



Pairwise Wilcoxon Signed-Rank Tests (Bonferroni Correction)

| Comparison | p-value | Significant (p<0.05) |
|------------------------------------|---------|----------------------|
| Claude 3.7 Sonnet vs Llama 3.1 8B | 0.0000 | ✓ |
| Claude 3.7 Sonnet vs Llama 3.3 70B | 0.0013 | ✓ |
| Claude 3.7 Sonnet vs Ministral 8B | 1.0000 | |
| Claude 3.7 Sonnet vs Qwen 2.5 72B | 1.0000 | |
| Llama 3.1 8B vs Llama 3.3 70B | 0.0010 | ✓ |
| Llama 3.1 8B vs Ministral 8B | 0.0000 | ✓ |
| Llama 3.1 8B vs Qwen 2.5 72B | 0.0000 | ✓ |
| Llama 3.3 70B vs Ministral 8B | 0.0001 | ✓ |
| Llama 3.3 70B vs Qwen 2.5 72B | 0.0021 | ✓ |
| Ministral 8B vs Qwen 2.5 72B | 1.0000 | |