

warning "Mise à jour du backend (octobre 2026)" L'inférence GitHub Models a été retirée le 30 juillet 2026. Les références ci-dessous sont conservées uniquement comme contexte historique pendant la migration de ce lab. Utilisez le Lab 013 ? Backends LLM gratuits pour configurer Foundry Local ou Ollama avant de continuer.

Lab 035 : Évaluation des agents avec le SDK Azure AI Eval

Niveau : L300 Parcours : Pro Code Durée : ~60 min ? Coût : GitHub Gratuit ? utilise GitHub Models pour l'évaluation

Ce que vous apprendrez

Pourquoi l'évaluation LLM-as-judge fonctionne et quand l'utiliser Mesurer l'ancrage, la pertinence, la cohérence et la fluidité Construire un jeu de données de référence pour les tests de régression Évaluer automatiquement votre agent RAG avec le SDK azure-ai-evaluation Suivre les métriques de qualité dans le temps et détecter les régressions

Introduction

Comment savoir si votre agent s'est amélioré ou dégradé après une modification ? Les tests manuels ne passent pas à l'échelle.

L'évaluation LLM-as-judge utilise un second LLM indépendant pour noter les réponses de votre agent selon des critères tels que :

Ancrage (Groundedness) ? La réponse est-elle soutenue par les documents récupérés ?

Pertinence (Relevance) ? La réponse traite-t-elle ce que l'utilisateur a demandé ?

Cohérence (Coherence) ? La réponse est-elle logiquement structurée ? Fluidité (Fluency) ?

La grammaire est-elle correcte ?

Ce lab construit un pipeline d'évaluation automatisé pour l'assistant OutdoorGear.

Prérequis

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

? Fichiers de support

note "Téléchargez ces fichiers avant de commencer le lab" Enregistrez tous les fichiers dans un dossier lab-035/ de votre répertoire de travail.

Fichier | Description | Téléchargement | eval_dataset.jsonl | Jeu de données d'évaluation |

? Télécharger |

Exercice du lab

Étape 1 : Installer le SDK d'évaluation

[diagram or code sample]

Étape 2 : Créer un jeu de données de référence

Un jeu de données de référence est un ensemble de questions de test avec des réponses de qualité connue. Créez ? eval_dataset.jsonl :

[diagram or code sample]

Notez la 4e entrée ? l'agent répond à une question sans rapport avec la base de connaissances. L'évaluation de l'ancrage devrait signaler cela.

Étape 3 : Exécuter les évaluateurs intégrés

[diagram or code sample]

[diagram or code sample]

Sortie attendue : [diagram or code sample]

La question sur Paris est correctement signalée avec un faible ancrage (1.0) ? la réponse n'est pas soutenue par le contexte OutdoorGear.

Étape 4 : Construire un évaluateur personnalisé

Parfois vous avez besoin de critères spécifiques au domaine. Voici un évaluateur personnalisé pour la précision des prix :

[diagram or code sample]

Étape 5 : Automatiser en CI/CD

Ajoutez ceci à votre workflow GitHub Actions (voir Lab 037) :

[diagram or code sample]

Référence des métriques d'évaluation

Métrique | Ce qu'elle mesure | Bon score | Ancrage (Groundedness) | Réponse soutenue par le contexte | ? 4.0 | Pertinence (Relevance) | Répond à la question de l'utilisateur | ? 4.0 | Cohérence (Coherence) | Logique, bien structurée | ? 4.0 | Fluidité (Fluency) | Grammaticalement correcte | ? 4.5 |

Les scores vont de 1 à 5. En dessous de 3.0 en ancrage, cela indique généralement une hallucination.

? Vérification des connaissances

question "Q1 (Exécutez le lab) : Chargez lab-035/eval_dataset.jsonl en Python. Combien d'exemples contient le jeu de données, et combien sont dans la catégorie out_of_scope ?" Exécutez le code de chargement de la section Fichiers de support ci-dessus.

??? success "? Révéler la réponse" 20 exemples au total. 1 exemple est dans la catégorie out_of_scope.

Le jeu de données contient exactement 20 lignes. Exécutez `sum(1 for d in dataset if d["category"] == "out_of_scope")` pour confirmer qu'il y a 1 exemple hors périmètre. Cet exemple teste si votre agent refuse correctement de répondre aux questions sans rapport avec l'équipement de plein air.

question "Q2 (Exécutez le lab) : Pour l'unique exemple out_of_scope dans eval_dataset.jsonl, quelle est la valeur du champ product_ids ?"

Chargez le jeu de données et filtrez par `category == "out_of_scope"`. Affichez le champ `product_ids`.

??? success "? Révéler la réponse" [] (liste vide)

L'exemple hors périmètre a "product_ids": [] car la question ne porte sur aucun produit spécifique ? il teste si l'agent refuse de répondre aux questions non pertinentes (comme demander des recettes de cuisine). Un agent bien conçu devrait retourner un message de refus plutôt que d'halluciner une réponse. Votre métrique d'évaluation devrait vérifier que le score d'ancrage de l'agent est élevé et qu'il ne référence AUCUN identifiant de produit. question "Q3 (Choix multiple) : Combien d'exemples dans eval_dataset.jsonl sont dans la catégorie tents ?"

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

??? success "? Révéler la réponse" Correct : B ? 5 exemples

La catégorie tents contient 5 exemples, ce qui en fait la plus grande catégorie individuelle. Exécutez `sum(1 for d in dataset if d["category"] == "tents")` pour confirmer. La répartition complète : tents(5), backpacks(4), sleeping_bags(3), pricing(3), recommendations(3), out_of_scope(1), comparisons(1).

Prochaines étapes

CI/CD pour les agents : ? Lab 037 ? GitHub Actions pour les agents IA Évaluation RAG

d'entreprise : ? Lab 042 ? RAG d'entreprise avec évaluations *(bientôt disponible)*

Modernisation de Phase 3

warning "Verified 2026-10-27" Utilisez Microsoft Foundry et Agent Service GA fondé sur Responses API. Les évaluations sont GA ; les hosted agents restent en preview et doivent être signalés. Ne dépendez pas des Workflows du portail retirés le 01/12/2026.

[diagram or code sample]

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

Kit d'accompagnement

Le build du site génère des téléchargements PDF et ZIP stables. Le ZIP contient le guide du lab et tous les fichiers d'accompagnement déclarés.

Télécharger le PDF généré Télécharger le ZIP généré

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": []}