Planner Agent Test Suite

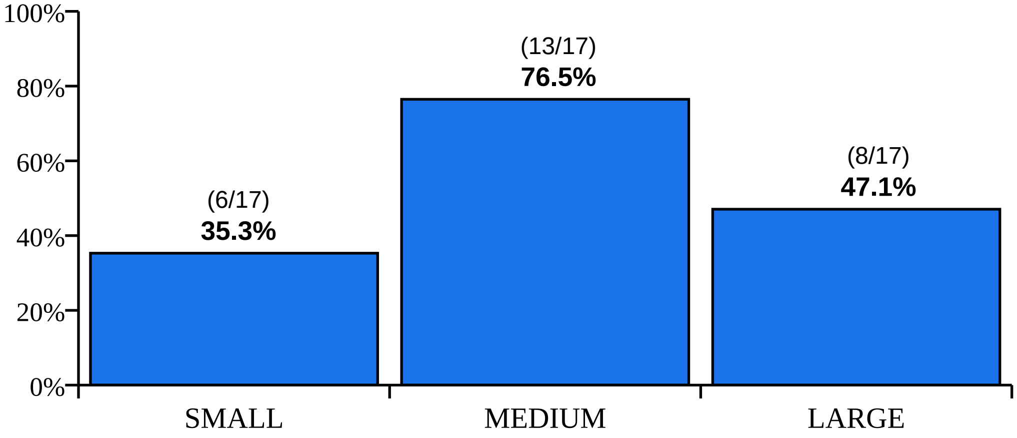
Multi-Model Database Selection Analysis

Agent: planner_multi_model_20250825
Version: 1.0.0
Generated: 2025-08-25 13:29:34

Executive Summary

Model	Tests	Passed	Success Rate	Avg Latency	Total Cost
SMALL	17	6/17	35.3%	1.88s	$0.0054
MEDIUM	17	13/17	76.5%	2.05s	$0.0155
LARGE	17	8/17	47.1%	3.18s	$0.0527
OVERALL	**51**	**27/51**	**52.9%**	**2.37s**	**$0.0736**

Database Selection Accuracy by Model



Detailed Test Results

Scenario 1: Management Commentary - Digital Transformation

Query: [{ "role": "user", "content": "What did RBC management say about digital transformation in Q1 2025?" }]
Expected Databases: transcripts

Model	Status	Databases Selected	Latency	Cost
SMALL	✓ Pass	transcripts	3.73s	$0.00041
MEDIUM	✓ Pass	transcripts	1.51s	$0.00099
LARGE	✓ Pass	transcripts	1.91s	$0.00328

Scenario 2: Efficiency Ratio Comparison

Query: [{ "role": "user", "content": "Compare the efficiency ratios between RBC and TD for Q2 2025" }]
Expected Databases: benchmarking, rts

Model	Status	Databases Selected	Latency	Cost
SMALL	✗ Fail	benchmarking	1.43s	$0.00034
MEDIUM	✓ Pass	benchmarking, rts	2.01s	$0.00102
LARGE	✗ Fail	benchmarking	2.90s	$0.00346

Scenario 3: Capital Ratios Query

Query: [{ "role": "user", "content": "What are the CET1 ratios for the Big Six banks in Q3 2024?" }]
Expected Databases: benchmarking, pillar3

Model	Status	Databases Selected	Latency	Cost
SMALL	✗ Fail	benchmarking	2.38s	$0.00037
MEDIUM	✓ Pass	benchmarking, pillar3	2.32s	$0.00109
LARGE	✗ Fail	benchmarking	4.51s	$0.00371

Scenario 4: Comprehensive Credit Analysis

Query: [{ "role": "user", "content": "Analyze RBC's credit performance in Q1 2025 - what did management say and what are the actual numbers?" }]
Expected Databases: benchmarking, rts, transcripts

Model	Status	Databases Selected	Latency	Cost
SMALL	✗ Fail	transcripts	3.13s	$0.00038
MEDIUM	✓ Pass	transcripts, benchmarking, rts	2.73s	$0.00103
LARGE	✗ Fail	benchmarking	2.77s	$0.00369

Scenario 5: NIM Trend Analysis

Query: [{ "role": "user", "content": "Show me the NIM trend for TD Bank from Q1 to Q4 2024" }]
Expected Databases: benchmarking, rts

Model	Status	Databases Selected	Latency	Cost

SMALL	✗ Fail	benchmarking	1.49s	$0.00036
MEDIUM	✓ Pass	benchmarking, rts	6.35s	$0.00106
LARGE	✗ Fail	benchmarking	3.79s	$0.00356

Scenario 6: Business Segment Analysis

Query: [{ "role": "user", "content": "What's the performance of RBC's wealth management segment in Q2 2025?" }]
Expected Databases: benchmarking, rts

Model	Status	Databases Selected	Latency	Cost
SMALL	✗ Fail	benchmarking	1.42s	$0.00033
MEDIUM	✓ Pass	benchmarking, rts	1.87s	$0.00100
LARGE	✓ Pass	benchmarking, rts	4.18s	$0.00338

Scenario 7: Forward Guidance Query

Query: [{ "role": "user", "content": "What guidance did BMO provide for fiscal 2025 in their latest call?" }]
Expected Databases: transcripts

Model	Status	Databases Selected	Latency	Cost
SMALL	✓ Pass	transcripts	1.93s	$0.00023
MEDIUM	✓ Pass	transcripts	1.33s	$0.00065
LARGE	✓ Pass	transcripts	3.24s	$0.00219

Scenario 8: Credit Risk Analysis

Query: [{ "role": "user", "content": "Analyze Scotia's credit risk metrics and management commentary for Q1 2025" }]
Expected Databases: benchmarking, pillar3, rts, transcripts

Model	Status	Databases Selected	Latency	Cost
SMALL	✗ Fail	benchmarking, transcripts	2.41s	$0.00035
MEDIUM	✗ Fail	benchmarking, pillar3, transcripts	2.88s	$0.00105
LARGE	✓ Pass	benchmarking, rts, pillar3, transcripts	9.68s	$0.00381

Scenario 9: ROE Ranking Query

Query: [{ "role": "user", "content": "Rank all Canadian banks by ROE for Q2 2025" }]
Expected Databases: benchmarking, rts

Model	Status	Databases Selected	Latency	Cost
SMALL	✗ Fail	benchmarking	2.50s	$0.00037
MEDIUM	✓ Pass	benchmarking, rts	2.24s	$0.00109
LARGE	✗ Fail	benchmarking	5.61s	$0.00368

Scenario 10: Detailed Earnings Query

Query: [{ "role": "user", "content": "Get CIBC's detailed earnings breakdown from their Q3 2025 report to shareholders" }]

Expected Databases: rts

Model	Status	Databases Selected	Latency	Cost
SMALL	✗ Fail	-	1.83s	$0.00023
MEDIUM	✗ Fail	-	1.73s	$0.00068
LARGE	✗ Fail	-	2.96s	$0.00223

Scenario 11: Transcript Call Summary Request

Query: [{ "role": "user", "content": "Show me the transcript call summary report for RBC's Q2 2025" }]
Expected Databases: reports

Model	Status	Databases Selected	Latency	Cost
SMALL	✓ Pass	reports	1.13s	$0.00034
MEDIUM	✓ Pass	reports	1.21s	$0.00097
LARGE	✓ Pass	reports	1.27s	$0.00322

Scenario 12: Key Themes Analysis

Query: [{ "role": "user", "content": "What are the key themes from TD's latest earnings calls?" }]
Expected Databases: reports

Model	Status	Databases Selected	Latency	Cost
SMALL	✓ Pass	reports	1.63s	$0.00023
MEDIUM	✓ Pass	reports	0.94s	$0.00064
LARGE	✓ Pass	reports	1.84s	$0.00215

Scenario 13: NIM Analysis with Context

Query: [{ "role": "user", "content": "What's RBC's NIM for Q1 2025 and how did management explain the changes?" }]
Expected Databases: benchmarking, rts, transcripts

Model	Status	Databases Selected	Latency	Cost
SMALL	✗ Fail	benchmarking	1.28s	$0.00034
MEDIUM	✓ Pass	benchmarking, rts, transcripts	2.59s	$0.00103
LARGE	✗ Fail	benchmarking	3.79s	$0.00353

Scenario 14: US Bank 10-Q Analysis

Query: [{ "role": "user", "content": "Get JPMorgan's Q2 2025 10-Q filing details" }]
Expected Databases: rts

Model	Status	Databases Selected	Latency	Cost
SMALL	✗ Fail	-	1.21s	$0.00022
MEDIUM	✗ Fail	-	1.82s	$0.00066
LARGE	✗ Fail	-	1.23s	$0.00217

Scenario 15: Simple Line Item Query

Query: [{ "role": "user", "content": "What was Scotia's net income for Q3 2025?" }]
Expected Databases: benchmarking, rts

Model	Status	Databases Selected	Latency	Cost
SMALL	✗ Fail	reports	1.28s	$0.00023
MEDIUM	✗ Fail	reports	1.02s	$0.00063
LARGE	✗ Fail	transcripts	1.79s	$0.00217

Scenario 16: CM Readthrough Report Request

Query: [{ "role": "user", "content": "Get me the CM readthrough report for RBC Q2 2025" }]
Expected Databases: reports

Model	Status	Databases Selected	Latency	Cost
SMALL	✓ Pass	reports	1.22s	$0.00034
MEDIUM	✓ Pass	reports	1.14s	$0.00097
LARGE	✓ Pass	reports	1.22s	$0.00324

Scenario 17: RTS Blackline Report Request

Query: [{ "role": "user", "content": "Show the RTS blackline comparison for TD's Q1 2025" }]
Expected Databases: reports

Model	Status	Databases Selected	Latency	Cost
SMALL	✓ Pass	reports	1.99s	$0.00034
MEDIUM	✓ Pass	reports	1.21s	$0.00097
LARGE	✓ Pass	reports	1.42s	$0.00322

Failed Test Analysis

Efficiency Ratio Comparison

Model: SMALL
Error: Expected databases ['benchmarking', 'rts'], got ['benchmarking']
Expected: benchmarking, rts
Actual: benchmarking

Model: LARGE
Error: Expected databases ['benchmarking', 'rts'], got ['benchmarking']
Expected: benchmarking, rts
Actual: benchmarking

Capital Ratios Query

Model: SMALL
Error: Expected databases ['benchmarking', 'pillar3'], got ['benchmarking']
Expected: benchmarking, pillar3
Actual: benchmarking

Model: LARGE
Error: Expected databases ['benchmarking', 'pillar3'], got ['benchmarking']
Expected: benchmarking, pillar3
Actual: benchmarking

Comprehensive Credit Analysis

Model: SMALL
Error: Expected databases ['benchmarking', 'rts', 'transcripts'], got ['transcripts']
Expected: benchmarking, rts, transcripts
Actual: transcripts

Model: LARGE
Error: Expected databases ['benchmarking', 'rts', 'transcripts'], got ['benchmarking']
Expected: benchmarking, rts, transcripts
Actual: benchmarking

NIM Trend Analysis

Model: SMALL
Error: Expected databases ['benchmarking', 'rts'], got ['benchmarking']
Expected: benchmarking, rts
Actual: benchmarking

Model: LARGE
Error: Expected databases ['benchmarking', 'rts'], got ['benchmarking']
Expected: benchmarking, rts
Actual: benchmarking

Business Segment Analysis

Model: SMALL
Error: Expected databases ['benchmarking', 'rts'], got ['benchmarking']
Expected: benchmarking, rts
Actual: benchmarking

Credit Risk Analysis

Model: SMALL
Error: Expected databases ['benchmarking', 'pillar3', 'rts', 'transcripts'], got ['benchmarking', 'transcripts']
Expected: benchmarking, pillar3, rts, transcripts
Actual: benchmarking, transcripts

Model: MEDIUM
Error: Expected databases ['benchmarking', 'pillar3', 'rts', 'transcripts'], got ['benchmarking', 'pillar3', 'transcripts']
Expected: benchmarking, pillar3, rts, transcripts
Actual: benchmarking, pillar3, transcripts

ROE Ranking Query

Model: SMALL
Error: Expected databases ['benchmarking', 'rts'], got ['benchmarking']
Expected: benchmarking, rts
Actual: benchmarking

Model: LARGE
Error: Expected databases ['benchmarking', 'rts'], got ['benchmarking']
Expected: benchmarking, rts
Actual: benchmarking

Detailed Earnings Query

Model: SMALL
Error: Expected status 'success', got 'no_databases' | Expected databases ['rts'], got []
Expected: rts
Actual: None

Model: MEDIUM
Error: Expected status 'success', got 'no_databases' | Expected databases ['rts'], got []
Expected: rts
Actual: None

Model: LARGE
Error: Expected status 'success', got 'no_databases' | Expected databases ['rts'], got []
Expected: rts
Actual: None

NIM Analysis with Context

Model: SMALL
Error: Expected databases ['benchmarking', 'rts', 'transcripts'], got ['benchmarking']
Expected: benchmarking, rts, transcripts
Actual: benchmarking

Model: LARGE
Error: Expected databases ['benchmarking', 'rts', 'transcripts'], got ['benchmarking']
Expected: benchmarking, rts, transcripts
Actual: benchmarking

US Bank 10-Q Analysis

Model: SMALL
Error: Expected status 'success', got 'no_databases' | Expected databases ['rts'], got []
Expected: rts
Actual: None

Model: MEDIUM
Error: Expected status 'success', got 'no_databases' | Expected databases ['rts'], got []
Expected: rts
Actual: None

Model: LARGE
Error: Expected status 'success', got 'no_databases' | Expected databases ['rts'], got []
Expected: rts
Actual: None

Simple Line Item Query

Model: SMALL
Error: Expected databases ['benchmarking', 'rts'], got ['reports']
Expected: benchmarking, rts
Actual: reports

Model: MEDIUM
Error: Expected databases ['benchmarking', 'rts'], got ['reports']
Expected: benchmarking, rts
Actual: reports

Model: LARGE
Error: Expected databases ['benchmarking', 'rts'], got ['transcripts']
Expected: benchmarking, rts
Actual: transcripts