

Comparison of factual correctness across models
Friedman $\chi^2 = 45.1429$, $p = 0.0000$ (significant)

