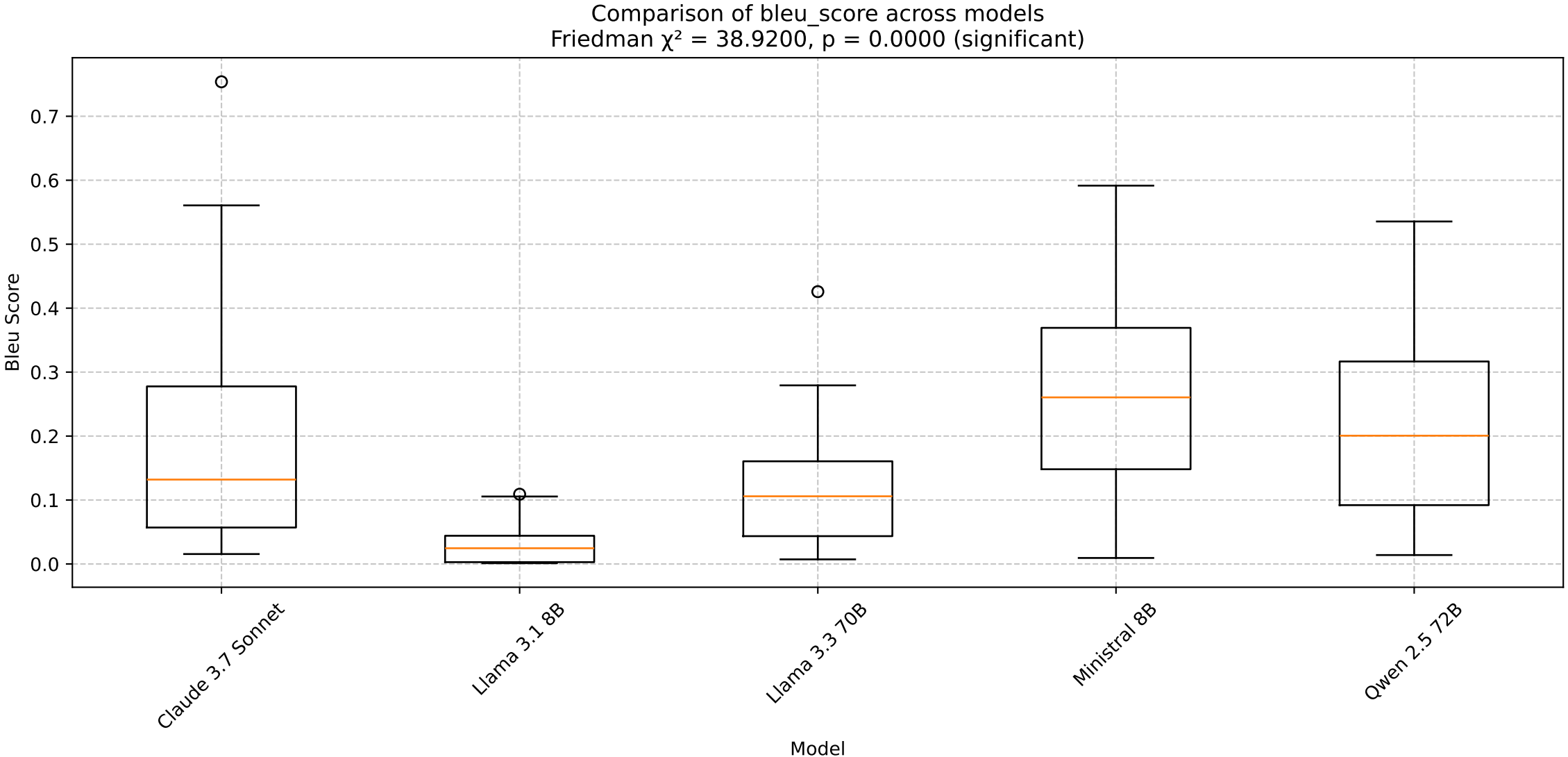
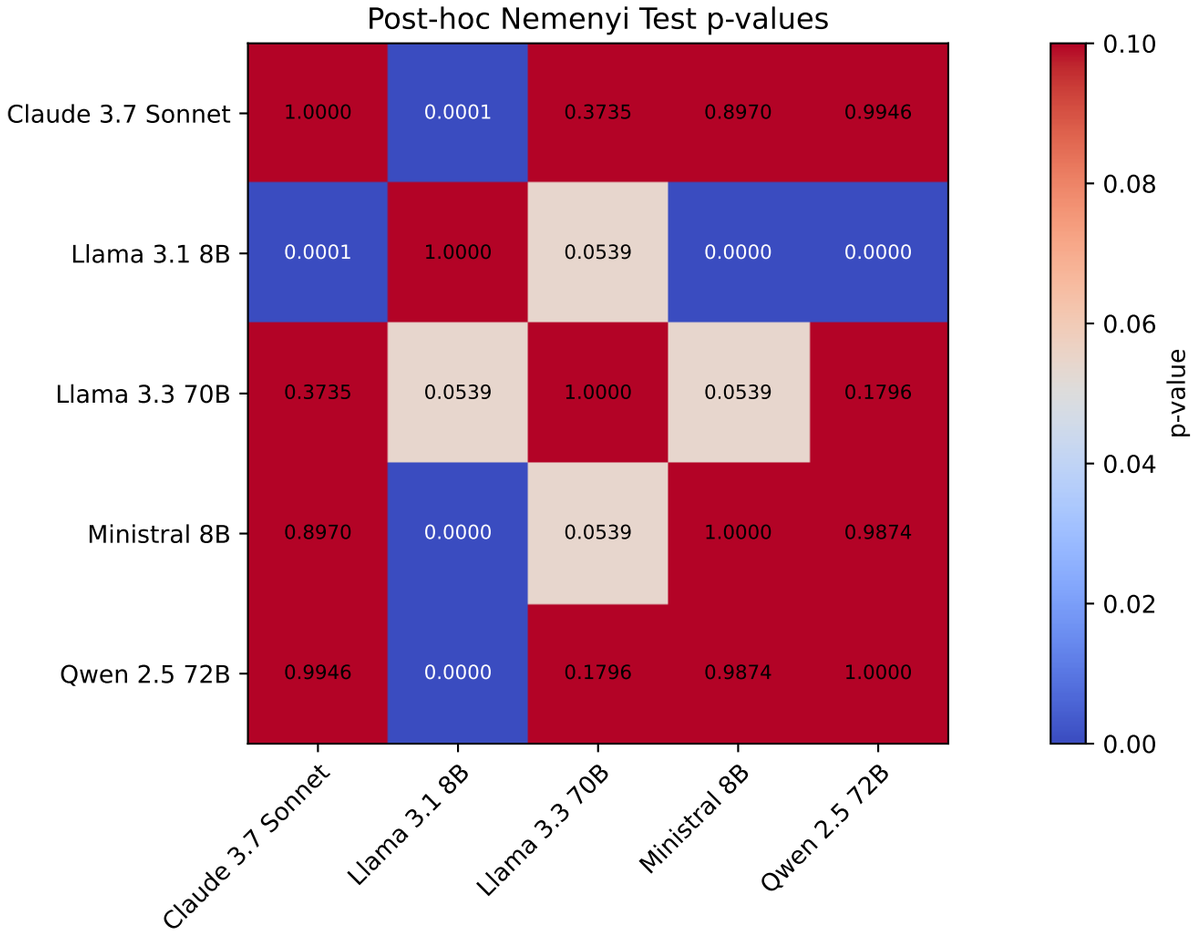

Comparison of bleu_score across models
Friedman χ² = 38.9200, p = 0.0000 (significant)
Bleu Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Claude 3.7 Sonnet
Llama 3.1 8B
Llama 3.3 70B
Ministral 8B
Qwen 2.5 72B
Model



Pairwise Wilcoxon Signed-Rank Tests

Comparison	Statistic	p-value	Significant (p<0.05)
Claude 3.7 Sonnet vs Llama 3.1 8B	5.0000	0.0000	✓
Claude 3.7 Sonnet vs Llama 3.3 70B	67.0000	0.1650	
Claude 3.7 Sonnet vs Ministral 8B	64.0000	0.1327	
Claude 3.7 Sonnet vs Qwen 2.5 72B	102.0000	0.9273	
Llama 3.1 8B vs Llama 3.3 70B	11.0000	0.0001	✓
Llama 3.1 8B vs Ministral 8B	2.0000	0.0000	✓
Llama 3.1 8B vs Qwen 2.5 72B	2.0000	0.0000	✓
Llama 3.3 70B vs Ministral 8B	19.0000	0.0006	✓
Llama 3.3 70B vs Qwen 2.5 72B	31.0000	0.0042	✓
Ministral 8B vs Qwen 2.5 72B	90.0000	0.5958	