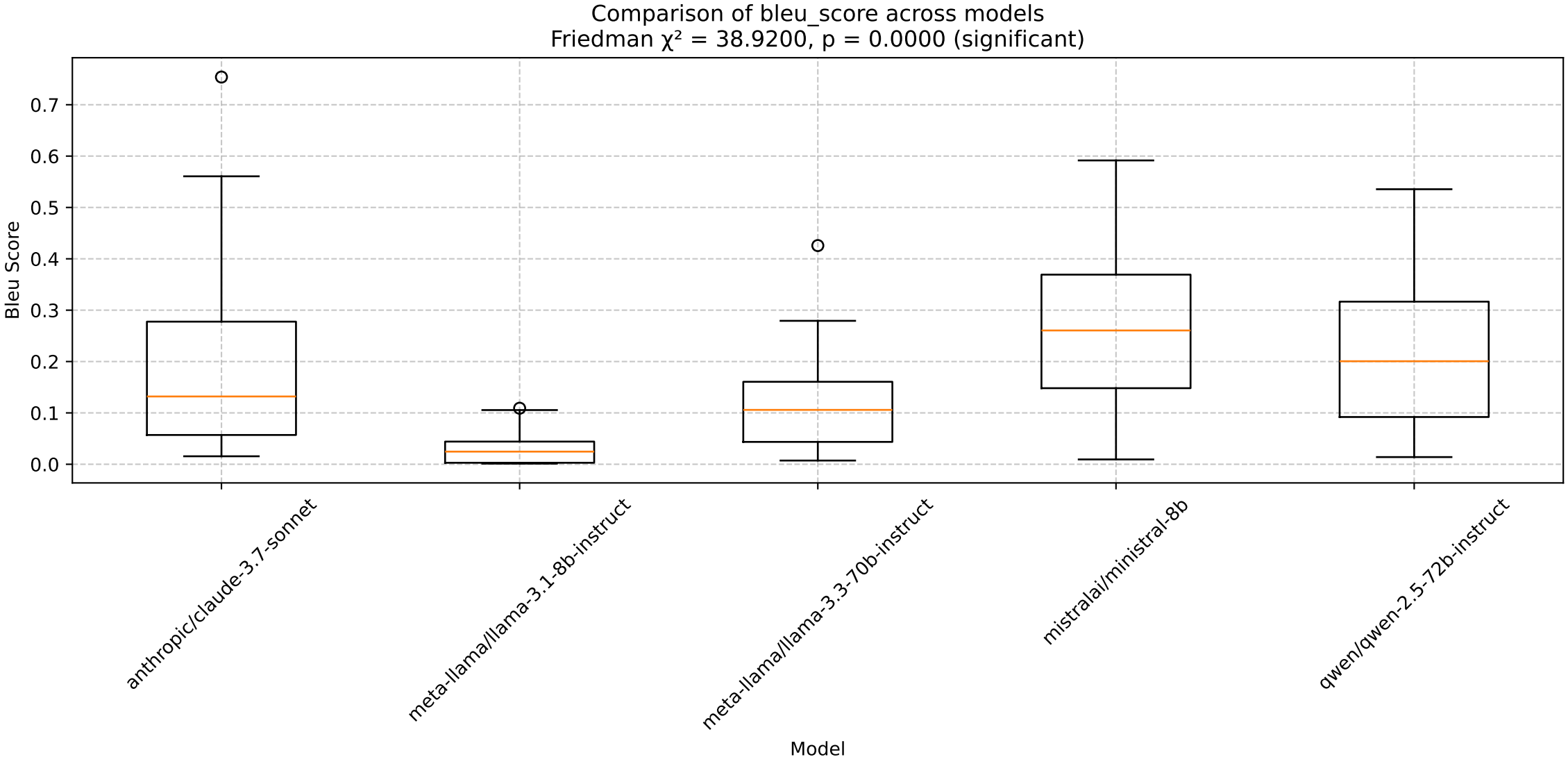
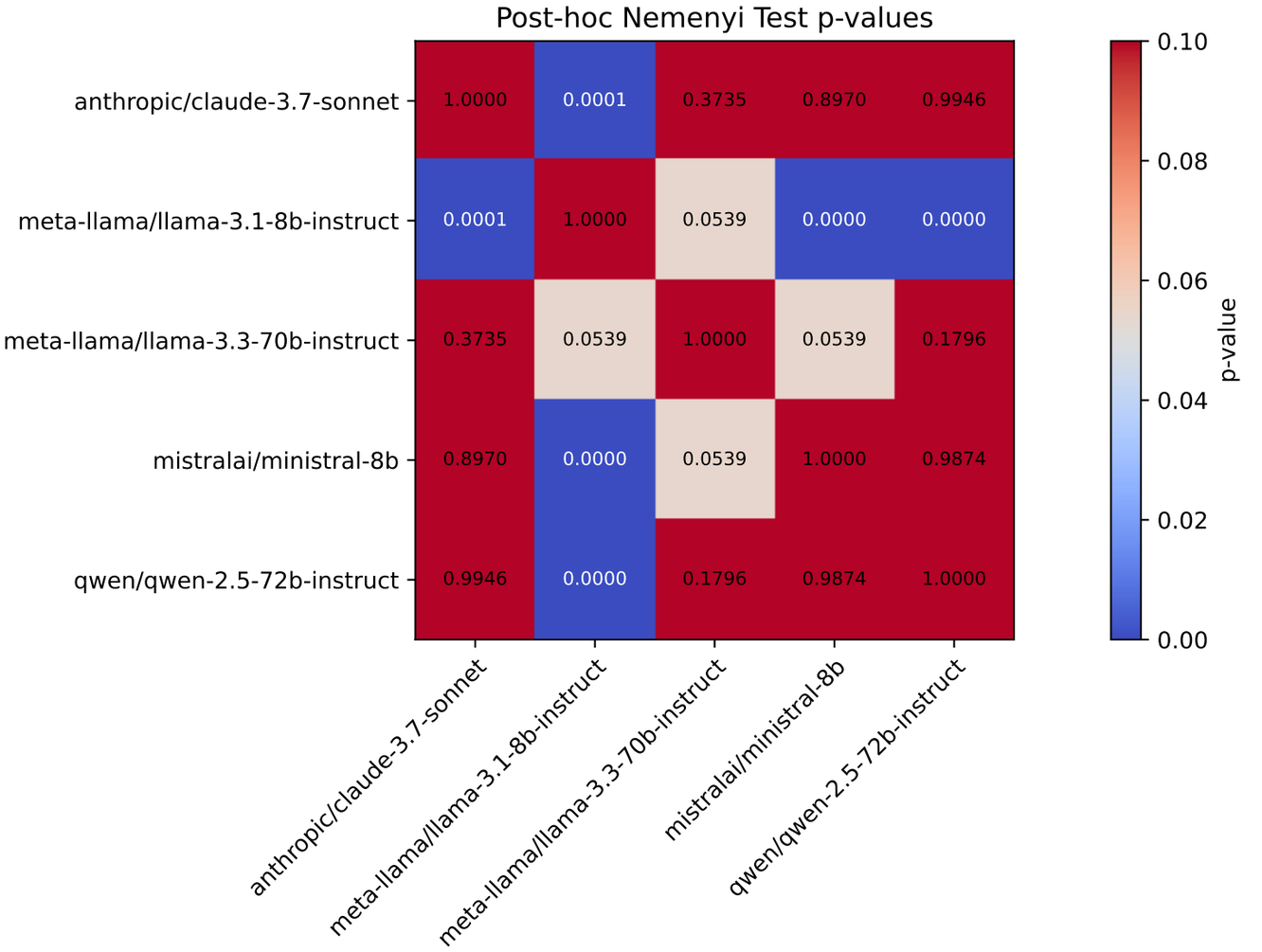

Comparison of bleu_score across models
Friedman χ² = 38.9200, p = 0.0000 (significant)
Bleu Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
anthropic/claude-3.7-sonnet
meta-llama/llama-3.1-8b-instruct
meta-llama/llama-3.3-70b-instruct
mistralai/ministral-8b
qwen/qwen-2.5-72b-instruct
Model




## Pairwise Wilcoxon Signed-Rank Tests

| Comparison | Statistic | p-value | Significant (p<0.05) |
|---|---|---|---|
| pic/claude-3.7-sonnet vs meta-llama/llama-3.1-8b | 5.0000 | 0.0000 | ✓ |
| pic/claude-3.7-sonnet vs meta-llama/llama-3.3-70 | 67.0000 | 0.1650 | |
| anthropic/claude-3.7-sonnet vs mistralai/ministral | 64.0000 | 0.1327 | |
| ropic/claude-3.7-sonnet vs qwen/qwen-2.5-72b-i | 102.0000 | 0.9273 | |
| a/llama-3.1-8b-instruct vs meta-llama/llama-3.3- | 11.0000 | 0.0001 | ✓ |
| ta-llama/llama-3.1-8b-instruct vs mistralai/minist | 2.0000 | 0.0000 | ✓ |
| lama/llama-3.1-8b-instruct vs qwen/qwen-2.5-72b | 2.0000 | 0.0000 | ✓ |
| a-llama/llama-3.3-70b-instruct vs mistralai/minist | 19.0000 | 0.0006 | ✓ |
| ama/llama-3.3-70b-instruct vs qwen/qwen-2.5-72 | 31.0000 | 0.0042 | ✓ |
| mistralai/ministral-8b vs qwen/qwen-2.5-72b-instr | 90.0000 | 0.5958 | |