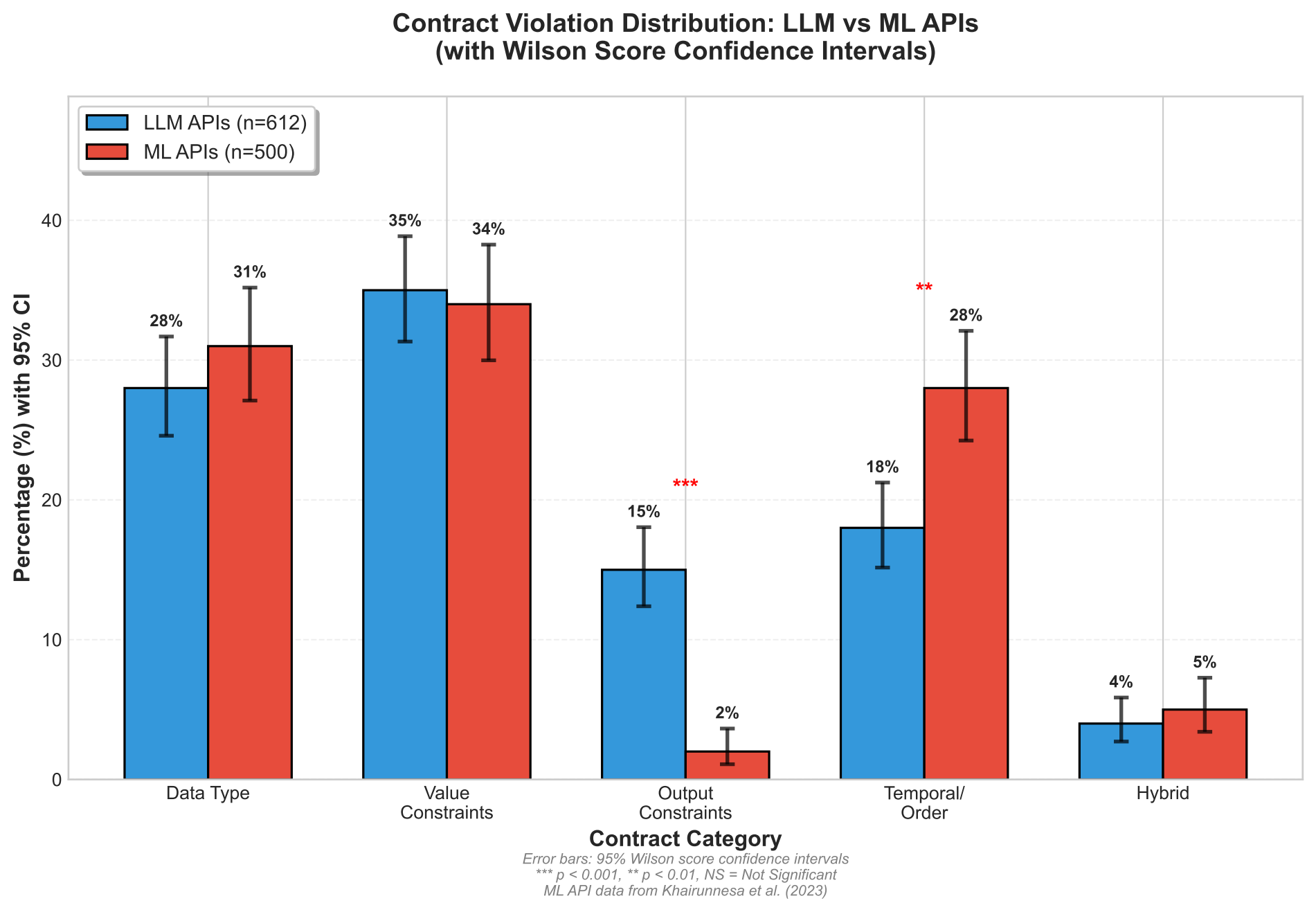
Contract Violation Distribution: LLM vs ML APIs
(with Wilson Score Confidence Intervals)
LLM APIs (n=612)
ML APIs (n=500)
Percentage (%) with 95% CI
0
10
20
30
40
28%
31%
35%
34%
15%
2%
18%
28%
4%
5%
***
**
Data Type
Value Constraints
Output Constraints
Temporal/ Order
Hybrid
Contract Category
Error bars: 95% Wilson score confidence intervals
*** p < 0.001, ** p < 0.01, NS = Not Significant
ML API data from Khairunnesa et al. (2023)