**Performance Degradation Across Benchmark Tiers
(Illustrative Results with 95% CI)**

Legend:
- Premium (Claude, GPT-4o)
- Mid-range (GPT-4o-mini, Gemini Flash)
- Open-source (Llama, Mistral)

*** p<0.001 (bootstrap test, n=1000)

Tier 1 (3-5 turns): Premium 68%, Mid-range 64%, Open-source 58%
Tier 2 (8-12 turns): Premium 61%, Mid-range 55%, Open-source 48%
Tier 3 (20+ turns): Premium 54%, Mid-range 46%, Open-source 38%

Y-axis: Average Performance Score (%)
X-axis: Benchmark Tier