```
╭──────────────────────────────────────────────────────╮
│ 🚀 Running Evaluation Mode: ALL                       │
╰──────────────────────────────────────────────────────╯
⚙️ Initializing Agent...

=== Running Hard Evals (Guardrails) ===
Running Hard Evals...

[1/17] Checking hard_001
Query: What is the deductible amount? (Only extract the
value)
Status: PASS | Router: needle
╭─ Agent Answer ─╮
│ $1,000.00      │
╰────────────────╯

[2/17] Checking hard_002
Query: Who is the policyholder? (Return JSON: {"name":
"..."})
Status: PASS
| Router:
needle
╭─ Agent A─╮
│ {"name": │
│ "Alex    │
│ Johnson" │
│ }        │
╰──────────╯

[3/17]
Checking
hard_003
Query: List
the claim
ID.
Status: PASS | Router: needle
╭──────── Agent Answer ────────╮
│ The claim ID is HO-2024-8892. │
╰───────────────────────────────╯

[4/17] Checking hard_004
Query: What is the total payout limit? Format: $XX,XXX
Status: PASS | Router: needle
╭─────────── Agent Answer ───────────╮
│ The total payout limit is $19,550.00. │
╰──────────────────────────────────────╯

[5/17] Checking hard_005
Query: What time did the 'ABNORMAL START' occur? (HH:MM:SS AM/PM)
Status: PASS | Router: needle
╭─────────────── Agent Answer ───────────────╮
│ The 'ABNORMAL START' occurred at 10:22:15 AM. │
╰──────────────────────────────────────────────╯

[6/17] Checking hard_006
Query: Who is the Senior Property Adjuster?
Status: PASS | Router: needle
╭────────────── Agent Answer ──────────────╮
│ The Senior Property Adjuster is Mike Ross. │
╰────────────────────────────────────────────╯

[7/17] Checking hard_007
Query: First find the incident location (City, State) and time from the documents. Then, what was the
ambient temperature (Amb Temp) at The Location at The Time? (Numeric only)
Status: PASS | Router: needle
╭───────────────────────────────────── Agent Answer ──────────────────────────────────────╮
│ The ambient temperature (Amb Temp) in Austin, TX on November 16, 2024, ranged between 12.1 °C and │
│ 27.0 °C. However, without a specific time of day, the exact ambient temperature at "The Time" cannot │
│ be given as a single numeric value.                                                        │
╰────────────────────────────────────────────────────────────────────────────────────────────╯

[8/17] Checking hard_008
Query: What is the Vendor ID for DryFast Inc?
Status: PASS | Router: needle
╭───────────── Agent Answer ─────────────╮
│ The Vendor ID for DryFast Inc is V-9982. │
```

~~~
[9/17] Checking hard_009
Query: Return the list of covered items as a JSON
array.
Status: PASS | Router: needle
                  Agent Answer
 The list of covered items is as follows:

 ```json
 [
     "Rug",
     "TV",
     "Sofa"
 ]
 ~~~

[10/17] Checking hard_010
Query: What city is the address 124 Maple Street
located in?
Status: PASS | Router: needle
                     Agent Answer
 The address 124 Maple Street is located in
 Austin.

[11/17] Checking hard_011
Query: What is the Licensed Plumber's number (Lic
#)?
Status: PASS | Router: needle
                     Agent Answer
 The Licensed Plumber's number (Lic #) is 4482.

[12/17] Checking hard_012
Query: Is the claim approved? (Yes/No)
Status: PASS | Router: needle
              Agent Answer
 Yes, the claim is approved.

[13/17] Checking hard_013
Query: First find the address of the incident
property. Then extract the state code from that
address.
Status: PASS | Router: needle
                     Agent Answer
 The state code from the incident property
 address is "TX" (Texas).

[14/17] Checking hard_014
Query: First identify the sensor or location for
the 10:45:00 AM reading. Then, what was the Audio
Level detected at that time?
Status: PASS | Router: needle
                                          Agent Answer
 The sensor or location for the 10:45:00 AM reading was Cam_LvRm_02, and the Audio Level detected at
 that time was 45 dB.

[15/17] Checking hard_015
Query: Is this document marked confidential? (Answer Yes or No)
Status: FAIL | Router: needle
                  Agent Answer
 Yes, the document is marked confidential.

   Reason: Failed to match regex: (?i)No

[16/17] Checking hard_016
Query: What is the coverage type?
Status: PASS | Router: needle
                             Agent Answer
 The coverage type is an HO-3 'Special Form' Homeowners Policy.

[17/17] Checking hard_017
```
```

Query: What is the Flow Meter ID?
Status: PASS | Router: needle

Agent Answer

The Flow Meter ID is Flow_Meter_01.

🛡️ Hard Evals (Guardrails) – Summary

ID	Type	Router	Status	Actual (Truncated)	Error
hard_001	regex	needle	**PASS**	$1,000.00	
hard_002	json	needle	**PASS**	{"name": "Alex Johnson"}	
hard_003	regex	needle	**PASS**	The claim ID is HO-2024-8892.	
hard_004	regex	needle	**PASS**	The total payout limit is $19,550.00.	
hard_005	regex	needle	**PASS**	The 'ABNORMAL START' occurred at 10:22:15 AM.	
hard_006	regex	needle	**PASS**	The Senior Property Adjuster is Mike Ross.	
hard_007	regex	needle	**PASS**	The ambient temperature (Amb Temp) in Austin, TX o...	
hard_008	regex	needle	**PASS**	The Vendor ID for DryFast Inc is V-9982.	
hard_009	json	needle	**PASS**	The list of covered items is as follows: ```json ...	
hard_010	regex	needle	**PASS**	The address 124 Maple Street is located in Austin.	
hard_011	regex	needle	**PASS**	The Licensed Plumber's number (Lic #) is 4482.	
hard_012	regex	needle	**PASS**	Yes, the claim is approved.	
hard_013	regex	needle	**PASS**	The state code from the incident property address ...	
hard_014	regex	needle	**PASS**	The sensor or location for the 10:45:00 AM reading...	
hard_015	regex	needle	**FAIL**	Yes, the document is marked confidential.	Failed to match regex: (?i)No
hard_016	regex	needle	**PASS**	The coverage type is an HO-3 'Special Form' Homeow...	
hard_017	regex	needle	**PASS**	The Flow Meter ID is Flow_Meter_01.	

Summary: **16**/**17** passed (**94.1**%)

=== Running LLM-as-a-Judge ===

🔍 **Query:** What was the date of the incident?
Using Tool: **needle**

🤖 **Agent Answer**

The date of the incident was November 16, 2024.

⚖️ Judge Results

Metric	Score	Explanation
Correctness	1	The actual answer contains the core correct fact, which is the date of the incident: November 16, 2024.
Relevancy	1	The agent answered the specific question asked by providing the correct date of the incident, which matches the expected answer. The information is relevant and contains the specific detail requested.
Recall	1	The Actual Answer contains all the key facts, numbers, dates, and entities present in the Expected Answer. Both specify the date of the incident as November 16, 2024.

🔍 **Query:** What is the total repair estimate cost?
Using Tool: **needle**

🤖 **Agent Answer**

The total repair estimate cost is $12,400.00.

⚖️ Judge Results

Metric	Score	Explanation

Correctness	1	The actual answer contains the core correct fact from the expected answer, which is the total repair estimate cost of $12,400.00.
Relevancy	1	The agent answered the specific question asked by providing the total repair estimate cost of $12,400.00, which matches the expected answer. The response is relevant and contains the specific detail requested.
Recall	1	The Actual Answer contains all the key facts and numbers present in the Expected Answer, specifically the total repair estimate cost of $12,400.00.

🔍 **Query:** Who is the policyholder?
Using Tool: **needle**

🤖 **Agent Answer**

The policyholder is Alex Johnson.

⚖️ *Judge Results*

Metric	Score	Explanation
Correctness	1	The actual answer contains the core correct fact that the policyholder is Alex Johnson, which matches the expected answer.
Relevancy	1	The agent answered the specific question asked by stating 'The policyholder is Alex Johnson,' which matches the expected answer. The response is relevant and contains the specific detail required, confirming it retrieved the correct information.
Recall	1	The Actual Answer contains all the key facts and entities present in the Expected Answer. Both mention 'Alex Johnson' as the policyholder, fulfilling the criteria for a correct response.

🔍 **Query:** What was the Total Vol recorded by Flow_Meter_01 at 11:15:00 AM?
Using Tool: **needle**

🤖 **Agent Answer**

The Total Vol recorded by Flow_Meter_01 at 11:15:00 AM was 448.5 gallons.

⚖️ *Judge Results*

Metric	Score	Explanation
Correctness	1	The actual answer contains the core correct facts from the expected answer, including the Total Vol value of 448.5 and the time of 11:15:00 AM. The use of 'gallons' instead of 'Gal' does not change the correctness of the information.
Relevancy	1	The agent answered the specific question asked by providing the Total Vol recorded by Flow_Meter_01 at 11:15:00 AM as 448.5 gallons. The answer contains specific details, such as the time and the volume, which implies it retrieved the correct information. The information is relevant to the query context.
Recall	1	The Actual Answer contains all the key facts, numbers, and entities present in the Expected Answer. The term 'Gal' in the Expected Answer is correctly interpreted as 'gallons' in the Actual Answer, which does not alter the meaning.

🔍 **Query:** Summarize the claim timeline. Include all dates, dollar amounts, and company names involved.
Using Tool: **summary**

🤖 **Agent Answer**

Here is a summary of the claim timeline:

- **Incident Date**: November 16, 2024
- **Emergency Mitigation**: Conducted by DryFast Inc.
- **Claim Adjustment**: Handled by Mike Ross
- **Total Payout**: $19,550.00
- **Restoration**: Completed by Austin Home Restorations LLC

These are the key dates, dollar amounts, and companies involved in the claim.

⚖️ *Judge Results*

Metric	Score	Explanation
Correctness	0	The actual answer does not include the valve shutoff on the same day as the incident or the inspection on November 18, which are core facts from the expected answer. Additionally, it introduces new information not present in the expected answer, such as the involvement of DryFast Inc., Mike Ross, the total payout, and Austin Home Restorations LLC.
Relevancy	0	The agent's response did not directly address the specific question asked, which was to summarize the claim timeline with all dates, dollar amounts, and

		company names involved. The expected answer focused on the incident date, valve shutoff, and inspection date, while the agent provided additional information not requested, such as emergency mitigation, claim adjustment, total payout, and restoration details. This indicates a failure to retrieve the correct claim documents/segments relevant to the query context.
Recall	0	The Actual Answer does not contain all the key facts present in the Expected Answer. Specifically, it omits the valve shutoff on the same day as the incident and the inspection on November 18. Additionally, it includes extra information not present in the Expected Answer, such as company names and a payout amount.

🔍 **Query:** What was the specific model of the TV claimed?
Using Tool: **needle**

🤖 **Agent Answer**

The specific model of the TV claimed is the Samsung QN90C Series.

⚖️ *Judge Results*

Metric	Score	Explanation
Correctness	1	The actual answer contains the core correct facts from the expected answer, as it correctly identifies the TV model as the Samsung QN90C Series.
Relevancy	1	The agent answered the specific question asked by providing the exact model of the TV, 'Samsung QN90C Series,' which matches the expected answer. The response is relevant and contains the specific detail requested.
Recall	1	The Actual Answer contains all the key facts and entities present in the Expected Answer, specifically mentioning the 'Samsung QN90C Series' model.

🔍 **Query:** Was the sofa replacement approved fully or partially?
Using Tool: **needle**

🤖 **Agent Answer**

The sofa replacement was approved partially with an amount of $250.00.

⚖️ *Judge Results*

Metric	Score	Explanation
Correctness	1	The actual answer contains the core correct fact that the sofa replacement was partially approved, which matches the expected answer.
Relevancy	1	The agent answered the specific question asked by stating that the sofa replacement was approved partially. It also provided additional detail about the amount approved, which implies it retrieved the correct claim documents. The information is relevant to the query context.
Recall	1	The Actual Answer contains all the key facts present in the Expected Answer. Both indicate that the sofa replacement was approved partially. The additional detail of the amount ($250.00) in the Actual Answer does not contradict the Expected Answer.

🔍 **Query:** What was the deductible amount?
Using Tool: **needle**

🤖 **Agent Answer**

The deductible amount was $1,000.00.

⚖️ *Judge Results*

Metric	Score	Explanation
Correctness	1	The actual answer contains the core correct fact from the expected answer, which is the deductible amount of $1,000. The inclusion of '.00' does not alter the correctness of the information.
Relevancy	1	The agent answered the specific question asked by providing the deductible amount as $1,000.00, which matches the expected answer of $1,000. The inclusion of '.00' does not affect the correctness or relevancy of the response.
Recall	1	The Actual Answer contains all the key facts and numbers present in the Expected Answer. Both mention the deductible amount as $1,000, with the Actual Answer providing an equivalent representation of the number.

📄 Detailed results saved to **evaluation_results.json**

```
=== Running Human-in-the-Loop ===
╭──────────────────────────────────────────────────────────────────────────────────╮
│ 👨‍🏫 Starting Human-in-the-Loop Grading Session                                   │
╰──────────────────────────────────────────────────────────────────────────────────╯
All cases already graded!
```