

QA Testing Suite Report

Insurance Claim Multi-Agent System

Report Generated: 2026-01-15
Version: 1.0.0

Executive Summary

Agent Performance Summary

Agent	Score	Status
Needle Agent	82.9%	✓ Excellent
Summary Agent	63.5%	~ Good
Routing Agent	100.0%	✓ Excellent

Agent Performance Details

Agent	Tests	Code Score	Model Score	Combined	HITL Tests	HITL Rating	HITL Score
Needle	20	80.8%	85.0%	82.9%	5	4.00/5	75.0%
Summary	15	-	63.5%	63.5%	5	3.60/5	65.0%
Routing	10	100.0%	-	100.0%	5	5.00/5	100.0%

Detailed Test Results

Needle Agent Tests

Summary: 20 tests | 15 passed (75.0%) | Average Score: 82.9%

Test needle_01 (Score: 100.0% ✓)

Question: What time did the collision occur?

Agent Answer:

The collision occurred on January 15, 2024, at 09:23:45 AM PST.

Ground Truth:

The collision occurred at 09:23:45 AM on January 15, 2024.

CODE GRADER (100.0%)

✓	Time Pattern	09:23:45
✓	Time Format	09:23:45 AM
✓	Date Pattern	January 15, 2024

MODEL GRADER (100.0%)

Factual Accuracy	100.0%
Completeness	100.0%
Precision	100.0%
No Hallucination	100.0%

Reasoning: The agent's answer accurately reflects all details from the ground truth, including the exact time and date of the collision, and it specifies the time zone without introducing any inaccuracies.

Sources:

Page 2: Initial Collision Dynamics | Page 1: Claim Introduction and Overview | Page 2: Initial Collision Dynamics

Test needle_02 (Score: 100.0% ✓)

Question: What is the claim number?

Agent Answer:

The claim number is CLM-2024-00789-AUTO.

Ground Truth:

The claim number is CLM-2024-00789-AUTO.

CODE GRADER (100.0%)

✓	Claim Id Pattern	CLM-2024-00789-AUTO
✓	Claim Format	CLM-2024-00789-AUTO

MODEL GRADER (100.0%)

Factual Accuracy	100.0%
Completeness	100.0%
Precision	100.0%
No Hallucination	100.0%

Reasoning: The agent's answer matches the ground truth exactly, providing all key information accurately and precisely without any hallucination.

Sources:

Page 1: Claim Introduction and Overview | Page 13: Claim Resolution and Final Summary | Page 10: Legal Documentation and Liability Determination

Test needle_03 (Score: 100.0% ✓)

Question: What was Sarah Mitchell's blood pressure during the emergency assessment?

Agent Answer:

Sarah Mitchell's blood pressure during the emergency assessment was recorded at 145/92 mmHg, indicating a hypertensive response.

Ground Truth:

Sarah Mitchell's blood pressure was 145/92 millimeters of mercury.

CODE GRADER (100.0%)

✓ Bp Pattern	145/92
✓ Bp Format	145/92

MODEL GRADER (100.0%)

Factual Accuracy	100.0%
Completeness	100.0%
Precision	100.0%
No Hallucination	100.0%

Reasoning: The agent's answer accurately reflects the ground truth with all key information included and precise details stated correctly, without any hallucinated information.

Sources:

Page 3: Emergency Response and Triage | Page 3: Emergency Response and Triage | Page 3: Emergency Response and Triage

Test needle_04 (Score: 100.0% ✓)

Question: How many feet were the skid marks?

Agent Answer:

The skid marks measured 47 feet, originating from the trajectory of Mr. Chen's vehicle.

Ground Truth:

The skid marks measured 47 feet.

CODE GRADER (100.0%)

✓ Distance Pattern	47
✓ Unit Pattern	feet

MODEL GRADER (100.0%)

Factual Accuracy	100.0%
Completeness	100.0%
Precision	100.0%
No Hallucination	100.0%

Reasoning: The agent's answer accurately reflects the ground truth, includes all key information, states specific details precisely, and does not contain any hallucinated information.

Sources:

Page 3: Emergency Response and Triage | Page 1: Claim Introduction and Overview | Page 12: Repair Completion and Quality Assurance

Test needle_05 (Score: 100.0% ✓)

Question: What was the speed of Robert Chen's vehicle when it entered the intersection?

Agent Answer:

Robert Chen's vehicle was traveling at a speed of 45 mph when it entered the intersection.

Ground Truth:

Robert Chen's vehicle was traveling at 45 miles per hour.

CODE GRADER (100.0%)

- ✓ Speed Pattern 45
- ✓ Unit Pattern mph

MODEL GRADER (100.0%)

Factual Accuracy	100.0%
Completeness	100.0%
Precision	100.0%
No Hallucination	100.0%

Reasoning: The agent's answer accurately reflects the ground truth with all key information included and stated precisely, without any hallucinations.

Sources:

Page 2: Initial Collision Dynamics | Page 7: Witness Testimony and Traffic Engineering | Page 3: Emergency Response and Triage

Test needle_06 (Score: 81.2% ✓)

Question: Who conducted the phone interview with Sarah Mitchell?

Agent Answer:

The phone interview with Sarah Mitchell was conducted by Claims Adjuster Linda Martinez at 16:30:00.

Ground Truth:

Claims Adjuster Linda Martinez conducted the phone interview.

CODE GRADER (100.0%)

- ✓ Name Pattern Linda Martinez
- ✓ Role Pattern Claims Adjuster

MODEL GRADER (62.5%)

Factual Accuracy	100.0%
Completeness	50.0%
Precision	100.0%
No Hallucination	100.0%

Reasoning: The agent's answer correctly identifies the claims adjuster but adds a specific time that is not in the ground truth, affecting completeness. All other details are accurate and precise.

Sources:

Page 3: Emergency Response and Triage | Page 10: Legal Documentation and Liability Determination | Page 4: Hospitalization and Diagnosis

Test needle_07 (Score: 12.5% X)

Question: What was the VIN (vehicle identification number) of Sarah Mitchell's vehicle?

Agent Answer:

The information is not available in the provided documents.

Ground Truth:

Sarah Mitchell's vehicle VIN was 1HGCV1F3XNA123456.

CODE GRADER (0.0%)

<input checked="" type="checkbox"/>	Vin Pattern	None
<input checked="" type="checkbox"/>	Vin Format	None

MODEL GRADER (25.0%)

Factual Accuracy	0.0%
Completeness	0.0%
Precision	0.0%
No Hallucination	100.0%

Reasoning: The agent's answer is factually incorrect and lacks key information about the VIN, but does not include hallucinated information.

Sources:

Page 10: Legal Documentation and Liability Determination | Page 12: Repair Completion and Quality Assurance | Page 1: Claim Introduction and Overview

Test needle_08 (Score: 100.0% ✓)

Question: How long was the ambulance transport to the hospital?

Agent Answer:

The ambulance transport to the hospital lasted 23 minutes, departing the scene at 09:52:00 and arriving at the Emergency Department bay at 10:15:33 AM.

Ground Truth:

The ambulance transport took 23 minutes, departing at 09:52:00 and arriving at 10:15:33 AM.

CODE GRADER (100.0%)

<input checked="" type="checkbox"/>	Duration Pattern	23 minutes
<input checked="" type="checkbox"/>	Time Pattern	09:52

MODEL GRADER (100.0%)

Factual Accuracy	100.0%
Completeness	100.0%
Precision	100.0%
No Hallucination	100.0%

Reasoning: The agent's answer accurately reflects all key details from the ground truth, including the duration, departure, and arrival times, without any hallucinated information.

Sources:

Page 4: Hospitalization and Diagnosis | Page 4: Hospitalization and Diagnosis | Page 4: Hospitalization and Diagnosis

Test needle_09 (Score: 100.0% ✓)

Question: Who was the attending physician at Seattle Medical Center?

Agent Answer:

The attending physician at Seattle Medical Center was Dr. Michael Patterson, a board-certified emergency physician with 15 years of trauma experience.

Ground Truth:

Dr. Michael Patterson was the attending physician.

CODE GRADER (100.0%)

✓ Name Pattern	Michael Patterson
✓ Title Pattern	Dr.

MODEL GRADER (100.0%)

Factual Accuracy	100.0%
Completeness	100.0%
Precision	100.0%
No Hallucination	100.0%

Reasoning: The agent's answer accurately states the attending physician's name and includes additional relevant details without introducing any hallucinated information.

Sources:

Page 4: Hospitalization and Diagnosis | Page 4: Hospitalization and Diagnosis | Page 3: Emergency Response and Triage

Test needle_10 (Score: 100.0% ✓)

Question: What was the make and model of Robert Chen's vehicle?

Agent Answer:

Robert Chen's vehicle was a 2019 Toyota Camry.

Ground Truth:

Robert Chen drove a Toyota Camry.

CODE GRADER (100.0%)

✓ Make Pattern	Toyota
✓ Model Pattern	Camry

MODEL GRADER (100.0%)

Factual Accuracy	100.0%
Completeness	100.0%
Precision	100.0%
No Hallucination	100.0%

Reasoning: The agent's answer accurately reflects the make and model of the vehicle, includes all key information, and specifies the year correctly without any hallucinated details.

Sources:

Page 2: Initial Collision Dynamics | Page 10: Legal Documentation and Liability Determination | Page 3: Emergency Response and Triage

Test needle_11 (Score: 75.0% ✓)

Question: How many physical therapy sessions were prescribed?

Agent Answer:

A total of 12 physical therapy sessions were prescribed over a six-week period.

Ground Truth:

12 physical therapy sessions were prescribed over 6 weeks.

CODE GRADER (50.0%)

- | | |
|--------------------|------|
| ✓ Sessions Pattern | 12 |
| ✗ Duration Pattern | None |

MODEL GRADER (100.0%)

Factual Accuracy	100.0%
Completeness	100.0%
Precision	100.0%
No Hallucination	100.0%

Reasoning: The agent's answer accurately reflects all key information from the ground truth, including the number of sessions and the time period, with precise language and no hallucinations.

Sources:

Page 6: Physical Therapy and Rehabilitation | Page 6: Physical Therapy and Rehabilitation | Page 6: Physical Therapy and Rehabilitation

Test needle_12 (Score: 50.0% X)

Question: What was Sarah Mitchell's heart rate during the emergency assessment?

Agent Answer:

Sarah Mitchell's heart rate during the emergency assessment was 98 beats per minute, indicating tachycardia.

Ground Truth:

Sarah Mitchell's heart rate was 92 beats per minute.

CODE GRADER (50.0%)

- | | |
|----------------|------------------|
| ✗ Hr Pattern | None |
| ✓ Unit Pattern | beats per minute |

MODEL GRADER (50.0%)

Factual Accuracy	0.0%
Completeness	50.0%
Precision	50.0%
No Hallucination	100.0%

Reasoning: The agent's answer contains an incorrect heart rate, which affects factual accuracy. It includes some relevant information but lacks the correct key detail, leading to a lower completeness score.

Sources:

[Page 3: Emergency Response and Triage](#) | [Page 3: Emergency Response and Triage](#) | [Page 4: Hospitalization and Diagnosis](#)

Test needle_13 (Score: 100.0% ✓)

Question: What was the intersection where the collision occurred?

Agent Answer:

The collision occurred at the intersection of Maple Avenue and 5th Street in Seattle, Washington 98101.

Ground Truth:

The collision occurred at the intersection of Maple Avenue and 5th Street in Seattle.

CODE GRADER (100.0%)

- | | |
|-------------------|--------------|
| ✓ Street1 Pattern | Maple Avenue |
| ✓ Street2 Pattern | 5th Street |
| ✓ City Pattern | Seattle |

MODEL GRADER (100.0%)

Factual Accuracy	100.0%
Completeness	100.0%
Precision	100.0%
No Hallucination	100.0%

Reasoning: The agent's answer is factually accurate, complete, precise, and does not contain any hallucinated information. It includes all relevant details from the ground truth.

Sources:

Page 2: Initial Collision Dynamics | Page 1: Claim Introduction and Overview | Page 2: Initial Collision Dynamics

Test needle_14 (Score: 100.0% ✓)

Question: What was the ambient temperature at the time of the collision?

Agent Answer:

The ambient temperature at the time of the collision was 42 degrees Fahrenheit.

Ground Truth:

The ambient temperature was 42 degrees Fahrenheit.

CODE GRADER (100.0%)

- | | |
|----------------|---------|
| ✓ Temp Pattern | 42 |
| ✓ Unit Pattern | degrees |

MODEL GRADER (100.0%)

Factual Accuracy	100.0%
Completeness	100.0%
Precision	100.0%
No Hallucination	100.0%

Reasoning: The agent's answer is identical to the ground truth, providing accurate, complete, and precise information without any hallucinations.

Sources:

[Page 2: Initial Collision Dynamics](#) | [Page 1: Claim Introduction and Overview](#) | [Page 1: Claim Introduction and Overview](#)

Test needle_15 (Score: 64.6% X)

Question: Which police officer documented the accident scene?

Agent Answer:

Officer James Wilson documented the accident scene, as indicated in the legal documentation regarding the liability determination.

Ground Truth:

Officer James Wilson from the Seattle Police Department documented the scene.

CODE GRADER (66.7%)

✓ Name Pattern	James Wilson
✓ Title Pattern	Officer
✗ Dept Pattern	None

MODEL GRADER (62.5%)

Factual Accuracy	100.0%
Completeness	50.0%
Precision	100.0%
No Hallucination	100.0%

Reasoning: The agent accurately identifies Officer James Wilson but omits the specific detail of the police department. The additional context about legal documentation is reasonable but not part of the ground truth.

Sources:

Page 10: Legal Documentation and Liability Determination | Page 3: Emergency Response and Triage | Page 1: Claim Introduction and Overview

Test needle_16 (Score: 100.0% ✓)

Question: What was the repair shop's name?

Agent Answer:

The repair shop's name is Premier Auto Body Shop.

Ground Truth:

The vehicle was repaired at Premier Auto Body Shop.

CODE GRADER (100.0%)

✓ Shop Pattern	Premier Auto Body Shop
----------------	------------------------

MODEL GRADER (100.0%)

Factual Accuracy	100.0%
Completeness	100.0%
Precision	100.0%
No Hallucination	100.0%

Reasoning: The agent's answer is factually accurate, complete, precise, and does not include any hallucinated information, matching the ground truth perfectly.

Sources:

Page 9: Parts Procurement and Repair Execution | Page 5: Technical Damage Assessment | Page 5:
Technical Damage Assessment

Test needle_17 (Score: 62.5% X)

Question: How many minutes did it take for emergency responders to arrive?

Agent Answer:

Emergency responders arrived on the scene at 09:31:22 AM, establishing a response time of 7 minutes and 37 seconds from the initial distress call received at 09:23:45 AM.

Ground Truth:

Emergency responders arrived within 8 minutes.

CODE GRADER (50.0%)

<input checked="" type="checkbox"/>	Time Pattern	None
<input checked="" type="checkbox"/>	Unit Pattern	minutes

MODEL GRADER (75.0%)

Factual Accuracy	50.0%
Completeness	50.0%
Precision	100.0%
No Hallucination	100.0%

Reasoning: The agent's answer provides a specific response time but does not match the ground truth of 'within 8 minutes,' leading to partial accuracy and completeness.

Sources:

Page 3: Emergency Response and Triage | Page 4: Hospitalization and Diagnosis | Page 3: Emergency Response and Triage

Test needle_18 (Score: 12.5% X)

Question: What was Sarah Mitchell's pain level rating?

Agent Answer:

The information is not available in the provided documents.

Ground Truth:

Sarah Mitchell reported a pain level of 6 out of 10.

CODE GRADER (0.0%)

<input checked="" type="checkbox"/>	Pain Pattern	None
<input checked="" type="checkbox"/>	Scale Pattern	None

MODEL GRADER (25.0%)

Factual Accuracy	0.0%
Completeness	0.0%
Precision	0.0%
No Hallucination	100.0%

Reasoning: The agent's answer is factually incorrect and lacks key information about Sarah Mitchell's pain level, but it does not include any hallucinated information.

Sources:

Page 3: Emergency Response and Triage | Page 6: Physical Therapy and Rehabilitation | Page 6: Physical Therapy and Rehabilitation

Test needle_19 (Score: 100.0% ✓)

Question: What was the deductible amount on Sarah Mitchell's policy?

Agent Answer:

The deductible amount on Sarah Mitchell's policy was \$500.00.

Ground Truth:

The deductible amount was \$500.

CODE GRADER (100.0%)

✓	Amount Pattern	\$500
✓	Currency Pattern	\$

MODEL GRADER (100.0%)

Factual Accuracy	100.0%
Completeness	100.0%
Precision	100.0%
No Hallucination	100.0%

Reasoning: The agent's answer is factually accurate, complete, precise, and does not contain any hallucinated information, matching the ground truth perfectly.

Sources:

Page 13: Claim Resolution and Final Summary | Page 10: Legal Documentation and Liability Determination | Page 1: Claim Introduction and Overview

Test needle_20 (Score: 100.0% ✓)

Question: What was the lifetime warranty certificate number issued by Premier Auto Body?

Agent Answer:

The lifetime warranty certificate number issued by Premier Auto Body is WAR-2024-0156.

Ground Truth:

The lifetime warranty certificate number was WAR-2024-0156.

CODE GRADER (100.0%)

✓	Cert Pattern	WAR-2024-0156
✓	Cert Format	WAR-2024-0156

MODEL GRADER (100.0%)

Factual Accuracy	100.0%
Completeness	100.0%
Precision	100.0%
No Hallucination	100.0%

Reasoning: The agent's answer is factually accurate, complete, precise, and does not include any hallucinated information, matching the ground truth perfectly.

Sources:

Page 12: Repair Completion and Quality Assurance | Page 9: Parts Procurement and Repair Execution | Page 5: Technical Damage Assessment

Summary Agent Tests

Summary: 15 tests | 4 passed (26.7%) | Average Score: 63.5%

Test summary_01 (Score: 100.0% ✓)

Question: What was the total estimated claim value?

Agent Answer:

The total estimated claim value for Claim ID CLM-2024-00789-AUTO was \$47,850. This amount includes \$44,010 for vehicle damage and \$3,840 for medical expenses incurred by policyholder Sarah Mitchell following the multi-vehicle collision on January 15, 2024.

Ground Truth:

The total estimated claim value was \$47,850, consisting of \$44,010 for vehicle repairs and \$3,840 for medical expenses.

MODEL GRADER (100.0%)

Comprehensiveness	100.0%
Coherence	100.0%
Synthesis	100.0%
Relevance	100.0%
Accuracy	100.0%

Reasoning: The agent's summary covers all major points accurately and is well-organized, presenting the information in a cohesive manner while remaining directly relevant to the question.

Sources:

Page 1: Claim Introduction and Overview | Page 13: Claim Resolution and Final Summary | Page 8: Comprehensive Financial Analysis

Test summary_02 (Score: 60.0% X)

Question: Summarize the events that led to the claim.

Agent Answer:

On January 15, 2024, at approximately 9:23 AM, Sarah Mitchell was involved in a multi-vehicle collision at the intersection of Maple Avenue and 5th Street in Seattle, Washington. The accident, which included three vehicles, resulted in significant property damage and minor injuries. Following the incident, Mitchell was evaluated at Seattle Medical Center, where she was diagnosed with a Grade 2 Whiplash Associated Disorder and a mild concussion. The total estimated claim value for the damages and medical expenses was determined to be \$47,850.

Ground Truth:

On January 15, 2024, at 09:23:45 AM, Sarah Mitchell was involved in a collision at Maple Avenue and 5th Street when Robert Chen's Toyota Camry ran a red light at 45 mph and struck Mitchell's Honda Accord. Emergency responders arrived within 8 minutes, and Mitchell was transported to Seattle Medical Center with whiplash and cervical spine strain. The incident occurred during morning rush hour on wet roads.

MODEL GRADER (60.0%)

Comprehensiveness	50.0%
Coherence	100.0%

Synthesis	50.0%
Relevance	100.0%
Accuracy	50.0%

Reasoning: The agent's summary covers some key points but misses details about the collision specifics and emergency response. It is well-organized and relevant, but some facts are inaccurate or incomplete.

Sources:

Page 1: Claim Introduction and Overview | Page 13: Claim Resolution and Final Summary | Page 11: Subrogation Process and Recovery

Test summary_03 (Score: 50.0% X)

Question: What was the outcome of Sarah Mitchell's medical treatment?

Agent Answer:

Sarah Mitchell's medical treatment following the auto accident included an initial evaluation for a Grade 2 Whiplash Associated Disorder and a mild concussion, diagnosed by Dr. Michael Patterson after her emergency room visit. She underwent a follow-up consultation where persistent cervical spine tenderness was noted, leading to a referral for 12 sessions of physical therapy aimed at improving her cervical mobility and strength. Although she scored 28/50 on the Neck Disability Index, indicating moderate disability, no specific costs for the physical therapy sessions were provided in the documentation. Overall, her treatment involved both emergency care and ongoing rehabilitation efforts.

Ground Truth:

Sarah Mitchell was diagnosed with Grade 2 whiplash, cervical spine strain, and mild concussion. She underwent emergency care, CT scan, and completed 12 physical therapy sessions over 6 weeks. Her pain improved from 6/10 to 4/10, and she received full medical clearance on February 3, 2024, with good prognosis for full recovery. Total medical costs were \$3,840.

MODEL GRADER (50.0%)

Comprehensiveness	50.0%
Coherence	50.0%
Synthesis	50.0%
Relevance	50.0%
Accuracy	50.0%

Reasoning: The agent's summary covers some key points but misses specific details like pain improvement and total costs. It is somewhat organized but lacks a clear flow, and while it presents accurate information, it does not fully align with the reference summary.

Sources:

Page 1: Claim Introduction and Overview | Page 13: Claim Resolution and Final Summary | Page 4: Hospitalization and Diagnosis

Test summary_04 (Score: 100.0% ✓)

Question: Who was determined to be at fault and why?

Agent Answer:

Mr. Robert Chen was determined to be at fault for the collision due to his failure to obey a traffic control device, specifically running a red light. This determination was supported by multiple pieces of evidence, including traffic camera footage, witness testimony from Mr. Marcus Thompson, and a signal timing analysis that confirmed a 2.277-second red light violation. The Seattle Police issued a traffic citation to Chen, and State Farm Insurance accepted 100% liability for the incident, confirming their responsibility for the damages incurred.

Ground Truth:

Robert Chen was determined to be 100% at fault. Evidence showed Chen ran a red light at 45 mph, entered the intersection 2.3 seconds after the light turned red, and failed to yield right of way. Video forensics, skid mark analysis, and witness testimony confirmed Chen violated traffic laws. Mitchell was following all regulations and had right of way.

MODEL GRADER (100.0%)

Comprehensiveness	100.0%
-------------------	--------

Coherence	100.0%
Synthesis	100.0%
Relevance	100.0%
Accuracy	100.0%

Reasoning: The agent's summary thoroughly covers all major points from the reference, is well-organized, integrates information cohesively, remains highly relevant, and presents accurate facts.

Sources:

Page 1: Claim Introduction and Overview | Page 13: Claim Resolution and Final Summary | Page 10: Legal Documentation and Liability Determination

Test summary_05 (Score: 60.0% X)

Question: What was the final resolution of the insurance claim?

Agent Answer:

The final resolution of Claim ID CLM-2024-00789-AUTO, filed by Sarah Mitchell, occurred on February 5, 2024. The claim, resulting from a multi-vehicle collision on January 15, 2024, was settled for a total of \$47,850, which included \$44,010 for vehicle damage and \$3,840 for medical expenses. State Farm Insurance accepted full liability for the incident, and payments were scheduled to ensure Ms. Mitchell received her deductible refund by February 5, 2024. The vehicle repairs were completed satisfactorily by January 30, 2024, and all financial obligations were met within the policy limits.

Ground Truth:

The claim was resolved on February 5, 2024, with total payout of \$47,850. Mitchell received \$43,510 for vehicle damage (after \$500 deductible) and \$3,840 for medical expenses. The vehicle was repaired and passed quality inspection (certificate WAR-2024-0156). Mitchell received full medical clearance and completed all therapy. The claim was processed within 21 days with all parties satisfied.

MODEL GRADER (60.0%)

Comprehensiveness	50.0%
Coherence	100.0%
Synthesis	50.0%
Relevance	100.0%
Accuracy	50.0%

Reasoning: The agent's summary covers some key points but has inaccuracies in the vehicle damage payout and lacks details about the quality inspection and therapy completion, affecting comprehensiveness and accuracy.

Sources:

Page 1: Claim Introduction and Overview | Page 13: Claim Resolution and Final Summary | Page 8: Comprehensive Financial Analysis

Test summary_06 (Score: 100.0% ✓)

Question: What were the key findings from the accident investigation?

Agent Answer:

The accident investigation revealed that Mr. Robert Chen ran a red light, entering the intersection at 45 mph while Sarah Mitchell was traveling northbound at 25 mph with a green signal. Eyewitness testimony from Mr. Marcus Thompson and traffic signal analysis confirmed a 2.277-second violation of the red light by Chen. The Seattle Police issued a traffic citation to Chen for failing to obey a traffic control device, and State Farm Insurance accepted 100% liability for the accident. This determination was supported by traffic camera footage and corroborating witness statements, establishing clear fault on Chen's part.

Ground Truth:

The investigation found Robert Chen ran a red light 2.3 seconds after it turned red, traveling at 45 mph in a 30 mph zone. Skid marks measured 47 feet on wet pavement. Video footage and witness testimony confirmed Chen's traffic violations. Weather conditions (42°F, wet roads from overnight rain) contributed to stopping distance but fault was entirely Chen's for running the red light.

MODEL GRADER (100.0%)

Comprehensiveness	100.0%
-------------------	--------

Coherence	100.0%
Synthesis	100.0%
Relevance	100.0%
Accuracy	100.0%

Reasoning: The agent's summary thoroughly covers all major points from the reference, is well-organized, integrates information effectively, remains highly relevant, and presents accurate facts.

Sources:

Page 1: Claim Introduction and Overview | Page 13: Claim Resolution and Final Summary | Page 7: Witness Testimony and Traffic Engineering

Test summary_07 (Score: 60.0% X)

Question: Describe the emergency response process.

Agent Answer:

On January 15, 2024, at 09:31 AM, the Seattle Emergency Communications Center received a distress call regarding the multi-vehicle collision involving Sarah Mitchell's 2022 Honda Accord. The Seattle Fire Department and King County Medic Unit 47 responded promptly, arriving at the scene to assess Ms. Mitchell, who was found to have elevated blood pressure and acute cervical pain. Officers from the Seattle Police Department secured the area and confirmed a traffic signal violation by another driver involved in the incident, which was crucial for understanding the circumstances of the accident. This coordinated emergency response ensured that medical assistance was provided swiftly and that the scene was managed effectively.

Ground Truth:

Emergency responders arrived within 8 minutes of the 09:23:45 AM collision. Medic Unit 47 with Paramedic Jennifer Ross assessed Mitchell at the scene (BP: 145/92, HR: 98 bpm, pain: 6/10). She was transported to Seattle Medical Center at 09:31:22 AM. Officer James Wilson secured the scene and documented evidence including skid marks, vehicle positions, and witness statements.

MODEL GRADER (60.0%)

Comprehensiveness	50.0%
Coherence	100.0%
Synthesis	50.0%
Relevance	100.0%
Accuracy	50.0%

Reasoning: The agent's summary covers some key points but misses specific details like the exact timing of the response and the pain scale. It is well-organized and relevant but lacks full accuracy and synthesis.

Sources:

Page 1: Claim Introduction and Overview | Page 13: Claim Resolution and Final Summary | Page 3: Emergency Response and Triage

Test summary_08 (Score: 50.0% X)

Question: What was the timeline from collision to claim resolution?

Agent Answer:

The timeline from the collision to claim resolution for Sarah Mitchell's insurance claim is as follows: - **January 15, 2024**: The multi-vehicle collision occurred at 09:23 AM, resulting in significant property damage and minor injuries. - **January 15, 2024**: Ms. Mitchell was evaluated for injuries at Seattle Medical Center later that day. - **January 17, 2024**: The Senior Claims Adjuster finalized the financial reserving and payment schedule, certifying the total claim value at \$47,850. - **January 20, 2024**: The Subrogation Department initiated recovery procedures against State Farm Insurance. - **February 2, 2024**: Payment for vehicle repairs was scheduled. - **February 5, 2024**: The claim was officially closed, marking the resolution of the claim process. Overall, the entire process took approximately three weeks from the date of the collision to the claim's closure.

Ground Truth:

The claim spanned January 15 to February 5, 2024 (21 days). Key milestones: collision (Jan 15, 09:23 AM), emergency response (8 minutes later), hospital admission (Jan 15), damage assessment (Jan 16), therapy start (Jan 22), repair completion (Jan 30), medical clearance (Feb 3), and final resolution (Feb 5).

MODEL GRADER (50.0%)

Comprehensiveness	50.0%
Coherence	50.0%
Synthesis	50.0%
Relevance	100.0%
Accuracy	50.0%

Reasoning: The agent's summary covers some key points but misses several milestones and has inaccuracies. It is somewhat organized but lacks a cohesive narrative, affecting overall clarity.

Sources:

Page 1: Claim Introduction and Overview | Page 13: Claim Resolution and Final Summary | Page 8: Comprehensive Financial Analysis

Test summary_09 (Score: 60.0% X)

Question: Describe the vehicle damage and repair process.

Agent Answer:

The vehicle damage to Sarah Mitchell's 2022 Honda Accord was assessed following a multi-vehicle collision on January 15, 2024. The inspection revealed critical unibody misalignment, particularly a 3.2-inch inward deformation of the driver's side A-pillar, leading to a total repair estimate of \$23,102.50. Premier Auto Body Shop initiated parts procurement on January 19, ordering 42 items valued at \$12,450.00, including expedited shipping for a critical backordered part. Repairs were completed on January 30, 2024, with a final quality assurance check confirming the vehicle's structural integrity and paint finish met factory standards, resulting in Ms. Mitchell's satisfaction with the work performed.

Ground Truth:

Sarah Mitchell's Honda Accord sustained driver's side door damage, frame misalignment, and side airbag deployment. Premier Auto Body Shop performed repairs including door replacement, frame straightening, paint matching, and safety system restoration. Repairs cost \$44,010, took approximately 10 days, and passed quality inspection (WAR-2024-0156). Mitchell paid \$500 deductible.

MODEL GRADER (60.0%)

Comprehensiveness	50.0%
Coherence	100.0%
Synthesis	50.0%
Relevance	100.0%
Accuracy	50.0%

Reasoning: The agent's summary covers some key points but misses critical details like the cost and deductible. It is well-organized and relevant, but the accuracy of the repair estimate and synthesis of information could be improved.

Sources:

Page 1: Claim Introduction and Overview | Page 13: Claim Resolution and Final Summary | Page 12: Repair Completion and Quality Assurance

Test summary_10 (Score: 50.0% X)

Question: Summarize the evidence collected at the accident scene.

Agent Answer:

The evidence collected at the accident scene included eyewitness testimony from Mr. Marcus Thompson, who reported that Mr. Chen ran a red light, which was corroborated by traffic signal timing analysis indicating a 2.277-second violation. Additionally, the Seattle Police Department conducted a video analysis that confirmed Mr. Chen's failure to yield the right of way. Traffic camera footage further supported the findings, establishing clear liability for the accident. First responders also documented Ms. Sarah Mitchell's condition, noting elevated blood pressure and acute cervical pain, which contributed to the medical assessment following the collision.

Ground Truth:

Evidence included: 47-foot skid marks from Chen's vehicle, traffic camera footage showing red light violation, witness testimony from Marcus Thompson, vehicle position documentation, weather conditions (42°F, wet pavement), traffic signal timing data, and physical measurements. Officer James Wilson documented all evidence. Engineering analysis confirmed Chen entered intersection 2.3 seconds after red light.

MODEL GRADER (50.0%)

Comprehensiveness	50.0%
Coherence	50.0%
Synthesis	50.0%
Relevance	50.0%
Accuracy	50.0%

Reasoning: The agent's summary covers some key points but misses others, such as specific evidence like skid marks and weather conditions. It is somewhat organized but lacks a cohesive flow and includes some irrelevant details.

Sources:

Page 1: Claim Introduction and Overview | Page 13: Claim Resolution and Final Summary | Page 4: Hospitalization and Diagnosis

Test summary_11 (Score: 60.0% X)

Question: What was the subrogation process outcome?

Agent Answer:

The subrogation process for Claim ID CLM-2024-00789-AUTO involved Progressive Auto Insurance seeking recovery from State Farm Insurance for a total of \$48,350, which included \$47,850 in damages and a \$500 deductible for Sarah Mitchell. State Farm acknowledged full liability on January 23, 2024, and agreed to issue payments of \$40,000 by February 1, 2024, and \$8,350 by February 15, 2024. Additionally, Mitchell was set to receive her deductible refund by February 5, 2024, completing the financial recovery process efficiently.

Ground Truth:

Progressive Insurance pursued subrogation against Robert Chen's insurer, State Farm. Due to Chen's 100% fault determination with clear evidence (video, witnesses, police report), State Farm accepted full liability. Progressive recovered the full claim amount of \$47,850 plus processing costs. The subrogation was completed efficiently without legal dispute.

MODEL GRADER (60.0%)

Comprehensiveness	50.0%
Coherence	100.0%
Synthesis	50.0%
Relevance	100.0%
Accuracy	50.0%

Reasoning: The agent's summary covers some key points but includes additional details not present in the reference, affecting comprehensiveness and accuracy. However, it is well-organized and relevant.

Sources:

Page 1: Claim Introduction and Overview | Page 13: Claim Resolution and Final Summary | Page 11: Subrogation Process and Recovery

Test summary_12 (Score: 20.0% X)

Question: Describe the weather and road conditions during the collision.

Agent Answer:

The claim documents do not provide specific information regarding the weather and road conditions at the time of the collision on January 15, 2024. However, the incident occurred during daylight hours at 09:23 AM, which typically suggests visibility was adequate. There is no mention of adverse weather conditions such as rain, snow, or fog that could have contributed to the accident. For a complete assessment, additional details on weather conditions would be necessary.

Ground Truth:

The collision occurred during morning rush hour with ambient temperature of 42°F. Roads were wet from overnight rainfall accumulating 0.3 inches, creating slippery conditions. Visibility was reduced to 200 feet. The wet pavement contributed to longer stopping distances, evidenced by 47-foot skid marks. However, weather was not a primary factor in fault determination as Chen ran a red light.

MODEL GRADER (20.0%)

Comprehensiveness	0.0%
Coherence	50.0%
Synthesis	0.0%

Relevance	50.0%
Accuracy	50.0%

Reasoning: The agent's summary lacks key details about weather and road conditions, making it incomplete. While it is somewhat organized, it fails to synthesize information effectively and does not fully address the question.

Sources:

Page 1: Claim Introduction and Overview | Page 13: Claim Resolution and Final Summary | Page 2: Initial Collision Dynamics

Test summary_13 (Score: 72.0% ✓)

Question: What were the total costs broken down by category?

Agent Answer:

The total costs for Claim ID CLM-2024-00789-AUTO, filed by Sarah Mitchell, were broken down into two main categories: vehicle damage and medical expenses. The vehicle damage amounted to \$44,010.00, which included a repair estimate of \$23,102.50 for the Honda Accord, covering OEM parts and labor. Additionally, medical expenses totaled \$3,840.00, bringing the overall claim value to \$47,850.00.

Ground Truth:

Total claim value: \$47,850. Vehicle repairs: \$44,010 (parts, labor, paint). Medical expenses: \$3,840 (emergency care, CT scan, 12 physical therapy sessions, physician consultations). Mitchell paid \$500 deductible. Progressive paid \$43,510 for vehicle and \$3,840 for medical. All costs were recovered via subrogation from State Farm.

MODEL GRADER (72.0%)

Comprehensiveness	50.0%
Coherence	100.0%
Synthesis	50.0%
Relevance	100.0%
Accuracy	100.0%

Reasoning: The agent's summary covers key points but omits details about the deductible and subrogation, affecting comprehensiveness and synthesis. However, it is coherent and accurate.

Sources:

Page 1: Claim Introduction and Overview | Page 13: Claim Resolution and Final Summary | Page 5: Technical Damage Assessment

Test summary_14 (Score: 60.0% X)

Question: Describe the medical diagnosis and treatment journey.

Agent Answer:

Sarah Mitchell was diagnosed with a Grade 2 Whiplash Associated Disorder and a mild concussion following a multi-vehicle collision on January 15, 2024. She was evaluated in the emergency room at Seattle Medical Center, where a CT scan confirmed no acute injuries, and her medical expenses from this visit totaled \$3,840. On January 22, 2024, she attended a follow-up consultation with Dr. Patterson, who noted ongoing cervical spine tenderness and limited range of motion. As a result, she was referred for 12 sessions of physical therapy at Seattle Rehabilitation Center, where her initial assessment indicated a moderate disability level.

Ground Truth:

Mitchell sustained Grade 2 whiplash injury, cervical spine strain, and mild concussion from the collision impact. Treatment included emergency room evaluation, CT scan to rule out serious injury, pain medication (Ibuprofen 600mg, Cyclobenzaprine 10mg), and 12 physical therapy sessions over 6 weeks. Dr. Michael Patterson provided care and final clearance. Pain decreased from 6/10 to 4/10 with full recovery expected.

MODEL GRADER (60.0%)

Comprehensiveness	50.0%
Coherence	100.0%

Synthesis	50.0%
Relevance	100.0%
Accuracy	50.0%

Reasoning: The agent's summary covers some key points but misses details about pain levels and the specific medications. It is well-organized and relevant, but the synthesis of information could be improved.

Sources:

Page 1: Claim Introduction and Overview | Page 13: Claim Resolution and Final Summary | Page 6: Physical Therapy and Rehabilitation

Test summary_15 (Score: 50.0% X)

Question: How long did the entire claim process take and what were the major phases?

Agent Answer:

The entire claim process for Claim ID CLM-2024-00789-AUTO took approximately 21 days, starting from the incident on January 15, 2024, and concluding with the claim closure on February 5, 2024. Major phases included the initial filing and assessment of the accident, which occurred on the same day as the incident, followed by medical evaluation and diagnosis on January 15. Financial analysis and reserving were finalized on January 17, with repair work completed by January 30. The subrogation process began shortly after, with State Farm acknowledging liability and agreeing to payment terms by January 23, leading to the claim's resolution.

Ground Truth:

The claim process took 21 calendar days (January 15 - February 5, 2024). Major phases: (1) Collision and emergency response (Jan 15), (2) Medical assessment and damage evaluation (Jan 15-17), (3) Treatment and repairs in parallel (Jan 19-30), (4) Medical clearance and quality inspection (Jan 30-Feb 3), (5) Final resolution and payment (Feb 5). Progressive met company service standards for processing time.

MODEL GRADER (50.0%)

Comprehensiveness	50.0%
Coherence	50.0%
Synthesis	50.0%
Relevance	100.0%
Accuracy	50.0%

Reasoning: The agent's summary covers some major points but misses key phases and has inaccuracies. It is somewhat organized but lacks a cohesive flow, making it less comprehensive overall.

Sources:

Page 1: Claim Introduction and Overview | Page 13: Claim Resolution and Final Summary | Page 8: Comprehensive Financial Analysis

Routing Agent Tests

Summary: 10 tests | 10 correct (100.0%) | Routing Accuracy: 100.0%

Test routing_01 (✓)

Question: What time did the collision occur?

Expected Route:	Needle
Actual Route:	Needle
Result:	CORRECT ✓

Test routing_02 (✓)

Question: Summarize the events that led to the claim.

Expected Route:	Summary
Actual Route:	Summary
Result:	CORRECT ✓

Test routing_03 (✓)

Question: What was Sarah Mitchell's blood pressure?

Expected Route:	Needle
Actual Route:	Needle
Result:	CORRECT ✓

Test routing_04 (✓)

Question: What was the total estimated claim value?

Expected Route:	Summary
Actual Route:	Summary
Result:	CORRECT ✓

Test routing_05 (✓)

Question: Who was the attending physician?

Expected Route: Needle

Actual Route: Needle

Result: CORRECT ✓

Test routing_06 (✓)

Question: Describe the emergency response process.

Expected Route:	Summary
Actual Route:	Summary
Result:	CORRECT ✓

Test routing_07 (✓)

Question: What was the claim number?

Expected Route:	Needle
Actual Route:	Needle
Result:	CORRECT ✓

Test routing_08 (✓)

Question: Who was determined to be at fault and why?

Expected Route:	Summary
Actual Route:	Summary
Result:	CORRECT ✓

Test routing_09 (✓)

Question: How many feet were the skid marks?

Expected Route:	Needle
Actual Route:	Needle
Result:	CORRECT ✓

Test routing_10 (✓)

Question: What was the timeline from collision to claim resolution?

Expected Route:	Summary
------------------------	---------

Actual Route: Summary

Result: CORRECT ✓

Human-in-the-Loop (HITL) Tests

Summary: 15 tests reviewed | Average Rating: 4.20/5 | Average Score: 80.0%

Test hitl_needle_01 (Rating: 4/5)

Question: What injuries did Sarah Mitchell sustain in the collision?

Agent Answer: Sarah Mitchell sustained signs of deceleration trauma, self-reporting neck pain measuring 6/10 on the visual analog pain scale, along with dizziness and spatial disorientation following the collision. Additionally, a clinical re-evaluation during her follow-up consultation on January 22, 2024, noted persistent cervical spine tenderness.

Rating: 4/5

Score: 75.0%

Query Type: Needle

Feedback: OK

Test hitl_needle_02 (Rating: 5/5)

Question: When and where did the collision take place?

Agent Answer: The collision took place on January 15, 2024, at 09:23:45 AM PST at the intersection of Maple Avenue and 5th Street in Seattle, Washington 98101.

Rating: 5/5

Score: 100.0%

Query Type: Needle

Feedback: good

Test hitl_needle_03 (Rating: 4/5)

Question: What was the weather like during the accident?

Agent Answer: The weather during the accident was characterized by wet road surfaces due to overnight rainfall accumulating 0.3 inches. The temperature was 42 degrees Fahrenheit, and visibility was reduced to 200 feet.

Rating: 4/5

Score: 75.0%

Query Type: Needle

Feedback: good

Test hitl_needle_04 (Rating: 3/5)

Question: Who were the key people involved in handling this claim?

Agent Answer: The key people involved in handling this claim include: 1. **Michael Torres** - Claims Representative from State Farm, a senior liability adjuster with 10 years of experience, assigned to investigate the liability exposure. 2. **Linda Martinez** - Claims Adjuster from Progressive Insurance, who communicated with Torres regarding liability determination and damage assessment. 3. **David Thompson** - Regional Claims Manager who closed the claim on February 5, 2024. 4. **Jennifer Wu** - Claims Manager at State Farm who signed the Liability Acceptance Letter on January 18, 2024. Additionally, **Sarah Mitchell** is the policyholder, and **Mr. Robert Chen** is the insured party under State Farm.

Rating:	3/5
Score:	50.0%
Query Type:	Needle
Feedback:	ok

Test hitl_needle_05 (Rating: 4/5)

Question: What were the vital signs measured at the accident scene?

Agent Answer: The initial set of vital signs recorded at the accident scene at 09:34:15 indicated physiological stress: Blood Pressure was elevated at 145/92 mmHg, Heart Rate was Tachycardic at 98 beats per minute, Respiratory Rate was 18 breaths per minute, and Oxygen Saturation (SpO2) was 98 percent on ambient air. Body temperature was 98.4°F.

Rating:	4/5
Score:	75.0%
Query Type:	Needle
Feedback:	good

Test hitl_summary_01 (Rating: 3/5)

Question: Provide an overview of the entire insurance claim from start to finish.

Agent Answer: The insurance claim CLM-2024-00789-AUTO was initiated by Sarah Mitchell on January 15, 2024, following a multi-vehicle collision involving her 2022 Honda Accord at the intersection of Maple Avenue and 5th Street in Seattle. The accident resulted in significant property damage estimated at \$47,850, which included \$44,010 for vehicle repairs and \$3,840 for medical expenses related to her diagnosis of Grade 2 Whiplash and a mild concussion. The claim was processed efficiently, with Senior Claims Adjuster Linda Martinez finalizing the payment schedule by January 17, 2024. Repairs were completed by Premier Auto Body by January 30, 2024, and were validated by a Progressive Field Inspector. The claim was closed on February 5, 2024, after Progressive initiated subrogation against State Farm Insurance, which acknowledged liability and agreed to reimburse the total claim amount, ensuring Mitchell received her deductible refund by

February 5, 2024.

Rating: 3/5

Score: 50.0%

Query Type: Summary

Feedback: good

Test hitl_summary_02 (Rating: 3/5)

Question: What factors contributed to the accident and how was fault determined?

Agent Answer: The accident involving Sarah Mitchell and Robert Chen was primarily caused by Chen running a red light, as confirmed by eyewitness testimony and traffic signal analysis. Mitchell was traveling northbound at 25 mph with a green signal when Chen, driving at 45 mph, entered the intersection after the light had turned red. The Seattle Police issued a traffic citation to Chen for failing to obey traffic control, and State Farm Insurance accepted full liability based on clear evidence from traffic camera footage and witness statements. This comprehensive investigation established that Chen's failure to yield the right of way was the critical factor in determining fault for the collision.

Rating:	3/5
Score:	50.0%
Query Type:	Summary
Feedback:	ok

Test hitl_summary_03 (Rating: 4/5)

Question: Describe Sarah Mitchell's journey from injury through recovery.

Agent Answer: Sarah Mitchell's journey began with a multi-vehicle collision on January 15, 2024, resulting in significant property damage to her 2022 Honda Accord and minor injuries, including a Grade 2 Whiplash Associated Disorder and a mild concussion. After being evaluated at Seattle Medical Center, where she incurred an emergency room bill of \$3,840, she was referred for physical therapy due to persistent neck pain and limited mobility. On January 22, she began physical therapy sessions at Seattle Rehabilitation Center, where her condition was assessed as 'Moderate Disability' on the Neck Disability Index. By January 30, her vehicle repairs were completed, and she expressed satisfaction with the quality of the work, marking a significant step towards her recovery.

Rating:	4/5
Score:	75.0%
Query Type:	Summary
Feedback:	good

Test hitl_summary_04 (Rating: 4/5)

Question: How did the insurance companies handle the subrogation process?

Agent Answer: The subrogation process for Claim ID CLM-2024-00789-AUTO was initiated by Progressive Auto Insurance on January 20, 2024, led by Specialist David Chen. They sought to recover a total of \$48,350 from State Farm Insurance, which included \$47,850 for damages and a \$500 deductible for the policyholder, Sarah Mitchell. State Farm acknowledged full liability on January 23, 2024, and agreed to a payment schedule, committing to pay \$40,000 by February 1, 2024, and the remaining \$8,350 by February 15, 2024. This arrangement ensured that Mitchell would receive her deductible refund by February 5, 2024.

Rating: 4/5
Score: 75.0%
Query Type: Summary
Feedback: good

Test hitl_summary_05 (Rating: 4/5)

Question: What was the financial impact of this claim on all parties involved?

Agent Answer: The financial impact of Claim ID CLM-2024-00789-AUTO primarily affected the policyholder, Sarah Mitchell, and the at-fault party's insurer, State Farm. The total estimated claim value was \$47,850, which included \$44,010 for vehicle damage and \$3,840 for medical expenses incurred by Mitchell due to her injuries from the accident. State Farm accepted full liability and agreed to pay these damages, with a payment schedule of \$40,000 by February 1, 2024, and \$8,350 by February 15, 2024. Additionally, Mitchell was set to receive a refund of her \$500 deductible by February 5, 2024, while Progressive Auto Insurance managed the subrogation process to recover the total claim amount from State Farm.

Rating: 4/5

Score: 75.0%

Query Type: Summary

Feedback: good

Test hitl_routing_01 (Routing: CORRECT)

Question: What were Sarah Mitchell's initial symptoms after the accident?

Routing Decision: NEEDLE

Human Feedback: CORRECT ✓

Score: 100.0%

Test hitl_routing_02 (Routing: CORRECT)

Question: Explain the overall progression of medical treatment and its outcomes.

Routing Decision: SUMMARY

Human Feedback: CORRECT ✓

Score: 100.0%

Test hitl_routing_03 (Routing: CORRECT)

Question: What was the exact settlement amount?

Routing Decision: NEEDLE

Human Feedback: CORRECT ✓

Score: 100.0%

Test hitl_routing_04 (Routing: CORRECT)

Question: Compare the initial claim assessment with the final outcome and discuss key differences.

Routing Decision: SUMMARY

Human Feedback: CORRECT ✓

Score: 100.0%

Test hitl_routing_05 (Routing: CORRECT)

Question: Who was the treating physician's name and contact information?

Routing Decision: NEEDLE

Human Feedback: CORRECT ✓

Score: 100.0%