

warning "Atualização de backend (outubro de 2026)" A inferência do GitHub Models foi descontinuada em 30 de julho de 2026. As referências abaixo são mantidas apenas como contexto histórico enquanto este lab é migrado. Use o Lab 013 ? Backends de LLM gratuitos para configurar o Foundry Local ou o Ollama antes de continuar.

Lab 035: Avaliação de Agentes com Azure AI Eval SDK

Nível: L300 Trilha: Pro Code Tempo: ~60 min ? Custo: GitHub Gratuito ? usa GitHub Models para avaliação

O que Você Vai Aprender

Por que a avaliação LLM-as-judge funciona e quando usá-la Medir fundamentação, relevância, coerência e fluência Construir um golden dataset para testes de regressão Avaliar seu agente RAG automaticamente com o SDK azure-ai-evaluation Rastrear métricas de qualidade ao longo do tempo e detectar regressões

Introdução

Como você sabe se seu agente melhorou ou piorou após uma mudança? Testes manuais não escalam.

A avaliação LLM-as-judge usa um segundo LLM independente para pontuar as respostas do seu agente com base em critérios como:

Fundamentação ? A resposta é suportada pelos documentos recuperados? Relevância ? A resposta aborda o que o usuário perguntou? Coerência ? A resposta é logicamente estruturada? Fluência ? A gramática está correta?

Este lab constrói um pipeline de avaliação automatizado para o assistente OutdoorGear.

Pré-requisitos

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

? Arquivos de Apoio

note "Baixe estes arquivos antes de iniciar o lab" Salve todos os arquivos em uma pasta lab-035/ no seu diretório de trabalho.

Arquivo | Descrição | Download | eval_dataset.jsonl | Conjunto de dados de avaliação | ? Download |

Exercício do Lab

Passo 1: Instalar o SDK de avaliação

[diagram or code sample]

Passo 2: Criar um golden dataset

Um golden dataset é um conjunto de perguntas de teste com respostas corretas conhecidas.

Crie ? eval_dataset.jsonl:

[diagram or code sample]

Observe a 4a entrada ? o agente responde uma pergunta não relacionada à base de conhecimento. A avaliação de fundamentação deve sinalizar isso.

Passo 3: Executar avaliadores integrados

[diagram or code sample]

[diagram or code sample]

Saída esperada: [diagram or code sample]

A pergunta sobre Paris é corretamente sinalizada com baixa fundamentação (1.0) ? a resposta não é suportada pelo contexto do OutdoorGear.

Passo 4: Construir um avaliador personalizado

Às vezes você precisa de critérios específicos do domínio. Aqui está um avaliador personalizado para precisão de preço:

[diagram or code sample]

Passo 5: Automatizar no CI/CD

Adicione isso ao seu workflow do GitHub Actions (veja Lab 037):

[diagram or code sample]

Referência de Métricas de Avaliação

Métrica | O que mede | Pontuação boa | Fundamentação | Resposta suportada pelo contexto | ?
4.0 | Relevância | Responde à pergunta do usuário | ? 4.0 | Coerência | Lógica, bem estruturada | ? 4.0 | Fluência | Gramaticalmente correta | ? 4.5 |

Pontuações são de 1 a 5. Abaixo de 3.0 em fundamentação geralmente indica alucinação.

? Verificação de Conhecimento

question "Q1 (Execute o Lab): Carregue lab-035/eval_dataset.jsonl em Python. Quantos exemplos existem no conjunto de dados, e quantos estão na categoria out_of_scope?"

Execute o código de carregamento da seção Arquivos de Apoio acima.

??? success "? Revelar Resposta" 20 exemplos no total. 1 exemplo está na categoria out_of_scope.

O conjunto de dados tem exatamente 20 linhas. Execute `sum(1 for d in dataset if d["category"] == "out_of_scope")` para confirmar que há 1 exemplo fora do escopo. Esse exemplo testa se seu agente recusa corretamente responder perguntas não relacionadas a equipamentos outdoor.

question "Q2 (Execute o Lab): Para o único exemplo out_of_scope em eval_dataset.jsonl, qual é o valor do campo product_ids?"

Carregue o conjunto de dados e filtre por `category == "out_of_scope"`. Imprima o campo `product_ids`.

??? success "? Revelar Resposta" [] (lista vazia)

O exemplo fora do escopo tem `"product_ids": []` porque a pergunta não é sobre nenhum produto específico ? está testando se o agente recusa responder perguntas irrelevantes (como pedir receitas culinárias). Um agente bem projetado deve retornar uma mensagem de recusa em vez de alucinar uma resposta. Sua métrica de avaliação deve verificar que a pontuação de groundedness do agente é alta e que ele NÃO referencia nenhum ID de produto.

question "Q3 (Múltipla Escolha): Quantos exemplos em eval_dataset.jsonl estão na categoria tents?"

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

??? success "? Revelar Resposta" Correto: B ? 5 exemplos

A categoria tents tem 5 exemplos, sendo a maior categoria individual. Execute `sum(1 for d in dataset if d["category"] == "tents")` para confirmar. A distribuição completa: tents(5), backpacks(4), sleeping_bags(3), pricing(3), recommendations(3), out_of_scope(1), comparisons(1).

Próximos Passos

CI/CD para agentes: ? Lab 037 ? GitHub Actions for AI Agents Avaliação RAG empresarial: ?

Lab 042 ? RAG Empresarial com Avaliações *(em breve)*

Modernização da Fase 3

warning "Verified 2026-10-27" Use o nome Microsoft Foundry e o Agent Service GA baseado na Responses API. Avaliações estão GA; hosted agents continuam preview e devem ser identificados assim. Não crie dependência nova de Workflows do portal, descontinuados em 01/12/2026.

[diagram or code sample]

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

Kit de apoio

O build do site gera downloads estáveis em PDF e ZIP. O ZIP contém o guia do lab e todos os

arquivos de apoio declarados para este lab.

Baixar PDF gerado Baixar ZIP gerado

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": []}