

warning "Backend update (October 2026)" GitHub Models inference was retired on July 30, 2026. Any references below are retained only for historical context while this lab is migrated. Use Lab 013 ? Free LLM backends to configure Foundry Local or Ollama before continuing.

Lab 035: Agent Evaluation with Azure AI Eval SDK

Level: L300 Path: Pro Code Time: ~60 min ? Cost: GitHub Free ? uses GitHub Models for evaluation

What You'll Learn

Why LLM-as-judge evaluation works and when to use it Measure groundedness, relevance, coherence, and fluency Build a golden dataset for regression testing Evaluate your RAG agent automatically with azure-ai-evaluation SDK Track quality metrics over time and detect regressions

Introduction

How do you know if your agent got better or worse after a change? Manual testing doesn't scale.

LLM-as-judge evaluation uses a second, independent LLM to score your agent's answers against criteria like:

Groundedness ? Is the answer supported by the retrieved documents? Relevance ? Does the answer address what the user asked? Coherence ? Is the answer logically structured? Fluency ? Is the grammar correct?

This lab builds an automated evaluation pipeline for the OutdoorGear assistant.

Prerequisites

Python 3.11+ pip install azure-ai-evaluation openai GITHUB_TOKEN set

? Supporting Files

note "Download these files before starting the lab" Save all files to a lab-035/ folder in your working directory.

File | Description | Download | eval_dataset.jsonl | Evaluation dataset | ? Download |

Lab Exercise

Step 1: Install the evaluation SDK

[diagram or code sample]

Step 2: Create a golden dataset

A golden dataset is a set of test questions with known-good answers. Create ?

eval_dataset.jsonl:

[diagram or code sample]

Note the 4th entry ? the agent answers a question unrelated to the knowledge base.

Groundedness evaluation should flag this.

Step 3: Run built-in evaluators

[diagram or code sample]

[diagram or code sample]

Expected output: [diagram or code sample]

The Paris question is correctly flagged with low groundedness (1.0) ? the answer is not supported by the OutdoorGear context.

Step 4: Build a custom evaluator

Sometimes you need domain-specific criteria. Here's a custom evaluator for price accuracy:

[diagram or code sample]

Step 5: Automate in CI/CD

Add this to your GitHub Actions workflow (see Lab 037):

[diagram or code sample]

Evaluation Metrics Reference

Metric | What it measures | Good score | Groundedness | Answer supported by context | ? 4.0
| Relevance | Answers the user's question | ? 4.0 | Coherence | Logical, well-structured |
? 4.0 | Fluency | Grammatically correct | ? 4.5 |

Scores are 1-5. Below 3.0 on groundedness usually indicates hallucination.

? Knowledge Check

question "Q1 (Run the Lab): Load lab-035/eval_dataset.jsonl in Python. How many examples are in the dataset, and how many are in the out_of_scope category?"

Run the loader code from the Supporting Files section above.

??? success "? Reveal Answer" 20 total examples. 1 example is in the out_of_scope category.

The dataset has exactly 20 lines. Run `sum(1 for d in dataset if d["category"] ==`

`"out_of_scope")` to confirm there is 1 out-of-scope example. That example tests whether your agent correctly refuses to answer questions unrelated to outdoor gear.

question "Q2 (Run the Lab): For the single out_of_scope example in eval_dataset.jsonl, what is the value of the product_ids field?"

Load the dataset and filter for `category == "out_of_scope"`. Print the `product_ids` field.

??? success "? Reveal Answer" [] (empty list)

The out-of-scope example has `"product_ids": []` because the question is not about any specific product ? it's testing whether the agent refuses to answer irrelevant questions (like asking for cooking recipes). A well-designed agent should return a refusal message rather than hallucinating an answer. Your evaluation metric should check that the agent's groundedness score is high and it does NOT reference any product IDs.

question "Q3 (Multiple Choice): How many examples in eval_dataset.jsonl are in the tents category?"

- A) 3 - B) 5 - C) 7 - D) 4

??? success "? Reveal Answer" Correct: B ? 5 examples

The tents category has 5 examples, making it the largest single category. Run `sum(1 for d in dataset if d["category"] == "tents")` to confirm. The full breakdown: tents(5), backpacks(4), sleeping_bags(3), pricing(3), recommendations(3), out_of_scope(1), comparisons(1).

Next Steps

CI/CD for agents: ? Lab 037 ? GitHub Actions for AI Agents Enterprise RAG evaluation: ? Lab 042 ? Enterprise RAG with Evaluations *(coming soon)*

Phase 3 modernization

warning "Verified 2026-10-27" Use the Microsoft Foundry name and GA Agent Service surfaces based on the Responses API. Evaluations are GA; hosted agents remain preview and must be labeled accordingly. Do not build new dependencies on portal Workflows scheduled for retirement on 2026-12-01.

[diagram or code sample]

Foundry Agent Service GA announcement Foundry Agent Service Foundry evaluation Foundry Workflows retirement

Supporting kit

The site build generates stable PDF and ZIP downloads. The ZIP contains the lab guide and every supporting file declared for this lab.

Download generated PDF Download generated ZIP

Supporting file: eval_dataset.jsonl

{ "query": "What tent should I buy for solo backpacking?", "ground_truth": "For solo

backpacking, the TrailBlazer Solo (P003) is the best choice at 850g for weight-conscious hikers, or the TrailBlazer Tent 2P (P001) for extra space.", "category": "tents", "product_ids": ["P003", "P001"]} {"query": "How cold can the ArcticDown sleeping bag handle?", "ground_truth": "The ArcticDown sleeping bag (P004) is rated to -20°C, making it suitable for winter camping and cold-weather expeditions.", "category": "sleeping_bags", "product_ids": ["P004"]} {"query": "What backpack fits most gear for a multi-day hike?", "ground_truth": "The Osprey Atmos 65L (P006) is ideal for multi-day hikes with its 65-liter capacity, anti-gravity suspension, and rain cover.", "category": "backpacks", "product_ids": ["P006"]} {"query": "I need a small pack for day hikes", "ground_truth": "The DayHiker 22L (P007) is perfect for day hikes at 580g with hydration compatibility and multiple pockets.", "category": "backpacks", "product_ids": ["P007"]} {"query": "Which tent handles snow and winter conditions?", "ground_truth": "The Summit Dome 4P (P002) is the 4-season option, rated for extreme weather with steel poles designed for snow load.", "category": "tents", "product_ids": ["P002"]} {"query": "What is the lightest tent you sell?", "ground_truth": "The TrailBlazer Solo (P003) is the lightest tent at 850g, designed for solo fast-and-light backpacking trips.", "category": "tents", "product_ids": ["P003"]} {"query": "Do you have a summer sleeping bag?", "ground_truth": "Yes, the SummerLight +5°C sleeping bag (P005) is designed for summer use with synthetic insulation and a side zipper for ventilation.", "category": "sleeping_bags", "product_ids": ["P005"]} {"query": "What's the difference between the two sleeping bags?", "ground_truth": "The ArcticDown (P004) uses 800-fill down and is rated to -20°C for winter use. The SummerLight (P005) uses synthetic insulation rated to +5°C for summer camping.", "category": "sleeping_bags", "product_ids": ["P004", "P005"]} {"query": "Which backpack has a hydration sleeve?", "ground_truth": "Both the Osprey Atmos 65L (P006) and the DayHiker 22L (P007) are hydration compatible with a reservoir sleeve.", "category": "backpacks", "product_ids": ["P006", "P007"]} {"query": "I want a tent for a family of 4", "ground_truth": "The Summit Dome 4P (P002) fits 4 people and is the most spacious option in our catalog, though it weighs 3.2kg.", "category": "tents", "product_ids": ["P002"]} {"query": "How much does the TrailBlazer Tent 2P cost?", "ground_truth": "The TrailBlazer Tent 2P (P001) costs \$249.99.", "category": "pricing", "product_ids": ["P001"]} {"query": "What is the price of the Osprey Atmos 65L?", "ground_truth": "The Osprey Atmos 65L (P006) is priced at \$289.99 and includes a rain cover.", "category": "pricing", "product_ids": ["P006"]} {"query": "Which products are suitable for cold weather?", "ground_truth": "For cold weather: the Summit Dome 4P (P002) tent for 4-season camping, and the ArcticDown -20°C sleeping bag (P004) for winter nights.", "category": "recommendations", "product_ids": ["P002", "P004"]} {"query": "What gear do I need for a solo overnight backpacking trip?", "ground_truth": "For a solo overnight trip: TrailBlazer Solo tent (P003), SummerLight +5°C sleeping bag (P005) for 3 seasons or ArcticDown (P004) for cold nights, and the Osprey Atmos 65L (P006) or DayHiker 22L (P007) for your pack.", "category": "recommendations", "product_ids": ["P003", "P005", "P006"]} {"query": "Do you have any backpacks with laptop sleeves?", "ground_truth": "Yes, the DayHiker 22L (P007) includes a padded 15-inch laptop sleeve, making it suitable for both hiking and everyday use.", "category": "backpacks", "product_ids": ["P007"]} {"query": "What is the weight of the Summit Dome?", "ground_truth": "The Summit Dome 4P (P002) weighs 3.2kg. It is heavier than our 3-season tents due to its steel poles and 4-season construction.", "category": "tents", "product_ids": ["P002"]} {"query": "Which tent is best for weekend camping?", "ground_truth": "For weekend camping, the TrailBlazer Tent 2P (P001) offers a great balance of weight (1.8kg), space, and weather protection at \$249.99.", "category": "tents", "product_ids": ["P001"]} {"query": "Is there a product under \$100?",

"ground_truth": "Yes, the DayHiker 22L daypack (P007) is priced at \$89.99, making it our most affordable product.", "category": "pricing", "product_ids": ["P007"]} {"query": "What's included with the Osprey Atmos 65L?", "ground_truth": "The Osprey Atmos 65L (P006) includes a rain cover, hipbelt pockets, hydration sleeve, and an anti-gravity suspension system. It fits torsos from 40-52cm.", "category": "backpacks", "product_ids": ["P006"]} {"query": "Tell me about products not sold by OutdoorGear", "ground_truth": "I can only provide information about OutdoorGear products. I don't have information about competitor products.", "category": "out_of_scope", "product_ids": []}