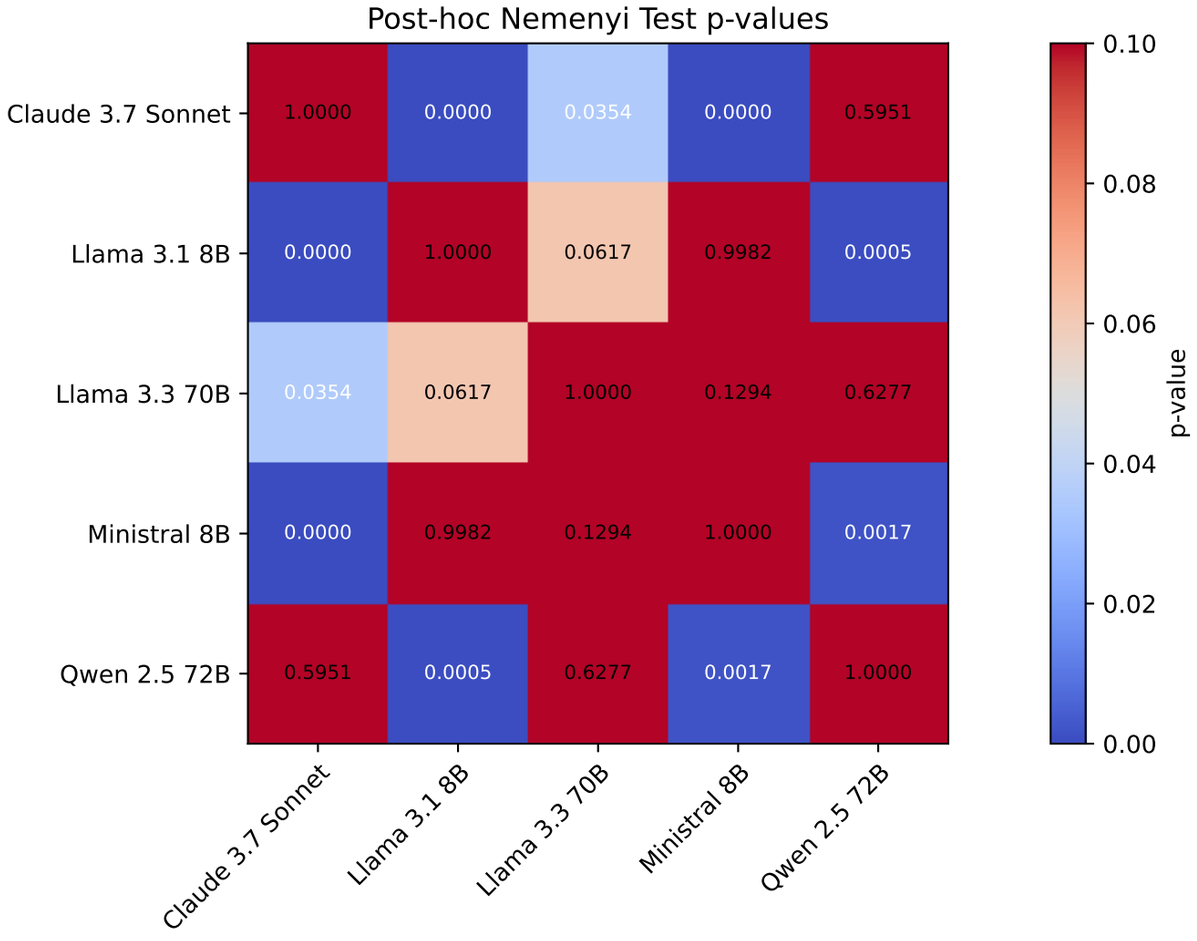

Pairwise Wilcoxon Signed-Rank Tests

Comparison	Statistic	p-value	Significant (p<0.05)
Claude 3.7 Sonnet vs Llama 3.1 8B	2.0000	0.0000	✓
Claude 3.7 Sonnet vs Llama 3.3 70B	6.0000	0.0000	✓
Claude 3.7 Sonnet vs Ministral 8B	2.0000	0.0000	✓
Claude 3.7 Sonnet vs Qwen 2.5 72B	25.0000	0.0017	✓
Llama 3.1 8B vs Llama 3.3 70B	22.0000	0.0010	✓
Llama 3.1 8B vs Ministral 8B	98.0000	0.8124	
Llama 3.1 8B vs Qwen 2.5 72B	13.0000	0.0002	✓
Llama 3.3 70B vs Ministral 8B	22.0000	0.0010	✓
Llama 3.3 70B vs Qwen 2.5 72B	37.0000	0.0196	
Ministral 8B vs Qwen 2.5 72B	13.0000	0.0002	✓