Clarifier Agent Test Suite

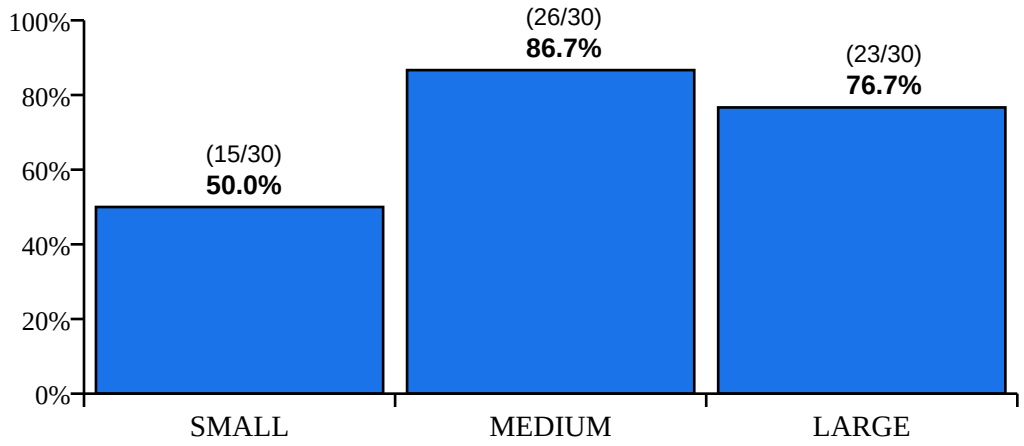
Multi-Model Performance Analysis

Agent: clarifier_multi_model
Version: 1.0.0
Generated: 2025-08-25 13:26:36

Executive Summary

Model	Tests	Passed	Success Rate	Avg Latency	Total Cost
SMALL	30	15/30	50.0%	1.84s	$0.0119
MEDIUM	30	26/30	86.7%	2.31s	$0.0341
LARGE	30	23/30	76.7%	6.44s	$0.1282
OVERALL	**90**	**64/90**	**71.1%**	**3.53s**	**$0.1742**

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']	1.60s	$0.00042
MEDIUM	✓ Pass	[1]	2024 ['Q3']	2.69s	$0.00125
LARGE	✓ Pass	[1]	2024 ['Q3']	2.46s	$0.00416

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']	2.58s	$0.00055
MEDIUM	✓ Pass	[1, 2, 3, 4, 5, 6]	2024 ['Q2']	2.05s	$0.00161
LARGE	✓ Pass	[1, 2, 3, 4, 5, 6]	2024 ['Q2']	2.23s	$0.00537

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	1.67s	$0.00038
MEDIUM	✓ Pass	Needs clarification	Needs clarification	2.50s	$0.00115
LARGE	✓ Pass	Needs clarification	Needs clarification	2.53s	$0.00383

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	1.53s	$0.00043
MEDIUM	✓ Pass	[2]	Needs clarification	2.72s	$0.00125
LARGE	✓ Pass	[2]	Needs clarification	1.96s	$0.00415

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	1.64s	$0.00042

MEDIUM	✓ Pass	[3]	2025 ['Q3']	5.81s	$0.00124
LARGE	✓ Pass	[3]	2025 ['Q3']	2.97s	$0.00415

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	1.25s	$0.00042
MEDIUM	✓ Pass	[4]	2025 ['Q1', 'Q2', 'Q3']	1.74s	$0.00125
LARGE	✓ Pass	[4]	2025 ['Q1', 'Q2', 'Q3']	2.39s	$0.00417

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	✗ Fail	[6]	Needs clarification	2.28s	$0.00042
MEDIUM	✓ Pass	[6]	2024 ['Q1']	1.58s	$0.00125
LARGE	✓ Pass	[6]	2024 ['Q1']	3.42s	$0.00416

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']	1.86s	$0.00017
MEDIUM	✓ Pass	[1, 2]	2023 ['Q4']	2.20s	$0.00049
LARGE	✓ Pass	[1, 2]	2023 ['Q4']	2.62s	$0.00441

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	✗ Fail	[5]	Needs clarification	2.96s	$0.00043
MEDIUM	✓ Pass	[5]	2023 ['Q1', 'Q2', 'Q3', 'Q4']	2.02s	$0.00125
LARGE	✓ Pass	[5]	2023 ['Q1', 'Q2', 'Q3', 'Q4']	1.95s	$0.00417

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	$0.00039
MEDIUM	✓ Pass	Needs clarification	Needs clarification	1.60s	$0.00113
LARGE	✓ Pass	Needs clarification	Needs clarification	2.24s	$0.00379

Scenario 11: Recent Performance Request

Query: Show me TD's recent performance
Expected: Status: success, Banks: [2], Year: 2025, Quarters: ['Q3']

Model	Status	Banks	Period	Latency	Cost
SMALL	✗ Fail	[2]	Needs clarification	1.36s	$0.00042
MEDIUM	✓ Pass	[2]	2025 ['Q3']	1.44s	$0.00124
LARGE	✓ Pass	[2]	2025 ['Q3']	7.07s	$0.00415

Scenario 12: Future Period Request

Query: What will RBC's Q1 2026 results be?
Expected: Status: needs_clarification, Banks: [1], Needs Clarification

Model	Status	Banks	Period	Latency	Cost
SMALL	✓ Pass	[1]	Needs clarification	2.95s	$0.00042
MEDIUM	✓ Pass	[1]	Needs clarification	1.79s	$0.00126
LARGE	✓ Pass	[1]	Needs clarification	8.70s	$0.00419

Scenario 13: Ambiguous Last Year

Query: Compare BMO's last year performance
Expected: Status: needs_clarification, Banks: [3], Needs Clarification

Model	Status	Banks	Period	Latency	Cost
SMALL	✓ Pass	[3]	Needs clarification	1.74s	$0.00042
MEDIUM	✓ Pass	[3]	Needs clarification	1.80s	$0.00127
LARGE	✗ Fail	[3]	2024 ['Q1', 'Q2', 'Q3', 'Q4']	7.93s	$0.00416

Scenario 14: QoQ Growth Request

Query: Show Scotia's QoQ growth
Expected: Status: success, Banks: [4], Year: 2025, Quarters: ['Q2', 'Q3']

Model	Status	Banks	Period	Latency	Cost
SMALL	✗ Fail	[4]	Needs clarification	1.50s	$0.00043
MEDIUM	✓ Pass	[4]	2025 ['Q2', 'Q3']	1.76s	$0.00125
LARGE	✓ Pass	[4]	2025 ['Q2', 'Q3']	8.24s	$0.00416

Scenario 15: Current Quarter Unreported

Query: How is CIBC doing this quarter?
Expected: Status: needs_clarification, Banks: [5], Needs Clarification

Model	Status	Banks	Period	Latency	Cost
SMALL	✓ Pass	[5]	Needs clarification	2.47s	$0.00041
MEDIUM	✓ Pass	[5]	Needs clarification	2.85s	$0.00129
LARGE	✓ Pass	[5]	Needs clarification	8.48s	$0.00421

Scenario 16: TTM Request

Query: RBC's TTM revenue
Expected: Status: success, Banks: [1]

Model	Status	Banks	Period	Latency	Cost
SMALL	✗ Fail	[1]	Needs clarification	2.08s	$0.00043
MEDIUM	✓ Pass	[1]	2025 ['Q4', 'Q1', 'Q2', 'Q3']	1.62s	$0.00125
LARGE	✓ Pass	[1]	-	7.91s	$0.00418

Scenario 17: Since Temporal Reference

Query: TD's performance since Q1 2024
Expected: Status: success, Banks: [2]

Model	Status	Banks	Period	Latency	Cost
SMALL	✗ Fail	[2]	Needs clarification	1.39s	$0.00042
MEDIUM	✓ Pass	[2]	2024 ['Q1', 'Q2', 'Q3', 'Q4']	1.64s	$0.00125
LARGE	✓ Pass	[2]	2024 ['Q1', 'Q2', 'Q3', 'Q4']	8.27s	$0.00418

Scenario 18: Cross-Year Comparison

Query: Compare National Bank Q4 2024 vs Q4 2023
Expected: Status: success, Banks: [6]

Model	Status	Banks	Period	Latency	Cost
SMALL	✓ Pass	[6]	2024 ['Q4']	1.42s	$0.00042
MEDIUM	✓ Pass	[6]	2024 ['Q4']	1.70s	$0.00125
LARGE	✓ Pass	[6]	2024 ['Q4']	8.42s	$0.00417

Scenario 19: Previous Quarter Request

Query: BMO's previous quarter results
Expected: Status: success, Banks: [3], Year: 2025, Quarters: ['Q3']

Model	Status	Banks	Period	Latency	Cost
SMALL	✗ Fail	[3]	Needs clarification	1.15s	$0.00042
MEDIUM	✗ Fail	[3]	Needs clarification	2.23s	$0.00126

LARGE	✓ Pass	[3]	2025 ['Q3']	7.90s	$0.00414

Scenario 20: Monthly to Quarterly Mapping

Query: Show me RBC's June performance
Expected: Status: needs_clarification, Banks: [1], Needs Clarification

Model	Status	Banks	Period	Latency	Cost
SMALL	✓ Pass	[1]	Needs clarification	1.87s	$0.00044
MEDIUM	✓ Pass	[1]	Needs clarification	1.62s	$0.00125
LARGE	✓ Pass	[1]	Needs clarification	9.00s	$0.00416

Scenario 21: Partial Bank Name

Query: Show Royal's Q2 2024 metrics
Expected: Status: success, Banks: [1], Year: 2024, Quarters: ['Q2']

Model	Status	Banks	Period	Latency	Cost
SMALL	✗ Fail	Needs clarification	Needs clarification	1.42s	$0.00039
MEDIUM	✓ Pass	[1]	2024 ['Q2']	1.57s	$0.00125
LARGE	✓ Pass	[1]	2024 ['Q2']	8.41s	$0.00415

Scenario 22: Informal Bank Reference

Query: How did the green bank do last quarter?
Expected: Status: needs_clarification, Needs Clarification

Model	Status	Banks	Period	Latency	Cost
SMALL	✓ Pass	Needs clarification	Needs clarification	1.49s	$0.00038
MEDIUM	✓ Pass	Needs clarification	Needs clarification	2.06s	$0.00116
LARGE	✓ Pass	Needs clarification	Needs clarification	7.84s	$0.00392

Scenario 23: Multiple Banks Listed

Query: Compare RBC, TD, and BMO for Q1 2025
Expected: Status: success, Banks: [1, 2, 3], Year: 2025, Quarters: ['Q1']

Model	Status	Banks	Period	Latency	Cost
SMALL	✓ Pass	[1, 2, 3]	2025 ['Q1']	2.62s	$0.00017
MEDIUM	✓ Pass	[1, 2, 3]	2025 ['Q1']	2.92s	$0.00049
LARGE	✓ Pass	[1, 2, 3]	2025 ['Q1']	8.44s	$0.00465

Scenario 24: Category With Exclusion

Query: Big Six except National Bank Q3 2024
Expected: Status: success, Banks: [1, 2, 3, 4, 5], Year: 2024, Quarters: ['Q3']

Model	Status	Banks	Period	Latency	Cost

SMALL	✗ Fail	[1, 2, 3, 4, 5, 6]	Needs clarification	1.93s	$0.00055
MEDIUM	✓ Pass	[1, 2, 3, 4, 5]	2024 ['Q3']	2.78s	$0.00049
LARGE	✗ Fail	[1, 2, 3, 4, 5]	2024 ['Q3']	9.62s	$0.00512

Scenario 25: Bank by Ranking

Query: The largest Canadian bank's latest results
Expected: Status: needs_clarification, Needs Clarification

Model	Status	Banks	Period	Latency	Cost
SMALL	✓ Pass	Needs clarification	Needs clarification	2.07s	$0.00039
MEDIUM	✓ Pass	Needs clarification	Needs clarification	2.42s	$0.00116
LARGE	✗ Fail	[1]	2025 ['Q3']	8.21s	$0.00414

Scenario 26: Bank Name Typo

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

Model	Status	Banks	Period	Latency	Cost
SMALL	✗ Fail	Needs clarification	Needs clarification	1.38s	$0.00038
MEDIUM	✗ Fail	Needs clarification	Needs clarification	2.70s	$0.00117
LARGE	✗ Fail	Needs clarification	Needs clarification	8.38s	$0.00387

Scenario 27: Bank Abbreviations

Query: BMO vs TD Q4 2024
Expected: Status: success, Banks: [3, 2], Year: 2024, Quarters: ['Q4']

Model	Status	Banks	Period	Latency	Cost
SMALL	✗ Fail	[3, 2]	2024 ['Q4']	2.07s	$0.00017
MEDIUM	✓ Pass	[3, 2]	2024 ['Q4']	2.46s	$0.00048
LARGE	✗ Fail	[3, 2]	2024 ['Q4']	8.30s	$0.00439

Scenario 28: Non-Existent Bank

Query: Wells Fargo's Canadian operations Q2 2025
Expected: Status: needs_clarification, Needs Clarification

Model	Status	Banks	Period	Latency	Cost
SMALL	✗ Fail	[9]	2025 ['Q2']	1.70s	$0.00042
MEDIUM	✗ Fail	[9]	2025 ['Q2']	1.76s	$0.00122
LARGE	✗ Fail	[9]	2025 ['Q2']	8.23s	$0.00408

Scenario 29: All Banks Request

Query: Show all banks Q3 2024 efficiency ratios

Expected: Status: success, Banks: [1, 2, 3, 4, 5, 6], Year: 2024, Quarters: ['Q3']

Model	Status	Banks	Period	Latency	Cost
SMALL	✗ Fail	Needs clarification	Needs clarification	1.55s	$0.00039
MEDIUM	✗ Fail	Needs clarification	Needs clarification	1.84s	$0.00114
LARGE	✗ Fail	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10	2024 ['Q3']	10.81s	$0.00601

Scenario 30: Pronoun Without Antecedent

Query: What were their results?
Expected: Status: needs_clarification, Needs Clarification

Model	Status	Banks	Period	Latency	Cost
SMALL	✓ Pass	Needs clarification	Needs clarification	1.97s	$0.00038
MEDIUM	✓ Pass	Needs clarification	Needs clarification	5.51s	$0.00114
LARGE	✓ Pass	Needs clarification	Needs clarification	8.40s	$0.00379

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]

Bank Alias - National Bank (SMALL)

Error: Expected status 'success', got 'needs_clarification'
Mismatch Details:
Expected: Banks=[6], Year=2024, Quarters=['Q1'], Status=success
Actual: Status=needs_clarification, Banks=[6]

Full Year Period (SMALL)

Error: Expected status 'success', got 'needs_clarification'
Mismatch Details:
Expected: Banks=[5], Year=2023, Quarters=['Q1', 'Q2', 'Q3', 'Q4'], Status=success
Actual: Status=needs_clarification, Banks=[5]

Recent Performance Request (SMALL)

Error: Expected status 'success', got 'needs_clarification'
Mismatch Details:
Expected: Banks=[2], Year=2025, Quarters=['Q3'], Status=success
Actual: Status=needs_clarification, Banks=[2]

QoQ Growth Request (SMALL)

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

TTM Request (SMALL)

Error: Expected status 'success', got 'needs_clarification'
Mismatch Details:
Expected: Banks=[1], Status=success
Actual: Status=needs_clarification, Banks=[1]

Since Temporal Reference (SMALL)

Error: Expected status 'success', got 'needs_clarification'
Mismatch Details:
Expected: Banks=[2], Status=success
Actual: Status=needs_clarification, Banks=[2]

Previous Quarter 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]

Partial Bank Name (SMALL)

Error: Expected status 'success', got 'needs_clarification' | Expected banks [1], got None
Mismatch Details:
Expected: Banks=[1], Year=2024, Quarters=['Q2'], Status=success
Actual: Status=needs_clarification

Category With Exclusion (SMALL)

Error: Expected status 'success', got 'needs_clarification' | Expected banks [1, 2, 3, 4, 5], got [1, 2, 3, 4, 5, 6]
Mismatch Details:
Expected: Banks=[1, 2, 3, 4, 5], Year=2024, Quarters=['Q3'], Status=success
Actual: Status=needs_clarification, Banks=[1, 2, 3, 4, 5, 6]

Bank Name Typo (SMALL)

Error: Expected status 'success', got 'needs_clarification' | Expected banks [4], got None
Mismatch Details:
Expected: Banks=[4], Year=2025, Quarters=['Q1', 'Q2', 'Q3'], Status=success
Actual: Status=needs_clarification

Bank Abbreviations (SMALL)

Error: Expected status 'success', got 'needs_clarification'
Mismatch Details:
Expected: Banks=[3, 2], Year=2024, Quarters=['Q4'], Status=success
Actual: Status=needs_clarification, Banks=[3, 2]

Non-Existent Bank (SMALL)

Error: Expected status 'needs_clarification', got 'success' | Expected needs_clarification=True, got False
Mismatch Details:
Expected: Status=needs_clarification
Actual: Status=success, Banks=[9], Year=2025, Quarters=['Q2']

All Banks Request (SMALL)

Error: Expected status 'success', got 'needs_clarification' | Expected banks [1, 2, 3, 4, 5, 6], got None
Mismatch Details:
Expected: Banks=[1, 2, 3, 4, 5, 6], Year=2024, Quarters=['Q3'], Status=success
Actual: Status=needs_clarification

Previous Quarter Request (MEDIUM)

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

Bank Name Typo (MEDIUM)

Error: Expected status 'success', got 'needs_clarification' | Expected banks [4], got None
Mismatch Details:
Expected: Banks=[4], Year=2025, Quarters=['Q1', 'Q2', 'Q3'], Status=success
Actual: Status=needs_clarification

Non-Existent Bank (MEDIUM)

Error: Expected status 'needs_clarification', got 'success' | Expected needs_clarification=True, got False
Mismatch Details:
Expected: Status=needs_clarification
Actual: Status=success, Banks=[9], Year=2025, Quarters=['Q2']

All Banks Request (MEDIUM)

Error: Expected status 'success', got 'needs_clarification' | Expected banks [1, 2, 3, 4, 5, 6], got None
Mismatch Details:
Expected: Banks=[1, 2, 3, 4, 5, 6], Year=2024, Quarters=['Q3'], Status=success
Actual: Status=needs_clarification

Ambiguous Last Year (LARGE)

Error: Expected status 'needs_clarification', got 'success' | Expected needs_clarification=True, got False
Mismatch Details:
Expected: Banks=[3], Status=needs_clarification
Actual: Status=success, Banks=[3], Year=2024, Quarters=['Q1', 'Q2', 'Q3', 'Q4']

Category With Exclusion (LARGE)

Error: Expected status 'success', got 'needs_clarification'
Mismatch Details:
Expected: Banks=[1, 2, 3, 4, 5], Year=2024, Quarters=['Q3'], Status=success
Actual: Status=needs_clarification, Banks=[1, 2, 3, 4, 5], Year=2024, Quarters=['Q3']

Bank by Ranking (LARGE)

Error: Expected status 'needs_clarification', got 'success' | Expected needs_clarification=True, got False
Mismatch Details:
Expected: Status=needs_clarification
Actual: Status=success, Banks=[1], Year=2025, Quarters=['Q3']

Bank Name Typo (LARGE)

Error: Expected status 'success', got 'needs_clarification' | Expected banks [4], got None
Mismatch Details:
Expected: Banks=[4], Year=2025, Quarters=['Q1', 'Q2', 'Q3'], Status=success
Actual: Status=needs_clarification

Bank Abbreviations (LARGE)

Error: Expected status 'success', got 'needs_clarification'
Mismatch Details:
Expected: Banks=[3, 2], Year=2024, Quarters=['Q4'], Status=success
Actual: Status=needs_clarification, Banks=[3, 2], Year=2024, Quarters=['Q4']

Non-Existent Bank (LARGE)

Error: Expected status 'needs_clarification', got 'success' | Expected needs_clarification=True, got False
Mismatch Details:
Expected: Status=needs_clarification
Actual: Status=success, Banks=[9], Year=2025, Quarters=['Q2']

All Banks Request (LARGE)

Error: Expected banks [1, 2, 3, 4, 5, 6], got [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]

Mismatch Details:
Expected: Banks=[1, 2, 3, 4, 5, 6], Year=2024, Quarters=['Q3'], Status=success
Actual: Status=success, Banks=[1, 2, 3, 4, 5, 6, 7, 8, 9, 10], Year=2024, Quarters=['Q3']