Ewaluacje i testy porównawcze

Biorąc pod uwagę niedawny wzrost zainteresowania GenAI, a zwłaszcza dużymi modelami językowymi (LLM), niezwykle ważne jest ostrożne i odpowiedzialne podejście do tych nowych i niepewnych funkcji. Wiele rankingów i badań wykazało, że LLM mogą dorównywać ludzkim możliwościom w różnych zadaniach, takich jak rozwiązywanie standardowych testów czy tworzenie dzieł sztuki, wzbudzając entuzjazm i uwagę. Jednak ich nowość i niepewność wymagają ostrożnego podejścia. Nie można przecenić roli benchmarkingu LLM we wdrożeniach produkcyjnych. Obejmuje ona ocenę wydajności, porównywanie modeli, kierowanie ulepszeniami, przyspieszanie postępu technologicznego, zarządzanie kosztami i opóźnieniami oraz zapewnienie efektywnego przepływu zadań w rzeczywistych aplikacjach. Chociaż ewaluacje są częścią LLMOps, ich kluczowe znaczenie w zapewnieniu, że LLM spełniają wymagania różnych aplikacji, uzasadnia osobne omówienie w tym rozdziale. Ewaluacja LLM nie jest prostym zadaniem, lecz złożonym i wieloaspektowym procesem, który wymaga podejścia ilościowego i jakościowego. Przy ocenie LLM należy stosować kompleksowe metody oceny obejmujące różne aspekty wydajności i skuteczności modelu. Dział sztucznej inteligencji zorientowanej na człowieka (HAI) Uniwersytetu Stanforda publikuje coroczny raport AI Index , którego celem jest zestawianie i śledzenie różnych punktów danych związanych ze sztuczną inteligencją. Jednym z najpoważniejszych wyzwań, z jakimi się mierzymy, jest brak standaryzowanych ocen, co sprawia, że systematyczne porównywanie różnych modeli jest niezwykle trudne, jeśli chodzi o ich możliwości oraz potencjalne ryzyko i szkody. Oznacza to, że nie dysponujemy obiektywną miarą tego, jak dobry lub inteligentny jest którykolwiek z tych konkretnych modeli, co podkreśla złożoność i znaczenie procesu oceny. Kiedy omawiamy oceny GenAI na tym początkowym etapie, większość dyskusji dotyczy ocen dokładności i wydajności, które oceniają, jak dobrze model językowy potrafi rozumieć i generować tekst przypominający język ludzki. Ten aspekt jest bardzo ważny dla aplikacji, które polegają na jakości i trafności generowanych treści, takich jak chatboty, tworzenie treści i zadania podsumowujące. Istnieją trzy ogólne rodzaje ewaluacji, które mogą mierzyć dokładność i wydajność: tradycyjne wskaźniki ewaluacji, które oceniają jakość języka, specyficzne dla danego zadania kryteria oceny LLM oraz ewaluacje dokonywane przez ludzi. Zacznijmy od zrozumienia, czym są ewaluacje LLM i poznajmy najlepsze praktyki związane z ewaluacją.

Ewaluacje LLM

Ocena LLM jest niezbędna, aby upewnić się, że są one wiarygodne i odpowiednie do rzeczywistych zastosowań. Solidna strategia ewaluacji obejmuje wskaźniki wydajności, takie jak dokładność, płynność, spójność i trafność. Wskaźniki te pomagają nam zrozumieć zalety i wady modelu w różnych kontekstach. Podsumowuję tutaj kilka obszarów jako najlepsze praktyki, które należy wziąć pod uwagę podczas oceny LLM:

* Aby wiarygodnie ocenić LLM, należy go przetestować w przypadkach użycia, do których został zaprojektowany, czyli wykorzystać model w różnych zadaniach przetwarzania języka naturalnego (NLP), takich jak podsumowywanie, odpowiadanie na pytania i tłumaczenie. Proces ewaluacji powinien wykorzystywać standardowe wskaźniki, takie jak ROUGE (Recall-Oriented Understudy for Gisting Evaluation), do podsumowania, aby zachować wiarygodność i porównywalność.

* Innym ważnym aspektem ewaluacji LLM jest tworzenie podpowiedzi. Podpowiedzi muszą być jednoznaczne i sprawiedliwe, zapewniając rzetelną ocenę możliwości modelu. Gwarantuje to, że wyniki ewaluacji odzwierciedlają rzeczywistą wydajność modelu.

* Benchmarking to kluczowa praktyka, która umożliwia ocenę wyników LLM w oparciu o istniejące kryteria i inne modele. Pozwala to nie tylko śledzić postępy, ale także identyfikować obszary

wymagające poprawy. Ciągły proces ewaluacji, w połączeniu z ciągłymi praktykami rozwoju, pozwala na okresową ocenę i udoskonalanie LLM.

* Ocena LLM musi uwzględniać kwestie etyczne na każdym etapie. Proces ten musi weryfikować model pod kątem stronniczości, uczciwości i problemów etycznych, analizując dane szkoleniowe i wyniki. Ponadto, doświadczenie użytkownika powinno być kluczowym elementem ewaluacji, zapewniając, że wyniki modelu odpowiadają potrzebom i oczekiwaniom użytkownika.

* Ocena musi być transparentna na każdym etapie. Rejestrowanie kryteriów, metod i wyników umożliwia niezależną weryfikację i zwiększa zaufanie do umiejętności LLM. Wreszcie, wyniki ewaluacji powinny stanowić podstawę cyklu ciągłego doskonalenia, ulepszając model, dane szkoleniowe i proces ewaluacji w oparciu o wskaźniki wydajności i informacje zwrotne.

Praktyki te podkreślają wagę rygorystycznego i systematycznego podejścia do oceny LLM, zapewniając ich dokładność, ale także uczciwość, etyczność i przydatność do różnych zastosowań. Stosując te praktyki, przedsiębiorstwa mogą przeprowadzać rzetelne i skuteczne oceny, opracowując wiarygodne i pomocne LLM do różnych zastosowań. Teraz, gdy wiemy, czym są oceny, przyjrzyjmy się niektórym wskaźnikom, których powinniśmy używać. Można je podzielić na tradycyjne i nowsze wskaźniki oceny specyficzne dla LLM.

Tradycyjne wskaźniki oceny

BLEU (Bilingual Evaluation Understudy), ROUGE (Recall-Oriented Understudy for Gisting Evaluation) i BERTScore (BERT Similarity Score) to jedne z bardziej ujednoliconych wskaźników. Wskaźniki te pomagają określić jakość językową wyników modelu w odniesieniu do tekstów referencyjnych i są wykorzystywane do oceny jakości tekstu w zadaniach takich jak tłumaczenie maszynowe lub streszczanie tekstu. Mimo to różnią się one w swoich podejściach i koncentrują się na różnych aspektach tekstu. Tabela przedstawia szczegółowe wyjaśnienie, co oznacza każdy z trzech wyników. W następnej sekcji pokażemy, jak je obliczyć.

Metryka: Cel: Metoda: Ograniczenia

BLEU: Mierzy przede wszystkim precyzję, czyli odsetek słów w tekście generowanym maszynowo, które pojawiają się w tekście referencyjnym. Porównuje n-gramy (sekwencje słów) tłumaczenia kandydata z tłumaczeniem referencyjnym i zlicza dopasowania. Może nie uwzględniać znaczenia semantycznego, ponieważ nie uwzględnia synonimów ani kontekstu słów. Nie radzi sobie również dobrze z reorganizacją wyrazów.

ROUGE: Jest bardziej zorientowany na przypominanie, koncentrując się na odsetku słów z tekstu referencyjnego, które pojawiają się w tekście generowanym. Ma kilka wariantów, takich jak ROUGE-N, który porównuje n-gramy, i ROUGE-L, który bierze pod uwagę najdłuższy wspólny podciąg. Podobnie jak BLEU, ROUGE może pomijać podobieństwa semantyczne i parafrazowanie, ponieważ opiera się na dokładnych dopasowaniach słów.

BERTScore: Ocenia podobieństwo semantyczne, a nie opiera się na dokładnych dopasowaniach słów. Wykorzystuje kontekstowe osadzenia z modeli takich jak BERT do reprezentacji tekstu i oblicza cosinusowe podobieństwo między tymi osadzeniami. Potrafi uchwycić parafrazę i znaczenie semantyczne lepiej niż BLEU i ROUGE, ponieważ uwzględnia kontekst każdego słowa.

Wskaźniki takie jak ROUGE, BLEU i BERTScore porównują podobieństwa między tekstem generowanym przez LLM a tekstem referencyjnym napisanym przez człowieka. Są one powszechnie używane do oceny zadań takich jak streszczanie i tłumaczenie maszynowe.

BLEUv : BLEU (Bilingual Evaluation Understudy) [2] to algorytm służący do oceny jakości tekstu przetłumaczonego maszynowo z jednego języka naturalnego na inny. Jego główną ideą jest pomiar zgodności między wynikami maszyny a wynikami tłumacza. Innymi słowy, według BLEU, im bliższe jest tłumaczenie maszynowe profesjonalnemu tłumaczeniu ludzkiemu, tym jest lepsze. BLEU nie bierze pod uwagę zrozumiałości ani poprawności gramatycznej; koncentruje się na nakładaniu się treści.

ROUGE

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [3] to zestaw miar stosowanych w NLP do oceny skuteczności automatycznego streszczania tekstu i tłumaczenia maszynowego. Jego głównym celem jest porównanie streszczeń lub tłumaczeń generowanych przez maszyny z ludzkimi streszczeniami referencyjnymi. Ocenia on następujące aspekty:

* Recall (Pamięć) — ROUGE mierzy, ile streszczenia referencyjnego jest przechwytywane przez podsumowanie systemowe. Ocenia, jak dobrze system odzyskuje lub przechwytuje treść z referencyjnego.

* Precyzja — ocenia również, ile z podsumowania systemu jest istotne, potrzebne lub użyteczne.

* Miara F — łączy precyzję i przypominanie, aby zapewnić zrównoważony obraz wydajności systemu. ROUGE ma różne wersje, takie jak ROUGE-N (wykorzystująca n-gramy) i ROUGE-L (oparta na algorytmie najdłuższego wspólnego podciągu). Analizując pojedyncze słowa i sekwencje, ROUGE pomaga nam mierzyć skuteczność algorytmów NLP w zadaniach podsumowania i tłumaczenia. ROUGE ma jednak ograniczenia. Opiera się wyłącznie na nakładaniu się na poziomie powierzchniowym i nie uwzględnia znaczenia semantycznego ani płynności. Wrażliwość na słowa z przerwą, stemming i kolejność wyrazów może wpływać na wyniki. Chociaż ROUGE dostarcza cennych informacji, kluczowe jest uwzględnienie innych metryk ewaluacyjnych i osądu ludzkiego, aby kompleksowo ocenić jakość podsumowania. Badacze często używają kombinacji metryk do oceny modeli podsumowania.

BERTScore

BERTScore to miara jakości generowania tekstu. Wykorzystuje on wstępnie wytrenowane osadzenia modelu BERT do porównywania zdań kandydujących i referencyjnych. Celem jest znalezienie podobnych słów w zdaniach kandydujących i referencyjnych na podstawie podobieństwa cosinusowego. Ta metryka jest zgodna z opinią człowieka w ocenach na poziomie zdań i systemu. Składa się z następujących elementów:

* Osadzenia kontekstowe — BERTScore reprezentuje zarówno zdania kandydujące, jak i referencyjne, z osadzeniami uwzględniającymi kontekst każdego słowa.

* Podobieństwo cosinusowe — oblicza podobieństwo cosinusowe między osadzeniami tekstów kandydujących i referencyjnych.

* Dopasowanie tokenów — aby obliczyć wyniki precyzji i odtworzenia, każdy token w tekście kandydującym pasuje do najbardziej podobnego tokenu w tekście referencyjnym.

* Wynik F1 — precyzja i odtworzenie są łączone w celu obliczenia wyniku F1, co zapewnia pojedynczą miarę jakości. Kluczową zaletą BERTScore w porównaniu z tradycyjnymi metrykami, takimi jak BLEU, jest jego zdolność do wychwytywania podobieństwa semantycznego. Oznacza to, że potrafi rozpoznać, kiedy różne słowa mają podobne znaczenie, a kiedy te same słowa są używane w różnych kontekstach.

Przykład tradycyjnej oceny metryk

Połączmy to wszystko i urzeczywistnijmy na prostym przykładzie. Mamy tu dwa podsumowania informacji i możemy ocenić, które z nich może być lepsze. W tym przykładzie bierzemy zasady rozwoju sztucznej inteligencji Fundacji Billa i Melindy Gatesów za artykuł, który chcemy przeanalizować i zrozumieć. Artykuł jest dostępny pod adresem https://mng.bz/vJe4. Na podstawie artykułu tworzymy dwa podsumowania, które porównamy. W tym przypadku jedno zostało stworzone przez NLTK, a drugie przez innego LLM (GPT-3.5). Mogą to być również dwie różne wersje napisane przez ludzi lub dowolna inna kombinacja. Używamy pakietów newspaper3K i bert_score do pobrania artykułu oraz pakietu Hugging Face Evaluate do oceny. Można je zainstalować w systemie conda za pomocą polecenia conda install -c conda-forge journal3k evaluate bert_score. W pip użyj polecenia pip install evaluate journal3k bert_score. Najpierw używamy journal3k do pobrania i przeanalizowania artykułu. Następnie stosujemy funkcję nlp(), aby przetworzyć artykuł i uzyskać podsumowanie z właściwości summary. Przed użyciem NLP musimy upewnić się, że artykuł został pobrany i przeanalizowany; należy pamiętać, że działa to tylko w przypadku języków zachodnich. Używamy podsumowania utworzonego przez NLP jako podsumowania referencyjnego, a biblioteki Evaluate do obliczenia konkretnych metryk. Listing przedstawia kod implementujący tę funkcję.

Listing 1: Automatyczne metryki oceny

```
from openai import AzureOpenAI

import evaluate

from bert_score import BERTScorer

...
```

```
AOAI_API_KEY = os.getenv("AOAI_KEY")
AZURE_ENDPOINT = os.getenv("AOAI_ENDPOINT")
...

URL = "https://www.gatesfoundation.org/ideas/articles/
            ➥artificial-intelligence-ai-development-principles"

def get_article(URL, config):
    article = Article(URL, config=config)
    article.download(recursion_counter=2)
    article.parse()
    article.nlp()
    return article.text, article.summary

def generate_summary(client, article_text):
    prompt = f"Summarize the following article:\n\n{article_text}"
    conversation = [{"role": "system", "content":
                        ➥"You are a helpful assistant."}]
    conversation.append({"role": "user", "content": prompt})

    response = client.chat.completions.create(
        model=MODEL,
        messages = conversation,
        temperature = TEMPERATURE,
        max_tokens = MAX_TOKENS,
    )
    return response.choices[0].message.content.strip()

def calculate_scores(generated_summary, reference_summary):
    metric = evaluate.load("bleu", trust_remote_code=True)
    bleu_score = metric.compute(predictions=
                ➥[generated_summary], references=[reference_summary])

    metric = evaluate.load("rouge", trust_remote_code=True)
    rouge_score = metric.compute(predictions=
                ➥[generated_summary], references=[reference_summary])

    scorer = BERTScorer(lang="en")
    p1, r1, f1 = scorer.score([generated_summary], [reference_summary])
    bert_score = f"Precision: {p1} Recall: {r1} F1 Score: {f1.tolist()[0]}"

    return bleu_score, rouge_score, bert_score

# Main code
client = AzureOpenAI(
      azure_endpoint = AZURE_ENDPOINT,
      api_key=AOAI_API_KEY,
      api_version=API_VERSION
)

config = Config()
config.browser_user_agent = USER_AGENT
config.request_timeout = 10
article_text, reference_summary = get_article(URL, config)

generated_summary = generate_summary(client, article_text)
```

Sets up the OpenAI details

Function to download and parse the article; returns both the article text and a summary

Summarizes the article using OpenAI

Function to calculate metrics (BLEU, ROUGE, etc.)

Configures newspaper3k to allow downloading articles

bleu_score, rouge_score, bert_score = calculate_scores(

➥generated_summary, reference_summary)

print(f"BLEU:{bleu_score}, ROUGE:{rouge_score}, BERT: {bert_score}")

Oto wynik, który możemy zaobserwować po wykonaniu kodu:

BLEU score: {'bleu': 0.04699157347901134,

➥'precisions': [0.32857142857142857,

➥0.09352517985611511,

➥0.021739130434782608,

➥0.0072992700729927005],

➥'brevity_penalty': 1.0,

➥'length_ratio': 1.2727272727272727,

➡'translation_length': 140, [

➡'reference_length': 110}

ROUGE score: {'rouge1': 0.3463203463203463,]

➡'rouge2': 0.09606986899563319,

➡'rougeL': 0.1645021645021645,

➡'rougeLsum': 0.2683982683982684}

BERT score: Precision: tensor([0.8524])

➡Recall: tensor([0.8710])

➡F1 Score: 0.8616269826889038

Jak widzieliśmy, wynik BLEU składa się z kilku komponentów, które łącznie oceniają jakość tłumaczenia wygenerowanego maszynowo w porównaniu z zestawem tłumaczeń referencyjnych. Przyjrzyjmy się każdemu komponentowi i zobaczmy, co oznacza, zaczynając od wyniku BLEU przedstawionego w tabeli 2. Tabele 3 i 4 przedstawiają odpowiednio wyniki dla wyniku ROUGE i wyniku BERT.

Wartość składnika: Znaczenie

BLEU: 0,047 (4,7%): To ogólny wynik BLEU, który jest dość niski. Wyniki BLEU mieszczą się w przedziale od 0 do 1 (lub od 0% do 100%), przy czym wyższe wyniki oznaczają lepszą jakość tłumaczenia. Wynik poniżej 10% jest zazwyczaj uważany za słaby.

Precyzja: To wyniki precyzji n-gramowej dla dopasowań 1-gramowych, 2-gramowych, 3-gramowych i 4-gramowych. Nasze wyniki wskazują na przyzwoitą liczbę dopasowań 1-gramowych, ale niewiele dłuższych dopasowań, co sugeruje, że tłumaczenie zawiera kilka poprawnych słów, ale brakuje mu spójnych fraz i zdań.

Kara za zwięzłość: 1,0: Oznacza to, że nie naliczono kary za zwięzłość; długość tłumaczenia była odpowiednia w porównaniu z długością odniesienia.

Stosunek długości: 1,27: Tłumaczenie jest o 27% dłuższe niż odniesienie, co może sugerować pewną rozwlekłość.

Długość tłumaczenia: 140: Długość tekstu przetłumaczonego maszynowo

Długość odniesienia: 110: Długość tekstu odniesienia

ROUGE-1: 0,3463 (34,63%): Mierzy nakładanie się 1-gramów między danymi wyjściowymi systemu a podsumowaniem referencyjnym. Umiarkowany wynik wskazuje na znaczne nakładanie się.

ROUGE-2: 0,0961 (9,61%): Mierzy nakładanie się bigramów i jest bardziej rygorystycznym wskaźnikiem niż ROUGE-1. Niski wynik sugeruje, że system ma trudności z tworzeniem poprawnych fraz.

ROUGE-L: 0,1645 (16,45%): Mierzy najdłuższy wspólny podciąg, wskazując na płynność i kolejność słów. Wynik sugeruje ograniczoną płynność.

ROUGE-Lsum: 0,2684 (26,84%): Jest podobny do ROUGE-L, ale uwzględnia sumę najdłuższych wspólnych podciągów, co wskazuje na nieco lepsze zrozumienie struktury treści.

Wartość składnika: Znaczenie

Precyzja: 0,8524 (85,24%): Mierzy, ile słów w tekście kandydata jest istotnych lub potrzebnych.

Przypominanie: 0,8710 (87,10%): Mierzy, ile tekstu kandydata zawiera treść referencyjną.

Wynik F1: 0,8616 (86,16%): Ta średnia harmoniczna precyzji i przypominania daje jeden wynik, który równoważy oba. Wynik F1 bliższy 1 oznacza lepszą wydajność.

Wyniki BLEU i ROUGE sugerują, że tłumaczenie lub streszczenie ma pole do poprawy, szczególnie w zakresie tworzenia spójnych fraz i zdań. Jednak wynik BERT jest dość wysoki, co wskazuje, że tekst kandydata jest semantycznie podobny do tekstu referencyjnego i oddaje większość jego treści. Zatem, chociaż tłumaczenie może nie być dosłownie zgodne z tekstem referencyjnym, to całkiem dobrze oddaje to samo ogólne znaczenie. Chociaż wskaźniki takie jak BERTScore, ROUGE i BLEU pomagają porównywać podobne teksty, koncentrują się one przede wszystkim na podobieństwie na poziomie powierzchownym. Mogą one nie odzwierciedlać równoważności semantycznej ani ogólnej jakości wygenerowanego tekstu. Te bardziej tradycyjne wskaźniki często penalizują programy LLM, które mogą generować spójne i płynne generacje. W tym celu potrzebujemy benchmarków specyficznych dla zadań LLM.

Testy porównawcze LLM dla konkretnych zadań

Pomiar wydajności modeli LLM w różnych zadaniach NLP wymaga testów porównawczych dla konkretnych zadań. Służą one do testowania, jak dobrze modele rozumieją, rozumują i generują język naturalny dla określonych dziedzin lub zadań, zapewniając jasny sposób porównywania różnych modeli. Testy te mogą ujawnić możliwości i ograniczenia modelu, umożliwiając ukierunkowane ulepszenia. Testy porównawcze dla konkretnych zadań oceniają modele LLM w zakresie konkretnych zadań NLP, takich jak klasyfikacja tekstu, analiza sentymentu, odpowiadanie na pytania, podsumowywanie i inne. Testy te zazwyczaj składają się ze zbiorów danych z predefiniowanymi danymi wejściowymi i oczekiwanymi wynikami, umożliwiając ilościową ocenę wydajności modelu za pomocą wskaźników, takich jak dokładność, wynik F1 lub wynik BLEU, w zależności od zadania. Niektóre kluczowe testy porównawcze LLM to: ugruntowanie, trafność, spójność, płynność i podobieństwo GPT; te wskaźniki oceny przedstawiono w tabeli 5.

Metryka: Skupienie: Metoda: Kiedy stosować?

Ugruntowanie: Ocenia, jak dobrze odpowiedzi generowane przez model są zgodne z informacjami zawartymi w danych źródłowych (kontekst podany przez użytkownika). Ta metryka zapewnia, że kontekst potwierdza odpowiedzi generowane przez sztuczną inteligencję.: Ocenia, jak dobrze stwierdzenia w odpowiedzi wygenerowanej przez sztuczną inteligencję są zgodne z kontekstem źródłowym, zapewniając, że kontekst wspiera te stwierdzenia. Jest oceniana w skali od 1 (źle) do 5 (dobrze).: Jest używana, gdy chcemy sprawdzić, czy odpowiedzi sztucznej inteligencji są zgodne i potwierdzone w danym kontekście. Jest również używana, gdy ważna jest poprawność faktograficzna i precyzja kontekstowa, na przykład podczas wyszukiwania informacji, odpowiadania na pytania i podsumowywania treści.

Spójność: Ocenia zdolność modelu do generowania spójnych, naturalnych wyników, podobnych do języka ludzkiego.: Ocenia, jak dobrze generacja jest ustrukturyzowana i powiązana. Jest oceniana w skali od 1 (źle) do 5 (dobrze). : Użyj go, aby ocenić, jak łatwe i przyjazne dla użytkownika są odpowiedzi generowane przez Twój model w sytuacjach rzeczywistych.

Płynność: Mierzy biegłość gramatyczną i czytelność odpowiedzi generowanej przez model. : Miara płynności ocenia, jak dobrze wygenerowany tekst przestrzega reguł gramatycznych, struktur składniowych i odpowiedniego doboru słów. Jest punktowana od 1 (zła) do 5 (dobra). : To narzędzie ocenia poprawność językową wygenerowanego tekstu, upewniając się, że jest on zgodny z odpowiednimi regułami gramatycznymi, strukturami składniowymi i odpowiednim doborem słów.

Podobieństwo GPT: Porównuje, jak podobne jest zdanie z danych źródłowych (prawdy bazowej) do wyniku z modelu AI. : Ta ocena obejmuje tworzenie osadzeń na poziomie zdań zarówno dla prawdy bazowej, jak i prognozy modelu, które są wielowymiarowymi reprezentacjami wektorowymi kodującymi znaczenie semantyczne i kontekst zdań. :Użyj go, aby uzyskać obiektywny pomiar wydajności modelu, szczególnie w zadaniach generowania tekstu, w których dostępne są poprawne odpowiedzi. Pozwala nam to sprawdzić, jak bardzo wygenerowany tekst odpowiada zamierzonej treści, co pomaga nam ocenić jakość i dokładność modelu.

Aby zilustrować, jak to działa, zastosujemy metodę ewaluacji bez odniesień opartą na metodzie G-Eval. Metoda bez odniesień oznacza, że nie polegamy na porównywaniu wygenerowanego podsumowania z istniejącym już podsumowaniem referencyjnym. Zacznijmy od zrozumienia metody G-Eval.

G-Eval: Podejście pomiarowe do oceny NLG

G-Eval wprowadza nowe ramy pomiaru jakości tekstu generowanego przez systemy NLG. Wykorzystując modele LLM, G-Eval łączy metodę podpowiedzi myślowych z techniką wypełniania formularzy, aby zbadać różne aspekty wyników NLG, takie jak spójność, konsekwencja i trafność. G-Eval ocenia jakość generowanej treści na podstawie samego podpowiedzi i tekstu, bez żadnych tekstów referencyjnych, i dlatego jest uznawany za wolny od odniesień. Metoda ta jest szczególnie przydatna w przypadku nowych zestawów danych i zadań z niewielką liczbą dostępnych odniesień ludzkich. Ta elastyczność sprawia, że G-Eval nadaje się do różnych innowacyjnych zastosowań, zwłaszcza w dziedzinach, w których dane stale ewoluują lub są bardzo specyficzne. Oto kilka scenariuszy, w których G-Eval byłby przydatny:

* Generowanie raportów medycznych — w dziedzinie medycyny, gdzie zautomatyzowane systemy generują spersonalizowane raporty na podstawie różnych danych pacjentów, G-Eval może oceniać raporty pod kątem poprawności, spójności i trafności medycznej. Ponieważ scenariusze pacjentów są bardzo zróżnicowane, konwencjonalne wskaźniki oparte na referencjach nie zawsze się sprawdzają, co sprawia, że G-Eval jest bardziej elastycznym i odpowiednim rozwiązaniem, gwarantującym jakość i rzetelność raportów medycznych.

* Tworzenie dokumentów prawnych — Kiedy sztuczna inteligencja tworzy dokumenty prawne dostosowane do konkretnych przypadków, G-Eval ocenia, jak dobrze dokumenty spełniają wymogi prawne, jak są jasne i spójne oraz jak dobrze przestrzegają przepisów. Jest to ważne w sytuacjach prawnych, w których posiadanie precyzyjnych tekstów referencyjnych dla każdej sytuacji jest niemożliwe, a dokładność i zgodność z normami prawnymi są kluczowe.

* Ocena kreatywnych treści — Nowość jest niezbędna w dziedzinach wymagających kreatywności, takich jak reklama czy opowiadanie historii w grach wideo. G-Eval pomaga ocenić nowość, atrakcyjność i dopasowanie takich treści do grupy docelowej, zapewniając sposób oceny jakości kreatywności, która wykracza poza samo podobieństwo słów lub fraz.

* Moderacja treści oparta na sztucznej inteligencji — G-Eval może pomóc w weryfikacji, czy działania moderacyjne są odpowiednie i skuteczne, nawet w przypadku braku wiarygodnych danych

referencyjnych, wykorzystując systemy sztucznej inteligencji do moderowania zmieniających się treści online. Jest to szczególnie ważne w środowiskach online, gdzie kontekst i wrażliwość mają znaczenie.

Poniższe przykłady pokazują, jak G-Eval może oceniać jakość tekstu generowanego przez sztuczną inteligencję, stosując standardy zbliżone do ludzkich i elastycznie dostosowując się do zróżnicowanych potrzeb. Jest to istotne w przypadku aplikacji GenAI, w których konwencjonalne metryki są niewystarczające. G-Eval oferuje wiele zalet firmom, które chcą rozwijać i wykorzystywać efektywne rozwiązania NLG:

* G-Eval wykazuje znacznie lepszą zgodność z oceną ludzką niż konwencjonalne wskaźniki, takie jak BLEU i ROUGE. Jest to szczególnie widoczne w przypadku otwartych i kreatywnych zadań NLG, w których konwencjonalne wskaźniki często zawodzą. Zapewniając dokładniejszy pomiar jakości systemu NLG, G-EVAL pomaga przedsiębiorstwom podejmować mądre decyzje dotyczące ich rozwoju i wdrażania.

* G-Eval wykorzystuje prawdopodobieństwa tokenów wyjściowych z LLM do generowania precyzyjnych, ciągłych wyników. Umożliwia to wychwycenie drobnych różnic między generowanymi tekstami, zapewniając bardziej szczegółowe informacje zwrotne niż tradycyjne wskaźniki, które często opierają się na punktacji dyskretnej. Tak precyzyjne informacje zwrotne mogą być bardzo pomocne dla przedsiębiorstw, które dostosowują swoje systemy NLG w celu uzyskania najlepszej wydajności.

* Ciekawą cechą G-Eval jest możliwość dostosowania go do oceny różnych zadań NLG poprzez zmianę podpowiedzi i kryteriów oceny. Ta elastyczność eliminuje potrzebę zatrudniania ewaluatorów specyficznych dla danego zadania, ułatwiając proces oceny przedsiębiorstwom korzystającym z różnych aplikacji NLG.

Potencjalnym problemem dla ewaluatorów opartych na LLM jest jednak to, że mogą oni preferować tekst generowany przez LLM. Ten problem wymaga dalszych badań i rozwiązań, aby zapewnić rzetelną i poprawną ocenę, zwłaszcza w przypadku wykorzystania metryk opartych na LLM do ulepszania systemów NLG.

G-Eval oferuje potencjalną metodę oceny NLG w przedsiębiorstwach, która może pomóc w tworzeniu i wykorzystywaniu bardziej wydajnych i niezawodnych systemów NLG do różnych celów. Zobaczmy, jak możemy to wykorzystać. Możemy zademonstrować, jak bardzo pomocny może być G-Eval, na prostym przykładzie. Wyobraźmy sobie przedsiębiorstwo, które chce ocenić chatboty obsługi klienta. Te chatboty są zazwyczaj szkolone do radzenia sobie z wieloma rodzajami pytań i problemów klientów, a ich wydajność jest niezbędna do utrzymania satysfakcji i lojalności klientów. Wyobraźmy sobie na przykład skargę klienta dotyczącą usługi. Załóżmy, że klient pisze następującą skargę w e-mailu:

Jestem bardzo rozczarowany opóźnieniem w obsłudze. Obiecano mi dwudniową dostawę, a minął już tydzień bez żadnych aktualizacji. To niedopuszczalne.

Wyobraź sobie teraz dwie różne automatyczne odpowiedzi generowane przez boty obsługi klienta:

* Odpowiedź A (bardziej dosłowna i ogólna) – „Przepraszamy za wszelkie niedogodności. Twoja skarga została zarejestrowana. Wkrótce Cię poinformujemy”.

* Odpowiedź B (bardziej empatyczna i konkretna) – „Bardzo nam przykro z powodu opóźnienia i w pełni rozumiemy Twoją frustrację. Nie chcemy zapewniać Ci takiego doświadczenia. Nasz zespół traktuje to priorytetowo i skontaktujemy się z Tobą z aktualizacją dotyczącą dostawy do jutra rano”.

Konwencjonalne wskaźniki, takie jak BLEU i ROUGE, prawdopodobnie oceniałyby te odpowiedzi na podstawie tego, jak bardzo określone słowa lub frazy pasują do zestawu predefiniowanych poprawnych odpowiedzi. Odpowiedź A mogłaby uzyskać całkiem dobry wynik, gdyby odpowiedzi referencyjne faworyzowały ogólne podziękowania. Jednak te wskaźniki mogłyby nie uwzględniać niuansów w tonie i szczegółowości, które są kluczowe dla satysfakcji klienta. Podczas oceny za pomocą G-Eval bardziej prawdopodobne byłoby sprawdzenie treści, tonu, empatii i adekwatności odpowiedzi do konkretnej skargi. Uwzględniono by, jak skutecznie odpowiedź odnosi się do stanu emocjonalnego klienta i poruszonego problemu. W naszym przykładzie odpowiedź B prawdopodobnie uzyskałaby wyższy wynik w teście G-Eval, ponieważ uwzględnia odczucia klienta, zawiera konkretną obietnicę i jasno określa oczekiwania – a wszystko to jest ważne dla ludzkich sędziów (tj. klientów) przy ocenie jakości obsługi klienta. W przedsiębiorstwach, szczególnie w obszarach takich jak obsługa klienta, skuteczność automatycznych odpowiedzi może znacząco wpłynąć na zadowolenie i lojalność klientów. G-Eval lepiej wpisuje się w ocenę ludzką, ponieważ uwzględnia jakościowe aspekty komunikacji, które są ważne w rzeczywistych interakcjach – takie jak empatia, konkretność i poczucie bezpieczeństwa – ale które często są pomijane przez tradycyjne wskaźniki, takie jak BLEU i ROUGE.

Przykład metryk oceny opartych na LLM

W tym przykładzie wdrażamy podejście G-Eval, wykorzystując model GPT-4 platformy Azure OpenAI do pomiaru jakości streszczeń tekstowych. Wykorzystuje on następujące cztery kryteria: trafność, spójność, spójność i płynność. Mamy artykuł i dwa streszczenia na jego podstawie. Dodatkowo, używamy kodu do oceny każdego streszczenia według czterech kryteriów i wskazania, które jest lepsze. Jako przykład, użyliśmy zasad sztucznej inteligencji Fundacji Billa i Melindy Gates, które są wymienione jako „Pierwsze zasady kierujące naszą pracą ze sztuczną inteligencją" i są dostępne online pod adresem https://mng.bz/vJe4. Mamy dwa streszczenia utworzone na podstawie tego artykułu źródłowego, które chcemy porównać z artykułem. Jedno streszczenie tworzy biblioteka NLP, a drugie LLM (Gemini Pro 1.5 firmy Google). Wszystkie te streszczenia zapisaliśmy lokalnie dla łatwego dostępu i czytamy je stamtąd. Pełny kod, z którego pobieramy artykuł i tworzymy streszczenia, znajduje się w repozytorium GitHub książki. Listingi 2 i 3 pokazują kluczowe obszary tego przykładu wraz z pełnym kodem. (https://bit.ly/GenAIBook). Definiujemy metryki oceny, ich kryteria i kroki za pomocą narzędzi Rapid Engineering i RAG. Każdy z nich opisuje kryteria punktacji i kroki, które należy wykonać podczas oceny podsumowania. Należy pamiętać, że ze względu na zwięzłość nie prezentujemy całego kodu.

Wyszczególnienie 12.2: Metryki oceny oparte na LLM

```
EVALUATION_PROMPT_TEMPLATE = """
You will be given one summary written for an article. Your task
is to rate the summary using one metric. Make sure you read
and understand these instructions very carefully.

Evaluation Criteria:
{criteria}

Evaluation Steps:
{steps}
```

Evaluation prompt template based on G-Eval

```
Example:
Source Text:

{document}

Summary:
{summary}

Evaluation Form (scores ONLY):
- {metric_name}
"""
# Metric 1: Relevance
RELEVANCY_SCORE_CRITERIA = """
Relevance(1-5) - selection of important content from the source. \
The summary should include only important information from the
➥source document. \
Annotators were instructed to penalize summaries which contained
➥redundancies and excess information.
"""
RELEVANCY_SCORE_STEPS = """
1. Read the summary and the source document carefully.
2. Compare the summary to the source document and identify the
   ➥main points of the article.
3. Assess how well the summary covers the main points of the article,
   ➥and how much irrelevant or redundant information it contains.
4. Assign a relevance score from 1 to 5.
"""
# Metric 2: Coherence
COHERENCE_SCORE_CRITERIA = """
Coherence(1-5) - the collective quality of all sentences. \
...
COHERENCE_SCORE_STEPS = """
1. Read the article carefully and identify the main topic and key points.
2. Read the summary and compare it to the article. Check if the summary
...
# Metric 3: Consistency
CONSISTENCY_SCORE_CRITERIA = """
Consistency(1-5) - the factual alignment between the summary and the
➥summarized source.
...
CONSISTENCY_SCORE_STEPS = """
1. Read the article carefully and identify its main facts and details.
2. Read the summary and compare it to the article. Check if the
   ➥summary ...
# Metric 4: Fluency
FLUENCY_SCORE_CRITERIA = """
Fluency(1-3): the quality of the summary in terms of grammar, spelling,
     punctuation, word choice, and sentence structure.
...


FLUENCY_SCORE_STEPS = """
Read the summary and evaluate its fluency based on the given criteria.
➥Assign a fluency score from 1 to 3.
...
```

Defines the relevance metric as outlined by G-Eval

Outlines the rules of the relevance metrics and how to measure

Defines the coherence metric as outlined by G-Eval

Outlines the rules of the coherence metric and how to measure

Defines the consistency metric as outlined by G-Eval

Outlines the rules of the consistency metrics and how to measure

Defines the fluency metric as outlined by G-Eval

Outlines the rules of the fluency metrics and how to measure

Wyjaśniliśmy już monity definiujące metryki i zasady ich obliczania. Teraz spójrzmy na resztę kodu na listingu 3. To proste i wykonujemy następujące czynności:

* Użyj funkcji get_article(), aby pobrać artykuł i podsumowania.

* Użyj funkcji get_geval_score(), przeprowadź pętlę po metrykach oceny i podsumowaniach, wygeneruj wynik G-Eval dla każdej kombinacji i zapisz wyniki w słowniku.

* Na koniec przekonwertuj słownik na ramkę danych, aby móc ją przestawić i wyświetlić na konsoli.

UWAGA: Parametry Azure OpenAI są dość rygorystyczne: max_tokens ustawione na 5, temperature na 0, a top_p na 1.

Listing 3: Metryki oceny oparte na LLM

```
def get_article():
    with open('./data/gates_foundation_article.txt', 'r') as f:
        article = f.read()
    with open('./data/gates_foundation_summary1.txt', 'r') as f:
        nlp_summary = f.read()
    with open('./data/gates_foundation_summary2.txt', 'r') as f:
        llm_summary = f.read()
    return article, nlp_summary, llm_summary

def get_geval_score(criteria: str, steps: str, document: str,
➥summary: str, metric_name: str):
    prompt = EVALUATION_PROMPT_TEMPLATE.format(
        criteria=criteria,
        steps=steps,
        metric_name=metric_name,
        document=document,
        summary=summary,
    )

    response = client.chat.completions.create(
        model=MODEL,
        messages = [{"role": "user", "content": prompt}],
        temperature = 0,
        max_tokens = 5,
        top_p = 1,
        frequency_penalty = 0,
        presence_penalty = 0,
        stop = None
    )
    return response.choices[0].message.content
```

Function to load the files from disk

Function to calculate various evaluation metrics

Sets up the prompt for G-Eval

Completion API to run the evaluation

```
evaluation_metrics = {
    "Relevance": (RELEVANCY_SCORE_CRITERIA, RELEVANCY_SCORE_STEPS),
    "Coherence": (COHERENCE_SCORE_CRITERIA, COHERENCE_SCORE_STEPS),
    "Consistency": (CONSISTENCY_SCORE_CRITERIA, CONSISTENCY_SCORE_STEPS),
    "Fluency": (FLUENCY_SCORE_CRITERIA, FLUENCY_SCORE_STEPS),
}

# Main code
client = AzureOpenAI(
    azure_endpoint = AZURE_ENDPOINT,
    api_key=AOAI_API_KEY,
    api_version=API_VERSION
)

article_text, nlp_summary, llm_summary = get_article()

summaries = {"NLP Summary (1)": nlp_summary,
             "LLM Summary (2)": llm_summary}

data = {"Evaluation Score": [], "Summary": [], "Score": []}

print("Starting evaluation...")
for eval_type, (criteria, steps) in evaluation_metrics.items():
    for summ_type, summary in summaries.items():
        data["Evaluation Score"].append(eval_type)
        data["Summary"].append(summ_type)
        result = get_geval_score(criteria, steps, article_text,
                                 summary, eval_type)
        numeric_part = ''.join(filter(str.isdigit,
                        ➥result.strip()))
        if numeric_part:
            score_num = int(float(result.strip()))
            data["Score"].append(score_num)

max_values = {key: max(values) for key, values in data.items()}

df = pd.DataFrame(data)

pivot_df = df.pivot(index='Evaluation Score',
                    columns='Summary',
                    values='Score')

print(pivot_df)
```

Dictionary to store the evaluation results

Loops over the evaluation metrics and summaries

Checks if result is not empty and if it is a number

Evaluation result stored in a dictionary

Converts the dictionary to a Pandas DataFrame

Pivots the DataFrame to allow easy visualization

The output of this is as follows:

```
Starting evaluation...
Summary          LLM Summary (2) NLP Summary (1)
Evaluation Score
Coherence                     5*              1
Consistency                   5*              4
Fluency                       3               2
Relevance                     5*              2
```

Zobaczmy, jak interpretujemy te wyniki i co one oznaczają:

* Spójność mierzy, jak logicznie i płynnie idee przechodzą z jednego zdania do drugiego. Wynik 5 dla streszczenia LLM oznacza, że przedstawia ono informacje w sposób logiczny, jest dobrze zorganizowane i łatwe do zrozumienia dla czytelników. Tradycyjne streszczenie NLP, z wynikiem 1, prawdopodobnie ma problemy z rozproszonymi myślami lub brakuje mu logicznego przepływu, co utrudnia czytelnikom zrozumienie kolejności lub powiązania myśli.

* Spójność odnosi się do braku sprzeczności w tekście i utrzymania tych samych standardów w całym streszczeniu. Wysoki wynik LLM, wynoszący 5, sugeruje utrzymanie jednolitego tonu, stylu i dokładności faktograficznej. Chociaż dobry, wynik 4 dla tradycyjnego NLP wskazuje na drobne problemy z jednolitym utrzymaniem tych elementów.

* Płynność ocenia płynność tekstu i naturalność języka. Wynik 3 dla LLM oznacza umiarkowaną płynność; Język jest generalnie jasny, ale może zawierać pewne niezręczne sformułowania lub złożoność, które mogą utrudniać czytelność. Tradycyjne NLP, uzyskując niższe wyniki, może wykazywać poważniejsze problemy, takie jak błędy gramatyczne lub nienaturalne struktury zdań.

* Trafność mierzy, jak dobrze streszczenie odnosi się do głównych punktów i celu oryginalnej treści. Wynik 5 punktów w przypadku LLM sugeruje, że skutecznie ujmuje ono i koncentruje się na kluczowych elementach oryginalnego tekstu, dostarczając streszczenie, które spełnia potrzeby informacyjne czytelnika. Tradycyjne NLP, z wynikiem 2 punktów, prawdopodobnie zawiera pewne istotne informacje, ale pomija ważne szczegóły lub zawiera nieistotne treści.

Jak widzieliśmy, ocena LLM wymaga więcej niż tradycyjnych zadań i obejmuje trudniejsze testy, które mierzą wyższy poziom zrozumienia, logiki i umiejętności adaptacyjnych. Niektóre z tych testów, takie jak HELM, HEIM, HellaSWAG i MMLU (Massive Multitask Language Understanding), wyróżniają się trudnością i zakresem.

HELM

HELM (Holistic Evaluation of Language Models) [6] to holistyczne ramy oceny modeli fundamentalnych wprowadzone przez Uniwersytet Stanforda. Ramy te mają na celu kompleksową ocenę modeli językowych, koncentrując się na ich możliwościach, ograniczeniach i powiązanym ryzyku. Opracowane w celu zwiększenia przejrzystości modeli, HELM oferuje bardziej szczegółowe zrozumienie ich wydajności w różnych scenariuszach. Kategoryzuje szeroki zakres potencjalnych scenariuszy i metryk istotnych dla modeli językowych. Podzbiór tych scenariuszy i metryk jest następnie oceniany na podstawie ich zasięgu i praktyczności, co gwarantuje, że HELM jest praktycznym i użytecznym narzędziem dla przedsiębiorstw. Podejście HELM wykorzystuje ocenę multimetryczną, oceniając różne czynniki, takie jak dokładność, kalibracja, odporność, uczciwość, stronniczość, toksyczność i wydajność dla każdego wybranego scenariusza. Polega ono na przeprowadzaniu zakrojonych na szeroką skalę ocen różnych modeli językowych w standardowych warunkach, aby zagwarantować porównywalność. Ponadto HELM promuje otwartość, udostępniając publicznie wszystkie podpowiedzi i uzupełnienia modeli do dalszej analizy. Uzupełnieniem jest modułowy zestaw narzędzi, który ułatwia ciągłą benchmarking w ramach społeczności. Dla przedsiębiorstw rozważających wykorzystanie modeli językowych GenAI, HELM dostarcza cennych informacji, które pozwalają na podejmowanie świadomych wyborów. Umożliwia porównywanie różnych modeli pod kątem różnych metryk, pomagając w wyborze najodpowiedniejszego modelu. Ponadto HELM pomaga ograniczać ryzyko poprzez ocenę potencjalnych zagrożeń, takich jak stronniczość i toksyczność, umożliwiając przedsiębiorstwom identyfikację i rozwiązywanie problemów przed ich wdrożeniem w praktyce. Przejrzystość i zaufanie promowane przez HELM poprzez dostęp do surowych prognoz modeli i danych ewaluacyjnych dodatkowo zwiększają zrozumienie i pewność w korzystaniu z modeli językowych w organizacji.

HEIM

Uniwersytet Stanforda wprowadził test porównawczy o nazwie Holistyczna Ocena Modeli Tekstu na Obraz (HEIM), aby zapewnić kompletną analizę ilościową mocnych i słabych stron modeli tekstu na obraz. W przeciwieństwie do innych ocen mierzących dopasowanie tekstu do obrazu i jakość obrazu, HEIM bada 12 ważnych aspektów stosowania modeli w świecie rzeczywistym. Niektóre z tych aspektów to estetyka, oryginalność, rozumowanie, wiedza, stronniczość, toksyczność i wydajność. HEIM stosuje holistyczne podejście do oceny modeli tekstu na obraz, opracowując 62 scenariusze. To holistyczne podejście ujawnia, że żaden pojedynczy model nie jest najlepszy we wszystkich obszarach, podkreślając różne mocne i słabe strony różnych modeli. HEIM powinien być kluczowym kryterium oceny dla przedsiębiorstw, ponieważ zapewnia przejrzysty i ujednolicony sposób oceny modeli tekstu na obraz. Rozumiejąc mocne i słabe strony tych modeli, przedsiębiorstwa mogą podejmować świadome decyzje dotyczące tego, które modele wykorzystać do konkretnych zadań lub usług. Ponadto ocena pomaga zidentyfikować potencjalne zagrożenia, takie jak stronniczość lub toksyczność, które mogą mieć konsekwencje prawne i wizerunkowe dla firm. W związku z tym, dla przedsiębiorstw tworzących i wdrażających aplikacje GenAI w środowisku produkcyjnym, benchmark HEIM oferuje cenne informacje w następujących czterech wymiarach:

* Wybór modelu — HEIM podkreśla, że żaden pojedynczy model nie wyróżnia się we wszystkich aspektach. Przedsiębiorstwa muszą starannie oceniać i wybierać modele w oparciu o specyficzne wymagania swojej aplikacji. Na przykład aplikacje skoncentrowane na twórczości artystycznej mogą priorytetowo traktować estetykę i oryginalność, podczas gdy te wymagające dokładności faktograficznej mogą koncentrować się na spójności i wiedzy.

* Ograniczanie ryzyka — HEIM kładzie nacisk na ocenę stronniczości, toksyczności i uczciwości. Przedsiębiorstwa muszą zadbać o to, aby ich aplikacje były etycznie poprawne i unikały utrwalania szkodliwych stereotypów lub generowania nieodpowiednich treści. Wymaga to starannego doboru modelu, jego dopracowania i wdrożenia środków bezpieczeństwa.

* Optymalizacja wydajności — ocena racjonalności, solidności i wielojęzyczności ma kluczowe znaczenie dla zapewnienia niezawodności aplikacji i satysfakcji użytkowników. Przedsiębiorstwa muszą wybierać modele, które będą działać skutecznie w różnych scenariuszach i przy różnych wnioskach użytkowników.

* Kwestie wydajności — wydajność generowania obrazu wpływa na komfort użytkowania i koszty operacyjne. Przedsiębiorstwa powinny rozważyć kompromisy między rozmiarem modelu, szybkością i wymaganiami dotyczącymi zasobów przy wyborze i wdrażaniu modeli.

Pokażemy przykład i wyjaśnimy, jak można zastosować HEIM do analizy modeli. Jak wspomniano wcześniej, HEIM ocenia modele tekst-obraz, tworząc scenariusze obejmujące 12 mierzonych aspektów. Gdybyśmy chcieli przetestować jeden z tych 12 aspektów – aspekt dopasowania tekstu do obrazu – HEIM mógłby użyć scenariusza, w którym model otrzymuje złożony komunikat tekstowy, a następnie sprawdza, jak dobrze generowany obraz pasuje do szczegółów i kontekstu komunikatu. Oto możliwy scenariusz ewaluacji:

* Komunikat – „Futurystyczny pejzaż miejski o zmierzchu z latającymi samochodami i neonami odbijającymi się w wodzie poniżej”.

* Generowanie modelu – Model generuje obraz na podstawie komunikatu.

* Ewaluacja – Ewaluatorzy lub zautomatyzowane metryki oceniają obraz pod kątem różnych czynników:

– Czy obraz wiernie przedstawia pejzaż miejski o zmierzchu?

– Czy widać latające samochody i neony, tak jak opisano?

– Czy w wodzie widać odbicie i jak realistycznie ono wygląda?

– Jak dobrze obraz jest dopasowany do komunikatu?

Wydajność modelu w tym scenariuszu wpłynęłaby na jego ogólny wynik w zakresie dopasowania tekstu do obrazu. Podobne scenariusze zostałyby opracowane dla innych aspektów, takich jak jakość obrazu, oryginalność, racjonalność itd. Kompleksowa ocena obejmująca wszystkie 12 aspektów pozwala na wgląd w możliwości i ograniczenia modelu. Podejście HEIM gwarantuje, że modele są oceniane pod kątem ich zdolności do generowania atrakcyjnych wizualnie obrazów oraz zrozumienia tekstu, kreatywności oraz potencjalnych uprzedzeń lub kwestii etycznych. Ta holistyczna ocena ma kluczowe znaczenie dla przedsiębiorstw, ponieważ pomaga im wybrać modele zgodne z ich wartościami i potrzebami, jednocześnie uwzględniając związane z tym ryzyko.

HellaSWAG

HellaSWAG to wymagający test porównawczy, który testuje zdolność modeli sztucznej inteligencji do rozumowania opartego na zdrowym rozsądku. Jest on ulepszeniem poprzedniej wersji, SWAG, poprzez dodanie bardziej zróżnicowanego i złożonego zestawu pytań wielokrotnego wyboru, które wymagają od modeli czegoś więcej niż tylko przetwarzania języka. HellaSWAG to zadanie, w którym każde pytanie przedstawia scenariusz z czterema możliwymi zakończeniami, a LLM musi wybrać najbardziej pasujące zakończenie spośród opcji. Na przykład, poniższe pytanie pochodzi z zestawu danych HellaSWAG. Pytanie składa się z kontekstu podanego LLM i czterech opcji, które są możliwymi zakończeniami dla tego kontekstu. Tylko jedna z tych opcji ma sens w rozumowaniu opartym na zdrowym rozsądku. W tym przykładzie wyróżniono opcję C:

Kobieta jest na zewnątrz z wiadrem i psem. Pies biega, próbując uniknąć kąpieli. Ona…

A. spłukuje wiadro mydłem i suszy głowę psa suszarką.

B. używa węża ogrodowego, aby zapobiec namydleniu.

C. moczy psa, a ten znowu ucieka.

D. wchodzi do wanny z psem.

Wybory modelu są porównywane z poprawnymi odpowiedziami, aby zmierzyć jego skuteczność. Ta metoda testowania ocenia znajomość niuansów językowych przez LLM oraz jego głębsze zrozumienie logiki zdroworozsądkowej i złożoności sytuacji z życia wziętych. Dobry wynik w teście HellaSWAG oznacza, że model posiada zdolność rozumienia niuansów i rozumowania, co jest niezbędne w zastosowaniach wymagających zaawansowanego rozumienia kontekstu i logiki.

Rozumienie języka w ramach Massive Multitask

Test Massive Multitask Language Understanding (MMLU) [9] ocenia zakres i głębokość wiedzy absolwenta studiów magisterskich (LLM) na różne tematy i dziedziny. MMLU wyróżnia się tym, że obejmuje setki zadań powiązanych z różnymi obszarami wiedzy, od nauk ścisłych i literatury po historię i nauki społeczne. MMLU został opracowany w odpowiedzi na obserwację, że podczas gdy wiele modeli językowych osiągało znakomite wyniki w zadaniach NLP, miały trudności z rozumieniem języka naturalnego (NLU). Wcześniejsze testy, takie jak GLUE i SuperGLUE, zostały szybko opanowane przez absolwentów studiów magisterskich (LLM), co wskazywało na potrzebę bardziej rygorystycznych

testów. MMLU miał na celu wypełnienie tej luki, testując rozumienie języka i umiejętności rozwiązywania problemów z wykorzystaniem wiedzy zdobytej podczas szkolenia. Test obejmuje pytania z różnych przedmiotów, w tym nauk humanistycznych, społecznych, ścisłych i innych specjalistycznych, od podstawowego do zaawansowanego poziomu zawodowego. To podejście było unikalne, ponieważ większość testów NLU w tamtym czasie koncentrowała się na wiedzy elementarnej. MMLU dążył do przesunięcia granic, testując wiedzę specjalistyczną, co stanowiło nowe wyzwanie dla absolwentów studiów magisterskich (LLM2). Podobnie jak HellaSWAG, MMLU często wykorzystuje format testu wielokrotnego wyboru, w którym model musi wskazać właściwą odpowiedź spośród zestawu opcji. Ogólna dokładność w tych zróżnicowanych zadaniach pokazuje ogólną wydajność modelu, kompleksowo mierząc jego rozumienie języka i zastosowanie wiedzy w różnych dziedzinach. Wysoka wydajność w MMLU oznacza, że model dysponuje dużą ilością informacji i potrafi ją wykorzystać do poprawnego udzielania odpowiedzi na pytania i problemy. Ten szeroki zakres zrozumienia jest niezbędny do tworzenia programów LLM, które potrafią subtelnie i informacyjnie radzić sobie ze złożonością ludzkiej wiedzy i języka. Choć MMLU jest kompleksowy, napotyka na kilka ograniczeń. Po pierwsze, wydajność modeli językowych w tym teście może być ograniczona przez różnorodność i jakość danych treningowych. Jeśli pewne tematy są niedoreprezentowane, modele mogą nie działać prawidłowo w tych obszarach. Ponadto MMLU ocenia przede wszystkim, czy modele mogą generować poprawne odpowiedzi, ale nie ocenia sposobu, w jaki dochodzą do tych wniosków. Jest to kluczowe w zastosowaniach, w których proces rozumowania jest tak samo ważny jak wynik. Kolejnym istotnym problemem jest potencjalne stronniczość w teście. Ponieważ MMLU jest konstruowany w oparciu o różne źródła, może nieumyślnie uwzględniać stronniczość wynikającą z tych materiałów, wpływając na rzetelność ocen modeli, zwłaszcza w przypadku drażliwych tematów. Ponadto istnieje ryzyko, że modele mogą nadmiernie dopasowywać się do konkretnego formatu i stylu MMLU, optymalizując wydajność testu zamiast rzeczywistego zrozumienia i zastosowania w rzeczywistych scenariuszach. Co więcej, wymagania logistyczne związane z przeprowadzeniem tak kompleksowego testu są znaczne i wymagają znacznych zasobów obliczeniowych, które mogą nie być dostępne dla wszystkich badaczy. To ograniczenie może ograniczyć zakres spostrzeżeń uzyskanych z testu. Wreszcie, skalowalność wiedzy stanowi wyzwanie; wraz z ewolucją dziedzin, test musi być regularnie aktualizowany, aby zachować aktualność, co wymaga ciągłych inwestycji w zasoby. Czynniki te uwypuklają złożoność korzystania z MMLU jako benchmarku i podkreślają potrzebę ciągłego udoskonalania w celu utrzymania jego skuteczności i trafności.

Korzystanie z Azure AI Studio do ewaluacji

Jako klient platformy Azure możesz łatwo korzystać z tych ewaluacji, ponieważ są one już zintegrowane z Azure AI Studio. Zawiera ono narzędzia do rejestrowania, wyświetlania i analizowania szczegółowych metryk ewaluacji, z opcją korzystania z niestandardowych przepływów ewaluacji i uruchomień wsadowych bez ewaluacji. Dzięki AI Studio możemy przeprowadzić ewaluację z zestawu danych testowych lub przepływu z gotowymi metrykami ewaluacji. Aby uzyskać większą elastyczność, możemy utworzyć własny przepływ ewaluacji. Zanim będziesz mógł użyć metryk wspomaganych przez sztuczną inteligencję do ewaluacji swojego modelu, upewnij się, że masz przygotowane następujące rzeczy:

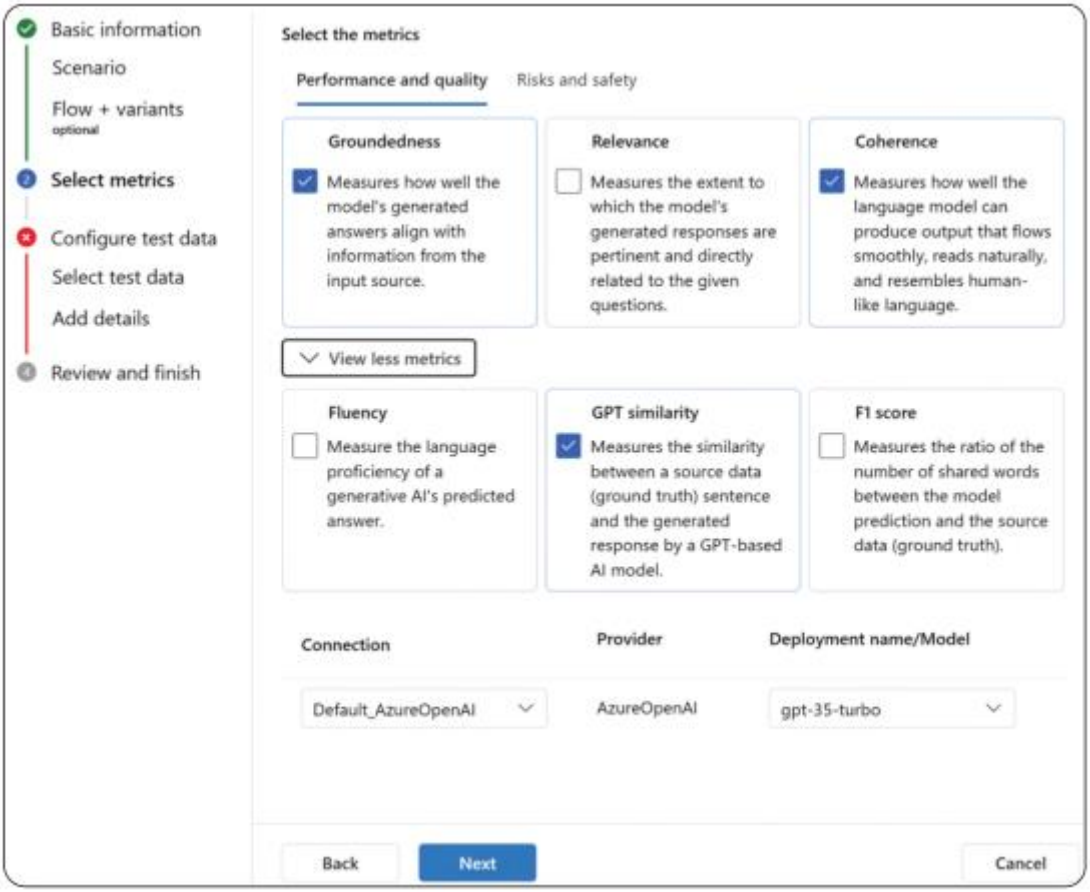
* Zestaw danych testowych w formacie CSV lub JSONL. Jeśli nie masz gotowego zestawu danych, możesz również wprowadzić dane ręcznie z interfejsu użytkownika.

* Wdrożono jeden z następujących modeli: modele GPT-3.5, modele GPT-4 lub modele Davinci.

* Środowisko wykonawcze instancji obliczeniowej do uruchomienia oceny.

Rysunek przedstawia przykład konfiguracji różnych testów.

eśli nie korzystasz z platformy Azure, dostępne są inne podobne opcje, takie jak DeepEval (patrz następna sekcja). To środowisko ewaluacyjne LLM o otwartym kodzie źródłowym umożliwia uruchamianie wielu metryk LLM i znacznie ułatwia ten proces.



DeepEval: Framework do oceny LLM

DeepEval to darmowy framework do oceny LLM, który działa podobnie jak pytest (popularny framework do testowania w Pythonie), ale koncentruje się na testowaniu jednostkowym wyników LLM. DeepEval wykorzystuje najnowsze badania do oceny wyników LLM w oparciu o metryki takie jak halucynacja, trafność odpowiedzi i RAGAS (Retrieval Augmented Generation Assessment). Metryki te opierają się na modelach LLM i innych modelach NLP, które są uruchamiane lokalnie na komputerze użytkownika. DeepEval obsługuje wiele przydatnych funkcji dla aplikacji korporacyjnych. Umożliwia jednoczesną ocenę całych zestawów danych, tworzenie niestandardowych metryk, porównywanie dowolnych modeli LLM z popularnymi benchmarkami oraz ocenę w czasie rzeczywistym w środowisku produkcyjnym. Ponadto współpracuje z narzędziami takimi jak LlamaIndex i Hugging Face i automatycznie łączy się z platformą Confident AI w celu bieżącej oceny aplikacji LLM przez cały okres jej użytkowania. Platforma oferuje również platformę do rejestrowania wyników testów, pomiaru zaliczeń/niezaliczeń metryk, wybierania i porównywania najlepszych hiperparametrów, organizowania przypadków testowych/zestawów danych ewaluacyjnych oraz monitorowania odpowiedzi LLM na żywo w środowisku produkcyjnym.

```
Command Prompt
========================================== test session starts ==========================================
platform win32 -- Python 3.11.3, pytest-8.0.0, pluggy-1.4.0 -- C:\Users\ambahree\Miniconda3\envs\book-test\python.exe
cachedir: .pytest_cache
rootdir: C:\src\Book-GenAI\book-GenAI-scratchpad\chapters\ch11
plugins: deepeval-0.21.0, anyio-4.2.0, Faker-23.2.1, repeat-0.9.3, subtests-0.11.0, xdist-3.5.0
collected 1 item

✨ You're running DeepEval's latest Answer Relevancy Metric! (using gpt-4-0125-preview, strict=False, async_mode=True)...
PASSEDRunning teardown with pytest sessionfinish ...


========================================== slowest 10 durations ==========================================
19.94s call     test_deepeval.py::test_answer_relevancy
0.80s setup    test_deepeval.py::test_answer_relevancy

(1 durations < 0.005s hidden.  Use -vv to show these durations.)
========================================== 1 passed, 5 warnings in 21.04s ==========================================
                                          Test Results
```

Test case	Metric	Score	Status	Overall Success Rate
test_answer_relevancy	Answer Relevancy	1.0 (threshold=0.5, evaluation model=gpt-4-0125-preview, reason=The score is 1.00 because the response directly addresses your concern without any irrelevant information. Great job on keeping the focus!)	PASSED	100.0%

```
✅ Tests finished! View results on
https://app.confident-ai.com/project/clu6j1clt0003jo0fjx43xchs/unit-tests/clu6jwscb0001l40g4umbcjiw/test-cases
```

Nowe testy porównawcze

W ciągu ostatnich 12–18 miesięcy zaobserwowaliśmy, że modele sztucznej inteligencji osiągnęły poziom nasycenia wydajnością w uznanych branżowych testach porównawczych, takich jak ImageNet, SQuAD i SuperGLUE, żeby wymienić tylko kilka. To skłoniło branżę do opracowania bardziej wymagających testów porównawczych. Niektóre z nowszych to SWE-bench do kodowania, MMMU do rozumowania ogólnego, MoCa do rozumowania moralnego i HaluEval do halucynacji.

SWE-bench

Aby zmierzyć postęp systemów GenAI, które potrafią kodować, potrzebujemy trudniejszych zadań do ich oceny. SWE-bench to zbiór danych zawierający niemal setki problemów inżynierii oprogramowania z rzeczywistych repozytoriów GitHub i Python. Stanowi on trudniejsze wyzwanie dla umiejętności kodowania sztucznej inteligencji, wymagając od systemów wprowadzania zmian w wielu funkcjach, obsługi różnych środowisk wykonawczych i przeprowadzania złożonego rozumowania. Zbiór danych SWE-bench ocenia zdolność systemów do automatycznego rozwiązywania problemów GitHub. Gromadzi 2294 pary problem-pull request z 12 popularnych repozytoriów Pythona. Ocena odbywa się poprzez weryfikację proponowanych rozwiązań za pomocą testów jednostkowych i porównanie ich z zachowaniem po PR jako rozwiązaniem referencyjnym. Podstawową miarą oceny dla SWE-bench jest odsetek rozwiązanych instancji zadań. Innymi słowy, mierzy ona skuteczność modelu w rozwiązywaniu danych problemów – im wyższy odsetek rozwiązanych instancji, tym lepsza wydajność modelu.

MMMU

MMMU (Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark) to nowy benchmark zaprojektowany do oceny możliwości modeli multimodalnych w zadaniach wymagających wiedzy przedmiotowej na poziomie uniwersyteckim i rozumowania na poziomie eksperckim z różnych dyscyplin. Zawiera 11,5 tys. pytań multimodalnych z egzaminów uniwersyteckich, quizów i podręczników, obejmujących sześć głównych dyscyplin: sztukę i design, biznes, nauki ścisłe, zdrowie i medycynę, nauki humanistyczne i społeczne oraz technologię i inżynierię. Pytania te obejmują 30 przedmiotów i 183 poddziedziny, obejmując 30 wysoce heterogenicznych typów obrazów, takich jak

wykresy, diagramy, mapy, tabele, nuty i struktury chemiczne. MMMU jest wyjątkowy, ponieważ koncentruje się na zaawansowanej percepcji i rozumowaniu, wykorzystując wiedzę z danej dziedziny i wymagające modele do wykonywania zadań, z którymi mierzą się eksperci. Test porównawczy został wykorzystany do oceny kilku otwartych programów LLM i zastrzeżonych modeli, takich jak GPT-4V, co uwydatnia istotne wyzwania, jakie stawia MMMU. Nawet zaawansowane modele osiągają dokładność jedynie na poziomie 56%–59%, co wskazuje na znaczne możliwości poprawy. Test działa poprzez ocenę zdolności absolwentów LLM do postrzegania, rozumienia i rozumowania w różnych dyscyplinach i poddziedzinach, z wykorzystaniem różnych typów obrazów. Test koncentruje się na trzech podstawowych umiejętnościach absolwentów LLM: percepcji, wiedzy i rozumowaniu. Warto zauważyć, że mogłoby się wydawać, że omówiony wcześniej model MMLU jest tym samym, co MMMU; jednak różnią się one od siebie. MMLU ocenia modele językowe w szerokim zakresie zadań tekstowych w różnych dziedzinach, koncentrując się wyłącznie na rozumieniu języka. Natomiast MMMU ocenia modele multimodalne, wymagając zarówno zrozumienia wizualnego, jak i tekstowego w różnych specjalistycznych dyscyplinach, co stanowi wyzwanie dla modeli o złożonej, specyficznej dla danej dziedziny treści multimodalnej. Test porównawczy MMMU stawia kilka kluczowych wyzwań dla modeli multimodalnych, które obejmują:

* Kompleksowość — ponieważ test obejmuje szeroki wachlarz 11,5 tys. problemów na poziomie uniwersyteckim z szeroko pojętych dyscyplin, modele muszą posiadać szeroką bazę wiedzy i zrozumienie w wielu dziedzinach.

* Wysoce heterogeniczne typy obrazów — pytania obejmują 30 różnych typów obrazów, takich jak wykresy, diagramy, mapy, tabele, nuty i struktury chemiczne. Oznacza to, że modele muszą potrafić interpretować i rozumieć różne informacje wizualne.

* Przeplatany tekst i obrazy — wiele pytań zawiera mieszankę tekstu i obrazów, co wymaga od modeli przetwarzania i integrowania informacji z obu modalności w celu uzyskania poprawnej odpowiedzi.

* Percepcja i rozumowanie na poziomie eksperckim — zadania wymagają głębokiej wiedzy przedmiotowej i rozumowania na poziomie eksperckim, podobnie jak wyzwania, przed którymi stoją eksperci w swoich dziedzinach.

Wyzwania te mają na celu poszerzenie granic istniejących modeli multimodalnych, testując ich zdolność do zaawansowanej percepcji, myślenia analitycznego i rozumowania specyficznego dla danej dziedziny. Pytania wymagają głębokiego zrozumienia tematu i umiejętności wykorzystania wiedzy w złożonych scenariuszach, nawet od ekspertów.

MoCa

Ramka MoCa (Measuring Human-Language Model Alignment on Causal and Moral Judgment Tasks) ocenia, jak dobrze LLM są zgodne z ludzkimi uczestnikami w wydawaniu osądów przyczynowych i moralnych dotyczących scenariuszy opartych na tekście. Modele sztucznej inteligencji (AI) mogą dobrze radzić sobie z zadaniami językowymi i wizualnymi, ale ich zdolność do podejmowania decyzji moralnych, zwłaszcza tych zgodnych z ludzkimi opiniami, jest niejasna. Aby zbadać ten temat, grupa naukowców ze Stanford stworzyła nowy zbiór danych (MoCa) ludzkich historii z aspektami moralnymi. Przyjrzymy się szczegółom każdej z nich:

* Sądy przyczynowe — ludzie intuicyjnie rozumieją wydarzenia, osoby i otaczający ich świat, organizując swoje rozumienie w intuicyjne teorie. Teorie te pomagają nam wnioskować o tym, jak obiekty i podmioty oddziałują na siebie, w tym o koncepcjach związanych z przyczynowością. Struktura MoCa gromadzi zbiór danych historii z prac z zakresu kognitywistyki i adnotuje każdą historię badanymi

przez nią czynnikami. Następnie sprawdza, czy osoby z LLM wydają sądy przyczynowe dotyczące scenariuszy tekstowych, które są zgodne z ludzkimi. Na poziomie zbiorczym, zgodność poprawiła się w przypadku nowszych LLM. Jednak analizy statystyczne ujawniają, że osoby z LLM ważą różne czynniki w inny sposób niż uczestnicy badań.

* Sądy moralne — zadania oceniają podmioty w tekście narracyjnym pod kątem rozumowania moralnego. Te zadania i zbiory danych różnią się strukturą, od swobodnych anegdot po bardziej ustrukturyzowane dane wejściowe. Model MoCa ocenia, jak dobrze modele LLM są zgodne z ludzkimi intuicjami moralnymi w tych scenariuszach.

Głównym wskaźnikiem używanym w MoCa jest krzywa obszaru pod krzywą charakterystyki operacyjnej odbiornika (AuROC), która mierzy zgodność modeli LLM z ludzkimi osądami. Co więcej, dokładność służy jako drugorzędny wskaźnik do porównywania modeli. Wyższy wynik wskazuje na bliższe dopasowanie do ludzkiego osądu moralnego. Badanie przyniosło intrygujące wyniki. Żaden model nie jest idealnie dopasowany do ludzkich systemów moralnych. Jednak nowsze, większe modele, takie jak GPT-4 i Claude, wykazują większą zgodność z ludzkimi nastrojami moralnymi niż mniejsze modele, takie jak GPT-3, co sugeruje, że wraz ze skalowaniem modeli sztucznej inteligencji, stają się one stopniowo bardziej zgodne z ludzkimi moralnymi przekonaniami.

Podsumowując, MoCa dostarcza wglądu w to, jak LLM radzą sobie z rozumowaniem przyczynowo-skutkowym i moralnym, rzucając światło na ich ukryte tendencje i zgodność (lub jej brak) z ludzką intuicją.

HaluEval

HaluEval to platforma ewaluacyjna na dużą skalę, zaprojektowana do oceny skuteczności LLM w rozpoznawaniu halucynacji. W tym kontekście halucynacje odnoszą się do treści generowanych przez LLM, które są sprzeczne ze źródłem lub nie mogą zostać zweryfikowane na podstawie wiedzy faktycznej. Platforma obejmuje zbiór wygenerowanych i opatrzonych adnotacjami ludzkimi próbek halucynacji. Do tworzenia tych próbek wykorzystuje się dwuetapowy framework obejmujący próbkowanie, a następnie filtrowanie, często w oparciu o odpowiedzi z modeli takich jak ChatGPT. Osoby oznaczające również wnoszą swój wkład poprzez adnotowanie halucynacji w odpowiedziach. Wyniki empiryczne HaluEval sugerują, że LLM, w tym ChatGPT, mogą generować halucynacje, szczególnie na określone tematy, poprzez tworzenie nieweryfikowalnych informacji. Badanie analizuje również skuteczność obecnych LLM w wykrywaniu halucynacji. Może to prowadzić do wykrywania halucynacji w zadaniach takich jak dialog oparty na wiedzy i pytaniach, a także streszczanie tekstu. Wyniki pokazują, że wielu LLM ma trudności z tymi zadaniami, co podkreśla, że halucynacje są poważnym, uporczywym problemem. Test HaluEval obejmuje 5000 ogólnych zapytań użytkowników z odpowiedziami ChatGPT oraz 30 000 przykładów dotyczących konkretnych zadań, z trzech zadań: odpowiadania na pytania, dialogu opartego na wiedzy i streszczania tekstu. To znaczący krok w kierunku zrozumienia i poprawy niezawodności LLM w generowaniu dokładnych i weryfikowalnych treści.

Ocena przez człowieka

Ocena przez człowieka odgrywa kluczową rolę w zrozumieniu jakości studiów magisterskich (LLM), ponieważ wychwytuje niuanse, kontekst i potencjalne błędy, które mogą zostać pominięte przez automatyczne metryki. Aby przedsiębiorstwa mogły skutecznie przeprowadzać oceny przez człowieka, muszą zacząć od zdefiniowania jasnych kryteriów, którymi będą się kierować przy ocenie wyników studiów magisterskich (LLM). Kryteria te powinny obejmować takie aspekty, jak dokładność, płynność, trafność i obecność błędów. Aby zapewnić spójność i obiektywizm, przedsiębiorstwa powinny

opracować kompleksowe wytyczne i rubryki, którymi będą kierować się ewaluatorzy. Wybierając odpowiednie metody oceny, przedsiębiorstwa mają kilka możliwości. Mogą zaangażować ekspertów dziedzinowych lub przeszkolonych adnotatorów do szczegółowych ocen albo skorzystać z platform crowdsourcingowych, takich jak Amazon Mechanical Turk , aby uzyskać dostęp do szerszego grona ewaluatorów. Kolejnym krokiem jest gromadzenie i adnotacja danych, co wymaga przyjaznych dla użytkownika interfejsów i jasnych instrukcji, aby zapewnić jakość i spójność. Ważne jest również zebranie wystarczającej ilości danych, aby uzyskać statystycznie istotne wyniki. Po zebraniu danych konieczna jest dogłębna analiza. Przedsiębiorstwa powinny stosować metody statystyczne do pomiaru zgodności między oceniającymi i potwierdzania wiarygodności ocen. Wnioski uzyskane w tym procesie powinny następnie zostać wykorzystane do iteracyjnego wprowadzania ulepszeń w modelach LLM, w tym do dostosowania danych szkoleniowych, architektury modelu i szybkiej inżynierii. Regularne oceny dokonywane przez ludzi są niezbędne do monitorowania postępów i wskazywania obszarów wymagających dalszych ulepszeń. Chociaż ocena dokonywana przez ludzi jest nieoceniona, wiąże się z pewnymi wyzwaniami. Może być kosztowna i czasochłonna, zwłaszcza w przypadku dużych zbiorów danych. Istnieje również ryzyko subiektywności i stronniczości w ludzkich osądach. Problemy te można jednak złagodzić, zapewniając jasne wytyczne, odpowiednie szkolenia dla oceniających oraz stosując właściwe metody agregacji. Dostępnych jest wiele narzędzi i platform, które usprawniają proces oceny dokonywanej przez ludzi. Platformy crowdsourcingowe zapewniają dostęp do zróżnicowanej kadry, a narzędzia do adnotacji oferują wydajne funkcje etykietowania danych. Dostępne są również ramy ewaluacyjne, w tym biblioteki metryk i skryptów zaprojektowane specjalnie do ewaluacji LLM i wspierania ewaluacji przez ludzi. Przykładami narzędzi wspomagających adnotacje są Label Studio (https://labelstud.io), które oferuje zarówno rozwiązania open source, jak i korporacyjne. Prodigy (https://prodi.gy) to kolejne narzędzie do adnotacji obsługujące tekst, obrazy, filmy i dźwięk. Istnieją również narzędzia do adnotacji wyłącznie tekstowych, takie jak Labellerr (https://www.labellerr.com) i LightTag (https://www.lighttag.io). Niektóre firmy specjalizują się w ewaluacji LLM, oferując solidne ramy testowe i różnorodne zasoby wspomagające proces ewaluacji. Dla przedsiębiorstw, które nie czują się komfortowo wdrażając swoje ramy ewaluacyjne za pomocą narzędzi takich jak Prompt-Flow, które omówiliśmy wcześniej w poprzednim rozdziale, Weights and Biases (https://wandb.ai) i tak dalej, istnieją nowe firmy specjalizujące się w ewaluacji LLM, takie jak Giskard (https://www.giskard.ai). Postępując zgodnie z tymi krokami i wykorzystując dostępne zasoby, przedsiębiorstwa mogą wdrożyć ustrukturyzowany i skuteczny proces oceny pracowników w ramach studiów magisterskich z prawa (LLM). Należy pamiętać, że jest to rozwijająca się dziedzina, a bycie na bieżąco z najnowszymi osiągnięciami i szkoleniami ma kluczowe znaczenie dla utrzymania jakości ocen.

Podsumowanie

* Systemy benchmarkingowe są niezbędne do weryfikacji wydajności GenAI i LLM, kierowania ulepszeniami i potwierdzania przydatności w warunkach rzeczywistych. Pomagają nam one w ocenie wydajności i gotowości generatywnej AI i LLM do wdrożenia w środowiskach produkcyjnych.

* Korelacja między ocenami a LLM to nowy i rozwijający się obszar. Powinniśmy wykorzystywać tradycyjne metryki, testy porównawcze specyficzne dla zadań LLM oraz ocenę ludzką, aby ocenić wydajność LLM i zapewnić jego przydatność do rzeczywistych zastosowań. G-Eval to metoda oceny bez odniesienia, wykorzystująca LLM do oceny spójności, spójności i trafności wygenerowanego tekstu.

* Konwencjonalne wskaźniki, takie jak BLEU, ROUGE i BERTScore, pomagają mierzyć jakość generowania tekstu i oceniać go numerycznie na podstawie dopasowania n-gramów lub podobieństwa semantycznego. Napotykają jednak pewne trudności w pełnym odzwierciedleniu znaczenia kontekstowego i parafrazowaniu.

* Testy porównawcze specyficzne dla studiów LLM mierzą, jak dobrze studia LLM wykonują zadania, takie jak klasyfikacja tekstu, analiza sentymentu i odpowiadanie na pytania. Wprowadzają nowe wskaźniki, takie jak ugruntowanie, spójność, płynność i podobieństwo GPT, które pomagają ocenić jakość wyników studiów LLM i ich zgodność ze standardami ludzkimi.

* Skuteczne metody ewaluacji rzetelnych ocen LLM obejmują testowanie w odpowiednich warunkach, tworzenie uczciwych podpowiedzi, przeprowadzanie przeglądów etycznych i ocenę doświadczenia użytkownika. Należą do nich zaawansowane testy porównawcze, takie jak HELM, HEIM, HellaSWAG i MMLU, które testują studia LLM w różnych scenariuszach i przy różnych możliwościach.

* Narzędzia takie jak Azure AI Studio i framework DeepEval umożliwiają skuteczną ewaluację studiów LLM w kontekście przedsiębiorstwa. Narzędzia te umożliwiają tworzenie niestandardowych przepływów pracy ewaluacyjnych, wykonywanie wsadowe i włączanie ocen w czasie rzeczywistym do ustawień produkcyjnych.