Clarifier Agent Test Suite

Multi-Model Performance Analysis

Agent: clarifier_multi_model
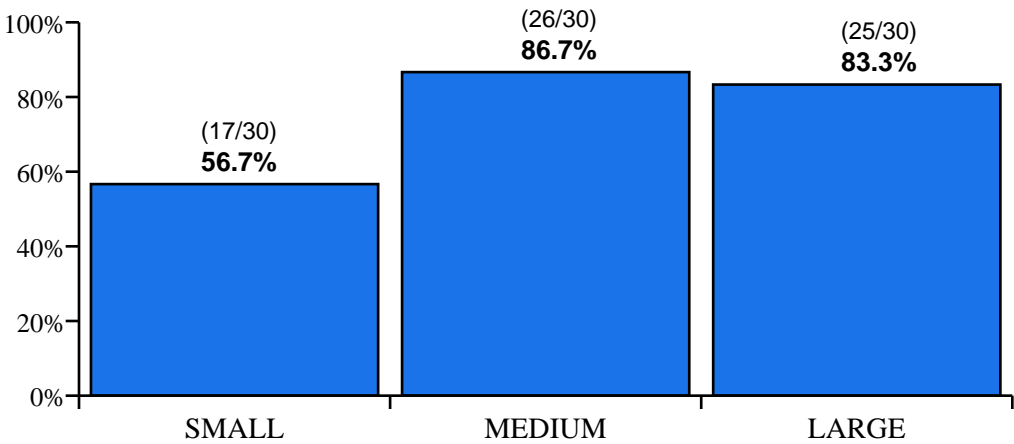**Version:** 1.0.0
Generated: 2025-08-25 00:49:34

Executive Summary

Model	Tests	Passed	Success Rate	Avg Latency	Total Cost
SMALL	30	17/30	56.7%	1.42s	$0.0119
MEDIUM	30	26/30	86.7%	2.01s	$0.0325
LARGE	30	25/30	83.3%	5.83s	$0.1188
OVERALL	**90**	**68/90**	**75.6%**	**3.09s**	**$0.1633**

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']	1.37s	$0.00040
MEDIUM	✓ Pass	[1]	2024 ['Q3']	1.47s	$0.00119
LARGE	✓ Pass	[1]	2024 ['Q3']	1.48s	$0.00397

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']	1.51s	$0.00054
MEDIUM	✓ Pass	[1, 2, 3, 4, 5, 6]	2024 ['Q2']	1.91s	$0.00155
LARGE	✓ Pass	[1, 2, 3, 4, 5, 6]	2024 ['Q2']	1.24s	$0.00518

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	1.24s	$0.00036
MEDIUM	✓ Pass	Needs clarification	Needs clarification	2.46s	$0.00110
LARGE	✓ Pass	Needs clarification	Needs clarification	1.78s	$0.00366

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	1.03s	$0.00040
MEDIUM	✓ Pass	[2]	Needs clarification	2.03s	$0.00119
LARGE	✓ Pass	[2]	Needs clarification	1.37s	$0.00397

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	1.12s	$0.00040

MEDIUM	✓ Pass	[3]	2025 ['Q3']	1.86s	$0.00119
LARGE	✓ Pass	[3]	2025 ['Q3']	1.38s	$0.00396

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	1.63s	$0.00041
MEDIUM	✓ Pass	[4]	2025 ['Q1', 'Q2', 'Q3']	2.05s	$0.00120
LARGE	✓ Pass	[4]	2025 ['Q1', 'Q2', 'Q3']	1.45s	$0.00398

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	1.24s	$0.00041
MEDIUM	✓ Pass	[6]	2024 ['Q1']	1.89s	$0.00119
LARGE	✓ Pass	[6]	2024 ['Q1']	1.28s	$0.00397

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']	1.84s	$0.00044
MEDIUM	✓ Pass	[1, 2]	2023 ['Q4']	2.45s	$0.00043
LARGE	✗ Fail	[1, 2]	2024 ['Q4']	1.36s	$0.00422

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	1.35s	$0.00041
MEDIUM	✗ Fail	[5]	Needs clarification	3.24s	$0.00123
LARGE	✓ Pass	[5]	2023 ['Q1', 'Q2', 'Q3', 'Q4']	1.44s	$0.00398

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	1.52s	$0.00037
MEDIUM	✓ Pass	Needs clarification	Needs clarification	1.87s	$0.00108
LARGE	✓ Pass	Needs clarification	Needs clarification	7.78s	$0.00361

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	0.92s	$0.00040
MEDIUM	✓ Pass	[2]	2025 ['Q3']	1.55s	$0.00119
LARGE	✓ Pass	[2]	2025 ['Q3']	7.58s	$0.00397

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	1.26s	$0.00040
MEDIUM	✓ Pass	[1]	Needs clarification	1.70s	$0.00120
LARGE	✓ Pass	[1]	Needs clarification	7.98s	$0.00401

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	1.20s	$0.00040
MEDIUM	✓ Pass	[3]	Needs clarification	1.94s	$0.00121
LARGE	✗ Fail	[3]	2024 ['Q1', 'Q2', 'Q3', 'Q4']	7.78s	$0.00398

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	1.13s	$0.00041
MEDIUM	✓ Pass	[4]	2025 ['Q2', 'Q3']	1.62s	$0.00119
LARGE	✓ Pass	[4]	2025 ['Q2', 'Q3']	7.89s	$0.00397

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	1.41s	$0.00041
MEDIUM	✓ Pass	[5]	Needs clarification	2.48s	$0.00122
LARGE	✓ Pass	[5]	Needs clarification	8.19s	$0.00402

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	1.56s	$0.00042
MEDIUM	✓ Pass	[1]	2025 ['Q4', 'Q1', 'Q2', 'Q3']	2.25s	$0.00119
LARGE	✓ Pass	[1]	-	7.89s	$0.00400

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	1.33s	$0.00040
MEDIUM	✓ Pass	[2]	2024 ['Q1', 'Q2', 'Q3', 'Q4']	1.74s	$0.00120
LARGE	✓ Pass	[2]	2024 ['Q1', 'Q2', 'Q3', 'Q4']	7.47s	$0.00399

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']	1.13s	$0.00041
MEDIUM	✓ Pass	[6]	2024 ['Q4']	1.43s	$0.00119
LARGE	✓ Pass	[6]	2024 ['Q4']	7.89s	$0.00397

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	1.02s	$0.00040
MEDIUM	✗ Fail	[3]	2025 ['Q2']	1.43s	$0.00119

LARGE	✓ Pass	[3]	2025 ['Q3']	7.81s	$0.00396

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	1.54s	$0.00041
MEDIUM	✓ Pass	[1]	Needs clarification	1.53s	$0.00120
LARGE	✓ Pass	[1]	Needs clarification	7.82s	$0.00398

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	1.43s	$0.00038
MEDIUM	✓ Pass	[1]	2024 ['Q2']	1.64s	$0.00119
LARGE	✓ Pass	[1]	2024 ['Q2']	7.92s	$0.00396

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	1.33s	$0.00037
MEDIUM	✓ Pass	Needs clarification	Needs clarification	2.46s	$0.00111
LARGE	✓ Pass	Needs clarification	Needs clarification	7.84s	$0.00375

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']	1.64s	$0.00047
MEDIUM	✓ Pass	[1, 2, 3]	2025 ['Q1']	2.32s	$0.00043
LARGE	✓ Pass	[1, 2, 3]	2025 ['Q1']	8.14s	$0.00446

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	1.54s	$0.00053
MEDIUM	✓ Pass	[1, 2, 3, 4, 5]	2024 ['Q3']	3.21s	$0.00043
LARGE	✓ Pass	[1, 2, 3, 4, 5]	2024 ['Q3']	10.04s	$0.00494

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	1.56s	$0.00037
MEDIUM	✓ Pass	Needs clarification	Needs clarification	2.47s	$0.00111
LARGE	✗ Fail	[1]	2025 ['Q3']	3.79s	$0.00396

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	1.10s	$0.00037
MEDIUM	✓ Pass	[4]	2025 ['Q1', 'Q2', 'Q3']	1.63s	$0.00119
LARGE	✗ Fail	Needs clarification	Needs clarification	6.17s	$0.00373

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	✓ Pass	[3, 2]	2024 ['Q4']	1.84s	$0.00015
MEDIUM	✓ Pass	[3, 2]	2024 ['Q4']	2.15s	$0.00043
LARGE	✓ Pass	[3, 2]	2024 ['Q4']	7.97s	$0.00421

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	✓ Pass	[9]	Needs clarification	3.07s	$0.00039
MEDIUM	✗ Fail	[9]	2025 ['Q2']	1.85s	$0.00117
LARGE	✓ Pass	Needs clarification	Needs clarification	9.71s	$0.00215

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	1.54s	$0.00037
MEDIUM	✗ Fail	Needs clarification	Needs clarification	1.43s	$0.00108
LARGE	✗ Fail	Needs clarification	Needs clarification	7.17s	$0.00364

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	1.33s	$0.00036
MEDIUM	✓ Pass	Needs clarification	Needs clarification	2.15s	$0.00108
LARGE	✓ Pass	Needs clarification	Needs clarification	7.34s	$0.00362

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

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

Full Year Period (MEDIUM)

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]

Previous Quarter Request (MEDIUM)

Error: Expected quarters ['Q3'], got ['Q2']
Mismatch Details:
Expected: Banks=[3], Year=2025, Quarters=['Q3'], Status=success
Actual: Status=success, Banks=[3], Year=2025, Quarters=['Q2']

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

Multiple Specific Banks (LARGE)

Error: Expected year 2023, got 2024
Mismatch Details:
Expected: Banks=[1, 2], Year=2023, Quarters=['Q4'], Status=success
Actual: Status=success, Banks=[1, 2], Year=2024, Quarters=['Q4']

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']

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

All Banks Request (LARGE)

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