

Model Pair	Nemenyi p-value	Nemenyi Sig (p<0.05)	Wilcoxon p-value	Wilcoxon Sig (p<0.05)
Claude 3.7 Sonnet vs Llama 3.1 8B	0.0000	✓	0.0001	✓
Claude 3.7 Sonnet vs Llama 3.3 70B	0.0354	✓	0.0003	✓
Claude 3.7 Sonnet vs Ministral 8B	0.0000	✓	0.0001	✓
Claude 3.7 Sonnet vs Qwen 2.5 72B	0.5951		0.0169	✓
Llama 3.1 8B vs Llama 3.3 70B	0.0617		0.0102	✓
Llama 3.1 8B vs Ministral 8B	0.9982		1.0000	
Llama 3.1 8B vs Qwen 2.5 72B	0.0005	✓	0.0017	✓
Llama 3.3 70B vs Ministral 8B	0.1294		0.0102	✓
Llama 3.3 70B vs Qwen 2.5 72B	0.6277		0.1959	
Ministral 8B vs Qwen 2.5 72B	0.0017	✓	0.0017	✓