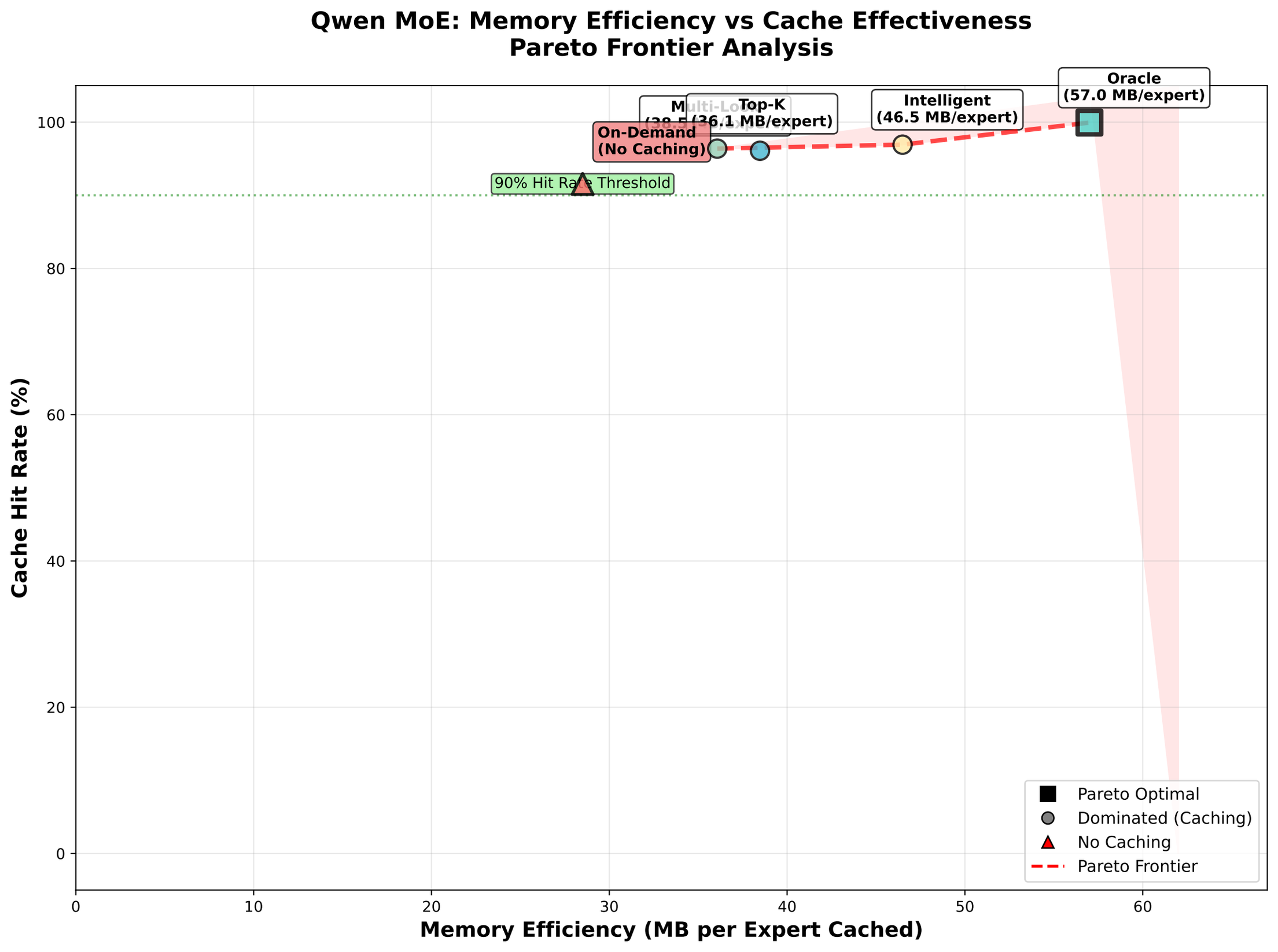
Qwen MoE: Memory Efficiency vs Cache Effectiveness
Pareto Frontier Analysis
Oracle
(57.0 MB/expert)
Intelligent
(46.5 MB/expert)
Top-K
(36.1 MB/expert)
On-Demand
(No Caching)
90% Hit Rate Threshold
Cache Hit Rate (%)
Memory Efficiency (MB per Expert Cached)
0
20
40
60
80
100
0
10
20
30
40
50
60
Pareto Optimal
Dominated (Caching)
No Caching
Pareto Frontier