OpenEvals 0.1.1

LangChain	**Evaluation of LLM applications**	
	Systeemvereisten:	Python, TypeScript
	Ontwikkeld door:	LangChain
Open source, MIT licentie	Contactpersoon:	Bert.Vanhalst@smals.be

Functionaliteiten

OpenEvals is een open-source package voor het evalueren van LLM toepassingen, of ze nu ontwikkeld zijn met LangChain, LangGraph of een ander framework. Het helpt ontwikkelaars om systematisch de kwaliteit van LLM-gebaseerde toepassingen te beoordelen. Het laat toe om problemen te identificeren en om de impact van verbeteringen te meten.

Concreet voorziet de package kant-en-klare LLM-as-judge evaluators voor het meten van *correctness*, *conciseness* en *hallucination detection*, string-gebaseerde evaluators voor *exact-matching* en *similarity*, en specifieke evaluators voor *Retrieval Augmented Generation* (RAG) toepassingen zoals *helpfulness*, *groundedness* en *retrieval relevance*. Daarnaast heeft OpenEvals ook mogelijkheden om gegenereerde code te evalueren met type-checking ondersteuning voor Python en TypeScript, en tools om multi-turn conversaties te simuleren.

Evaluators zijn aanpasbaar. Zo kan het scoring-mechanisme aangepast worden door gebruik te maken van continue waarden tussen 0 en 1 in plaats van binaire scores, of door specifieke keuzemogelijkheden te definiëren met aangepaste scoringscriteria. De onderliggende modellen kunnen eenvoudig gewisseld worden en de prompts kunnen volledig aangepast worden. OpenEvals biedt ook de mogelijkheid om volledig eigen evaluators te voorzien met aangepaste metrieken op basis van functies die voldoen aan de OpenEvals-interface zodat ze consistent geïntegreerd kunnen worden in het OpenEvals ecosysteem.

OpenEvals integreert met LangSmith, LangChain's platform voor het monitoren van LLM-gebaseerde toepassingen. Op die manier kunnen testresultaten bijgehouden worden over experimenten heen.

Conclusies & Aanbevelingen

Hoewel nog nieuw (v0.1.1) levert de OpenEvals package toch al enkele bruikbare out-of-the-box evaluators. De alignering van deze automatische evaluators met een menselijke score is misschien niet perfect, maar ze zijn wel al praktisch bruikbaar en kunnen nog gecustomiseerd worden door de prompt aan te passen. Goed om weten: naast OpenEvals biedt LangChain ook AgentEvals dat specifiek gericht is op het evalueren van agents.

Testen & Resultaten

We gaan aan de slag met OpenEvals om een Retrieval Augmented Generation (RAG) systeem voor het beantwoorden van algemene vragen over studentenarbeid te evalueren.

Vooraleer we aan de slag gaan met de automatische evaluatie wordt er eerst een manueel traject doorlopen. In deze manuele fase stelt een business expert een lijst op van 64 representatieve testvragen. Voor elk van die testvragen wordt een referentieantwoord opgesteld door de expert. Vervolgens laten we een antwoord genereren door de RAG-toepassing. Dit gegenereerde antwoord wordt vergeleken met het referentieantwoord en er wordt een score gegeven van 0 (fout), 1 (deels correct of onvolledig) of 2 (correct en volledig).

Daarna lanceren we de automatische evaluator. Concreet gaan we aan de slag met de *correctness* evaluator die nagaat of een gegenereerd antwoord factueel juist en volledig is ten opzichte van een referentieantwoord. In het voorbeeld hieronder is de input weergegeven voor de correctness evaluator: de gebruikersvraag, het gegenereerd antwoord en het referentieantwoord. De evaluator geeft als output niet alleen een score (true), maar ook een motivatie die ons meer inzicht geeft in mogelijke problemen met de kwaliteit van het antwoord.

Input

Input: *A partir de quel âge puis-je travailler comme étudiant?*

Generated output: *Tu peux t'engager comme jobiste dès que tu as 15 ans et que tu as suivi le premier degré de l'enseignement secondaire, ou dès que tu as 16 ans.*

Reference output: *Vous pouvez travailler comme étudiant dès que vous avez 16 ans ou si vous avez 15 ans et avez suivi les deux premières années de l'enseignement secondaire.*

Output

Score: *true*

Motivation: *The provided answer states that a student can work at 15 years old if they have completed the first cycle ("premier degré") of secondary education, equivalent to the two first years of secondary education mentioned in the reference. It also notes that one can work at 16 years old. This information is factually accurate, complete, and addresses the question using precise terminology. Thus, the score should be: true.*

We krijgen goede en bruikbare resultaten, mits we een reasoning model gebruiken (gpt-o3-mini) in plaats van een chat model als gpt-4o-mini. De *alignment* is 70%, wat betekent dat de score van de evalutor in 70% van de testcases gelijk is aan de manuele score. Dit is deels te wijten aan het verschil in scoring-mechanisme dat binair is voor de automatische evaluator, maar drieledig (0-1-2) bij de manuele score.

Andere metrieken, zoals *helpfulness*, *groundedness*, *hallucination detection* en *retrieval relevance* werden kort getest en kunnen nuttig zijn om het volledige systeem of aparte componenten (*retrieval* en *generation*) te debuggen.

Gebruiksvoorwaarden & Budget

OpenEvals is beschikbaar als open-source package onder MIT licentie. Er zijn geen directe kosten verbonden aan het gebruik ervan, maar gebruikers moeten rekening houden met mogelijke indirecte kosten zoals API-kosten voor het gebruik van externe LLM's.