| 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 | 0.0011 | ✓ | 0.0025 | ✓ |
| Claude 3.7 Sonnet vs Llama 3.3 70B | 0.0006 | ✓ | 0.0054 | ✓ |
| Claude 3.7 Sonnet vs Ministral 8B | 0.0000 | ✓ | 0.0020 | ✓ |
| Claude 3.7 Sonnet vs Qwen 2.5 72B | 0.2422 | | 0.0214 | ✓ |
| Llama 3.1 8B vs Llama 3.3 70B | 0.9999 | | 1.0000 | |
| Llama 3.1 8B vs Ministral 8B | 0.5624 | | 0.0832 | |
| Llama 3.1 8B vs Qwen 2.5 72B | 0.3735 | | 0.0484 | ✓ |
| Llama 3.3 70B vs Ministral 8B | 0.6598 | | 0.3854 | |
| Llama 3.3 70B vs Qwen 2.5 72B | 0.2909 | | 0.0193 | ✓ |
| Ministral 8B vs Qwen 2.5 72B | 0.0086 | ✓ | 0.0074 | ✓ |