Workshop AI (day 6) Microsoft-AGIC-UniPadova

Evaluation in AI Foundry

AI Foundry supporta **evaluation dei Large Language Model** con un focus sia sulla **performance e qualità**, sia sugli aspetti di **safety e security**. Di recente è stato introdotto anche il **red teaming**, ovvero una forma di stress test aggressivo sui modelli per identificarne punti deboli o comportamenti indesiderati.

Microsoft mette a disposizione **system prompt e metodi predefiniti** per facilitare lo sviluppatore nella valutazione dei propri chatbot. Gli evaluator si trovano direttamente nell'interfaccia utente di AI Foundry, nella sezione dedicata all'evaluation. All'interno di questa sezione è presente una voce chiamata **"Evaluator Library"**, che raccoglie tutti i valutatori disponibili.

Cliccando su uno di essi, si accede al relativo codice Python, all'interno dei quali si trovano commentati i **percorsi ai file prompt template (prompty)** utilizzati per condurre le valutazioni.

Categorie principali di Evaluator

1. Performance and Quality Evaluators

- Valutano la **qualità delle risposte**.
- Consentono il **confronto tra modelli** (es. GPT-4 vs GPT-3.5), usando eventualmente un **terzo LLM come giudice** (es. GPT-4.5).
- È anche possibile usare lo **stesso modello per auto-valutazione**.
- Per valutazioni più oggettive si possono usare metriche **deterministiche** come F1 Score, GLEU, N-Precision (NP)

Per la **groundedness** è spesso utile coinvolgere un **utente umano** che fornisca una ground truth di riferimento.

2. Risk and Safety Evaluators

- Utilizzano il servizio **multimodale** di **Content Service**.
- Servono a prevenire contenuti pericolosi o inappropriati secondo i seguenti criteri:
  - *Self-harm*
  - *Sexual content*
  - *Hate/unfairness*
  - *Violence*

Possono valutare **immagini fornite dall'utente** o **generate dall'assistente**.

3. Custom Evaluators

- Permettono allo sviluppatore di definire **classi personalizzate**, combinando:
  - Codice custom
  - Prompt custom

Quando avviene la valutazione

La valutazione va eseguita **dopo la generazione** da parte dell'assistente, se possibile **prima che l'utente veda la risposta**, permettendo un filtraggio o una correzione preventiva. È possibile valutare **conversazioni complete**, non solo singoli turni.