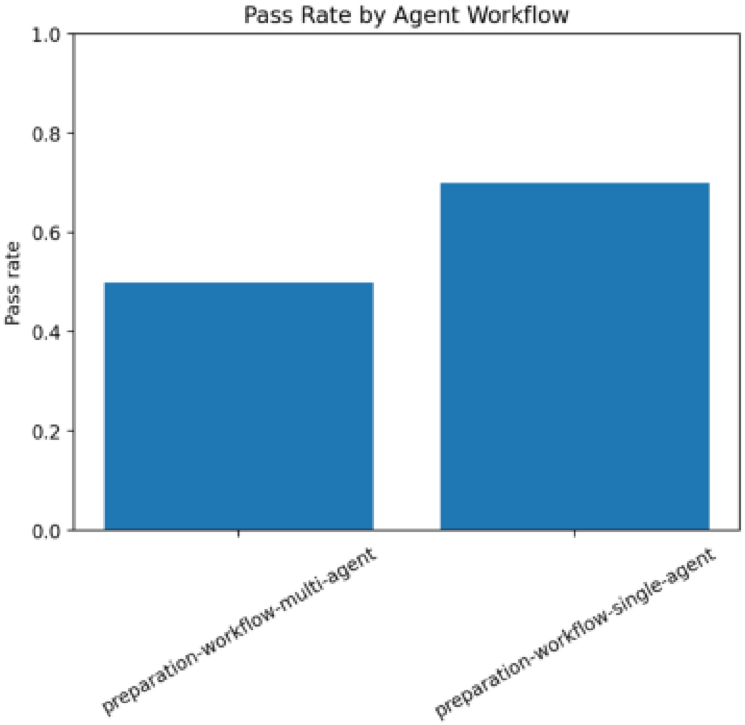
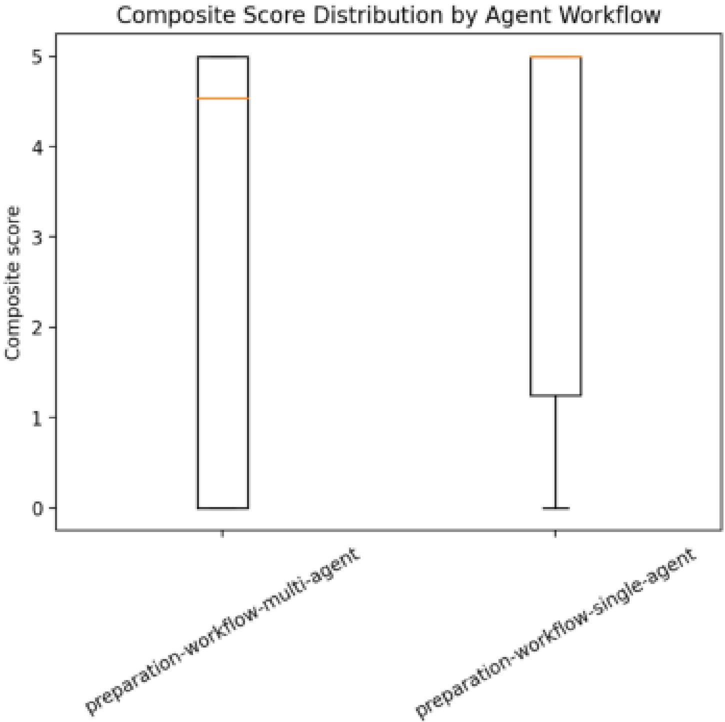
Evaluation Report Visualization

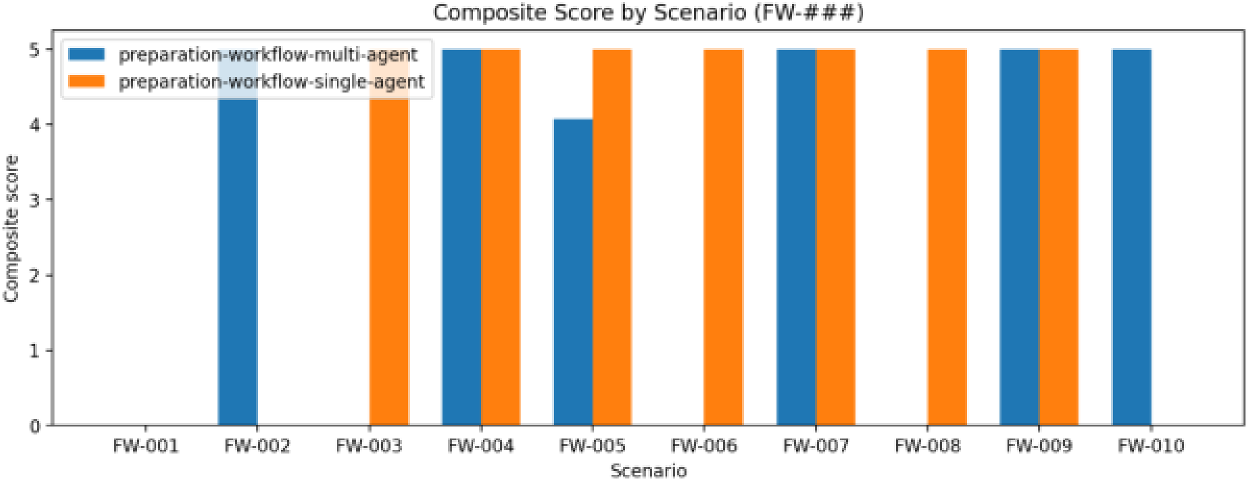
generatedAtUtc: 2026-02-25T00:18:35.0450425+00:00
totalCases: 20
totalPassed: 12
overallPassRate: 0.600

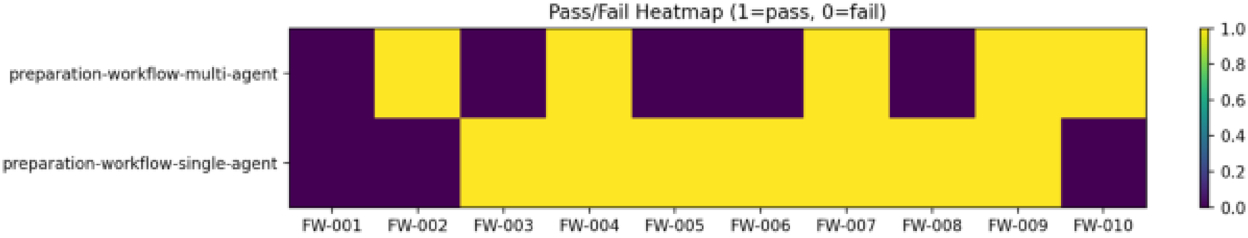
Agent summary:

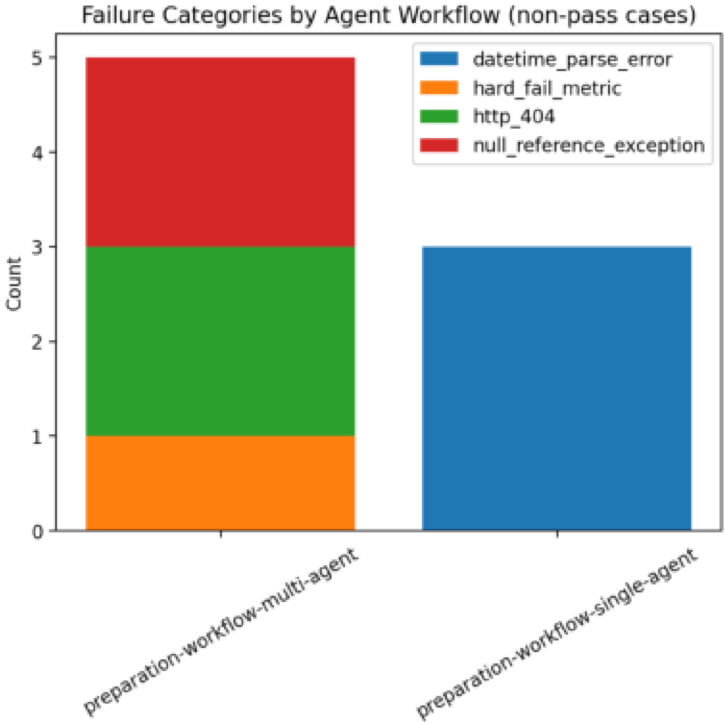
- preparation-workflow-multi-agent: passRate=0.500 passed=5/10, compositeAvg(report)=2.908, metricCoverage=0.600
- preparation-workflow-single-agent: passRate=0.700 passed=7/10, compositeAvg(report)=3.000, metricCoverage=0.700

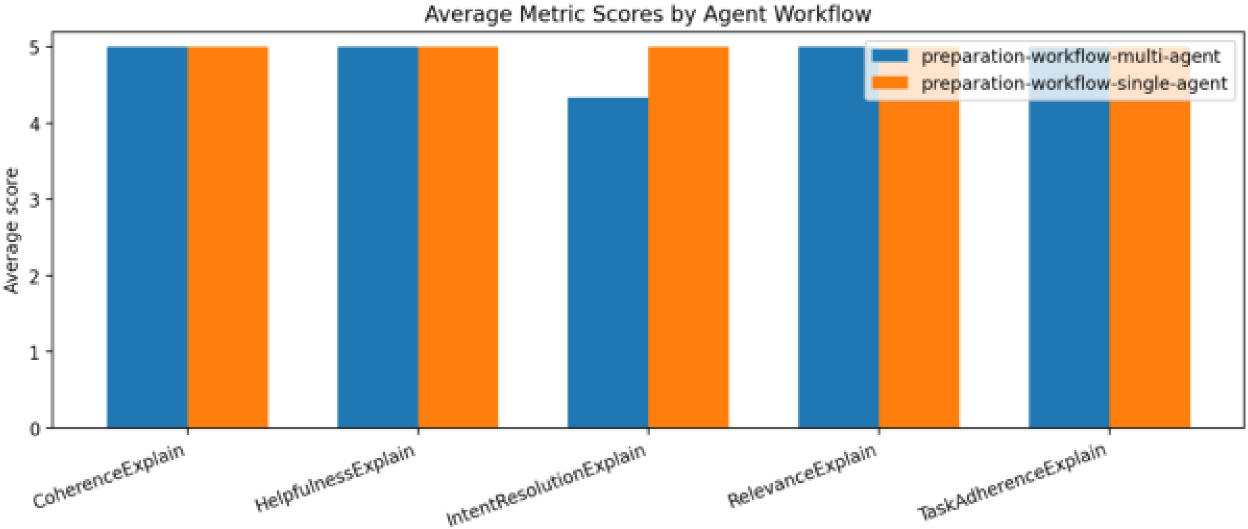

Pass Rate by Agent Workflow
Pass rate
1.0
0.8
0.6
0.4
0.2
0.0
preparation-workflow-multi-agent
preparation-workflow-single-agent

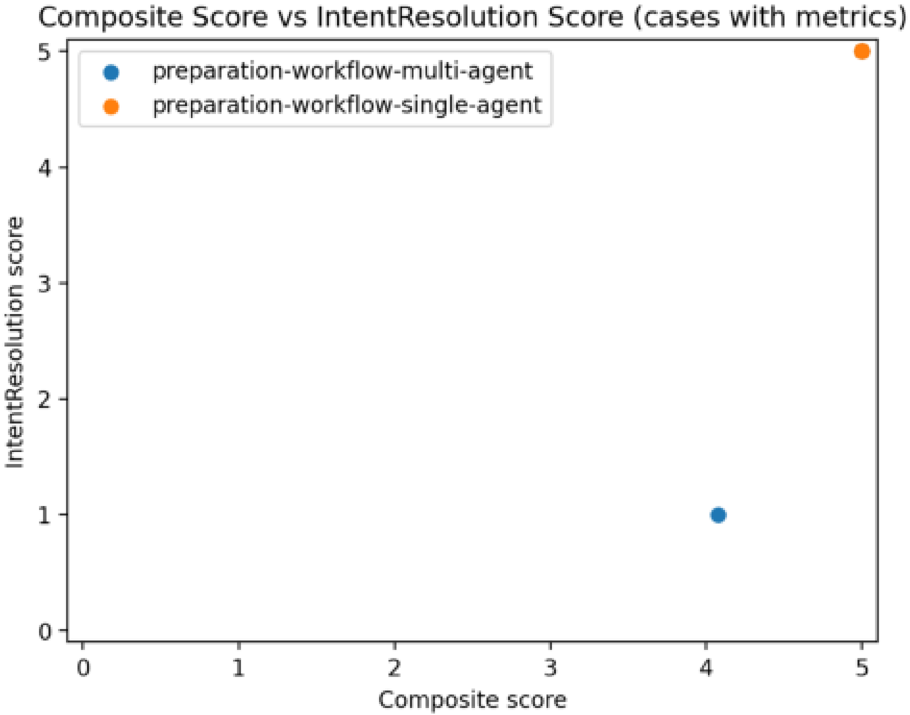

Composite Score Distribution by Agent Workflow
Composite score
5
4
3
2
1
0
preparation-workflow-multi-agent
preparation-workflow-single-agent

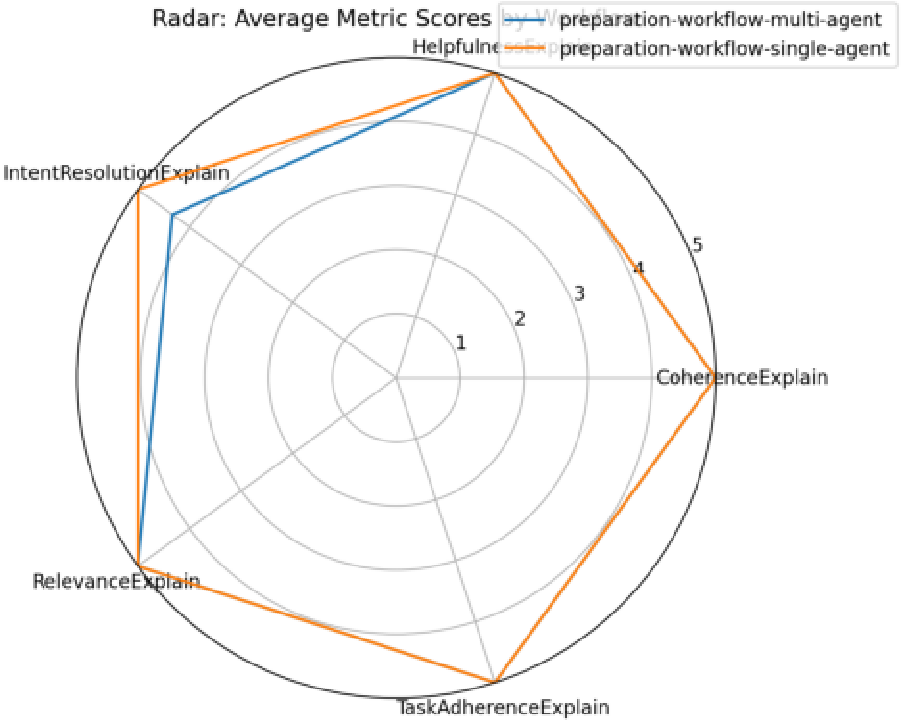

Composite Score by Scenario (FW-###)
preparation-workflow-multi-agent
preparation-workflow-single-agent
Composite score
0
1
2
3
4
5
FW-001
FW-002
FW-003
FW-004
FW-005
FW-006
FW-007
FW-008
FW-009
FW-010
Scenario

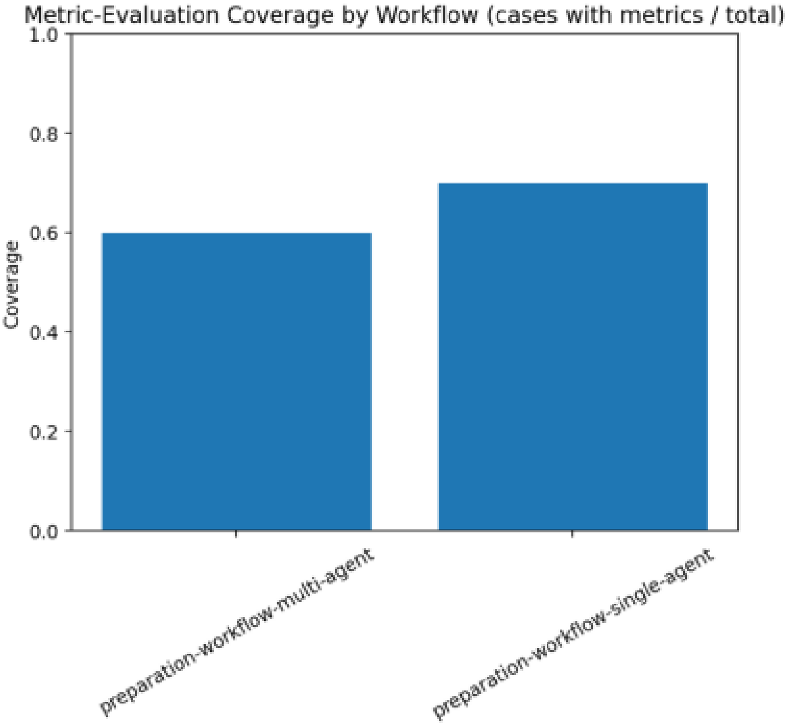

Pass/Fail Heatmap (1=pass, 0=fail)
preparation-workflow-multi-agent
preparation-workflow-single-agent
FW-001
FW-002
FW-003
FW-004
FW-005
FW-006
FW-007
FW-008
FW-009
FW-010
1.0
0.8
0.6
0.4
0.2
0.0


Failure Categories by Agent Workflow (non-pass cases)
datetime_parse_error
hard_fail_metric
http_404
null_reference_exception
Count
0
1
2
3
4
5
preparation-workflow-multi-agent
preparation-workflow-single-agent


Average Metric Scores by Agent Workflow
Average score
0
1
2
3
4
5
preparation-workflow-multi-agent
preparation-workflow-single-agent
CoherenceExplain
HelpfulnessExplain
IntentResolutionExplain
RelevanceExplain
TaskAdherenceExplain


Composite Score vs IntentResolution Score (cases with metrics)
preparation-workflow-multi-agent
preparation-workflow-single-agent
IntentResolution score
Composite score
0
1
2
3
4
5
0
1
2
3
4
5


Radar: Average Metric Scores by Workflow
preparation-workflow-multi-agent
preparation-workflow-single-agent
HelpfulnessExplain
IntentResolutionExplain
CoherenceExplain
RelevanceExplain
TaskAdherenceExplain
1
2
3
4
5


Metric-Evaluation Coverage by Workflow (cases with metrics / total)
Coverage
1.0
0.8
0.6
0.4
0.2
0.0
preparation-workflow-multi-agent
preparation-workflow-single-agent

Scenario Comparison (Pivoted)

scenario	compositeScore.preparation-workflo	compositeScore.preparation-workflo	failureCategory.preparation-workflo	failureCategory.preparation-workflo	passed.preparation-workflo	passed.preparation-workflow-singl
FW-001	0.0	0.0	null_reference_exception	datetime_parse_error	False	False
FW-002	5.0	0.0	none	datetime_parse_error	True	False
FW-003	0.0	5.0	http_404	none	False	True
FW-004	5.0	5.0	none	none	True	True
FW-005	4.076923076923077	5.0	hard_fail_metric	none	False	True
FW-006	0.0	5.0	null_reference_exception	none	False	True
FW-007	5.0	5.0	none	none	True	True
FW-008	0.0	5.0	http_404	none	False	True
FW-009	5.0	5.0	none	none	True	True
FW-010	5.0	0.0	none	datetime_parse_error	True	False