Clarifier Agent Test Suite

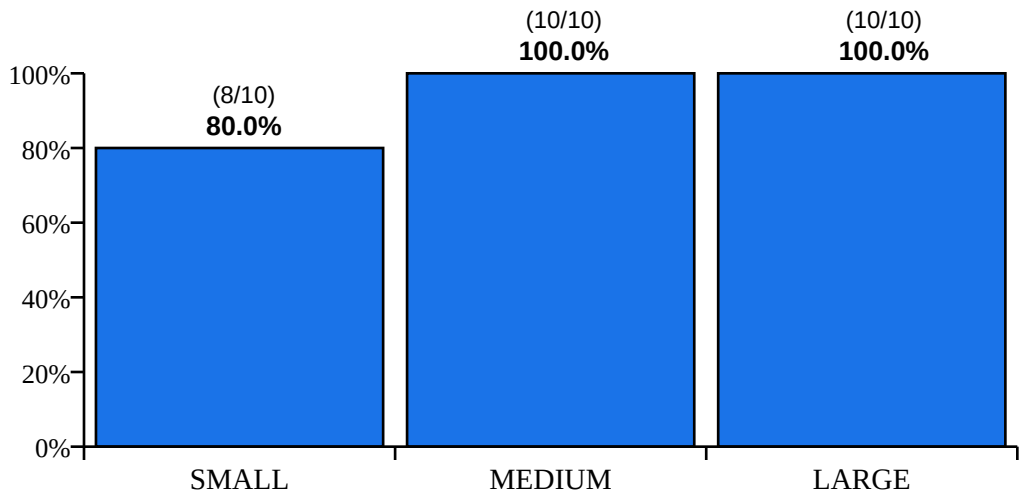
Multi-Model Performance Analysis

Agent: clarifier_multi_model
Version: 1.0.0
Generated: 2025-08-25 00:11:56

Executive Summary

Model	Tests	Passed	Success Rate	Avg Latency	Total Cost
SMALL	10	8/10	80.0%	1.67s	$0.0035
MEDIUM	10	10/10	100.0%	1.63s	$0.0097
LARGE	10	10/10	100.0%	1.66s	$0.0344
OVERALL	**30**	**28/30**	**93.3%**	**1.65s**	**$0.0476**

Success Rate by Model



Detailed Test Results

Scenario 1: Clear Bank and Period - RBC Q3 2024

Query: Show me RBC's Q3 2024 financial results
Expected: Status: success, Banks: [1], Year: 2024, Quarters: ['Q3']

Model	Status	Banks	Period	Latency	Cost
SMALL	✓ Pass	[1]	2024 ['Q3']	2.01s	$0.00035
MEDIUM	✓ Pass	[1]	2024 ['Q3']	1.50s	$0.00101
LARGE	✓ Pass	[1]	2024 ['Q3']	1.65s	$0.00336

Scenario 2: Multiple Banks - Big Six Q2 2024

Query: Compare the Big Six banks performance in Q2 2024
Expected: Status: success, Banks: [1, 2, 3, 4, 5, 6], Year: 2024, Quarters: ['Q2']

Model	Status	Banks	Period	Latency	Cost
SMALL	✓ Pass	[1, 2, 3, 4, 5, 6]	2024 ['Q2']	2.15s	$0.00048
MEDIUM	✓ Pass	[1, 2, 3, 4, 5, 6]	2024 ['Q2']	1.67s	$0.00137
LARGE	✓ Pass	[1, 2, 3, 4, 5, 6]	2024 ['Q2']	1.73s	$0.00458

Scenario 3: Ambiguous Bank - No Period

Query: What are the latest financial metrics?
Expected: Status: needs_clarification, Needs Clarification

Model	Status	Banks	Period	Latency	Cost
SMALL	✓ Pass	Needs clarification	Needs clarification	1.23s	$0.00030
MEDIUM	✓ Pass	Needs clarification	Needs clarification	1.45s	$0.00090
LARGE	✓ Pass	Needs clarification	Needs clarification	2.35s	$0.00305

Scenario 4: Clear Bank - Missing Period

Query: Show me TD Bank's revenue
Expected: Status: needs_clarification, Banks: [2], Needs Clarification

Model	Status	Banks	Period	Latency	Cost
SMALL	✓ Pass	[2]	Needs clarification	1.23s	$0.00034
MEDIUM	✓ Pass	[2]	Needs clarification	1.41s	$0.00101
LARGE	✓ Pass	[2]	Needs clarification	1.64s	$0.00337

Scenario 5: Latest Period Request

Query: Give me BMO's latest quarterly results
Expected: Status: success, Banks: [3], Year: 2025, Quarters: ['Q3']

Model	Status	Banks	Period	Latency	Cost
SMALL	✗ Fail	[3]	Needs clarification	1.23s	$0.00034

MEDIUM	✓ Pass	[3]	2025 ['Q3']	1.34s	$0.00101
LARGE	✓ Pass	[3]	2025 ['Q3']	1.33s	$0.00336

Scenario 6: YTD Period Request

Query: Show Scotia's YTD 2025 performance
Expected: Status: success, Banks: [4], Year: 2025, Quarters: ['Q1', 'Q2', 'Q3']

Model	Status	Banks	Period	Latency	Cost
SMALL	✗ Fail	[4]	Needs clarification	1.08s	$0.00034
MEDIUM	✓ Pass	[4]	2025 ['Q1', 'Q2', 'Q3']	1.73s	$0.00101
LARGE	✓ Pass	[4]	2025 ['Q1', 'Q2', 'Q3']	1.64s	$0.00338

Scenario 7: Bank Alias - National Bank

Query: What is National Bank's Q1 2024 net income?
Expected: Status: success, Banks: [6], Year: 2024, Quarters: ['Q1']

Model	Status	Banks	Period	Latency	Cost
SMALL	✓ Pass	[6]	2024 ['Q1']	1.48s	$0.00035
MEDIUM	✓ Pass	[6]	2024 ['Q1']	2.00s	$0.00101
LARGE	✓ Pass	[6]	2024 ['Q1']	1.43s	$0.00336

Scenario 8: Multiple Specific Banks

Query: Compare RBC and TD performance in Q4 2023
Expected: Status: success, Banks: [1, 2], Year: 2023, Quarters: ['Q4']

Model	Status	Banks	Period	Latency	Cost
SMALL	✓ Pass	[1, 2]	2023 ['Q4']	2.72s	$0.00038
MEDIUM	✓ Pass	[1, 2]	2023 ['Q4']	2.12s	$0.00043
LARGE	✓ Pass	[1, 2]	2023 ['Q4']	1.74s	$0.00361

Scenario 9: Full Year Period

Query: Show CIBC's full year 2023 results
Expected: Status: success, Banks: [5], Year: 2023, Quarters: ['Q1', 'Q2', 'Q3', 'Q4']

Model	Status	Banks	Period	Latency	Cost
SMALL	✓ Pass	[5]	2023 ['Q1', 'Q2', 'Q3', 'Q4']	1.68s	$0.00035
MEDIUM	✓ Pass	[5]	2023 ['Q1', 'Q2', 'Q3', 'Q4']	1.51s	$0.00101
LARGE	✓ Pass	[5]	2023 ['Q1', 'Q2', 'Q3', 'Q4']	1.55s	$0.00338

Scenario 10: Clear Period - Ambiguous Bank

Query: What was the efficiency ratio in Q2 2024?
Expected: Status: needs_clarification, Needs Clarification

Model	Status	Banks	Period	Latency	Cost
SMALL	✓ Pass	Needs clarification	Needs clarification	1.84s	$0.00031
MEDIUM	✓ Pass	Needs clarification	Needs clarification	1.54s	$0.00090
LARGE	✓ Pass	Needs clarification	Needs clarification	1.54s	$0.00300

Failed Test Analysis

Latest Period Request (SMALL)

Error: Expected status 'success', got 'needs_clarification'
Mismatch Details:
Expected: Banks=[3], Year=2025, Quarters=['Q3'], Status=success
Actual: Status=needs_clarification, Banks=[3]

YTD Period Request (SMALL)

Error: Expected status 'success', got 'needs_clarification'
Mismatch Details:
Expected: Banks=[4], Year=2025, Quarters=['Q1', 'Q2', 'Q3'], Status=success
Actual: Status=needs_clarification, Banks=[4]