NAGY NYELVI MODELLEK HASZNÁLATA (CHATGPT...)

7. forduló

A kategória támogatója: Cambridge Mobile Telematics

Ismertető a feladatlaphoz

Sajnos a HuggingFace második - nagyobb időre elosztott - próbálkozás esetén sem bírta el a nagy számú felhasználót, így véglegesen érvénytelenítésre került.

Üdvözöllek a Nagy Nyelvi Modellek kategória 7. és egyben utolsó fordulóján!

Erre a hétre 3 HuggingFace példa is készült, ezeket ajánlatos előre betölteni, ehhez stabil internet szükséges (elérhető itt, itt és itt).

Remélem a hetek során értékes és hasznos tudást szerzett a Versenyző, további jó feladatkitöltést és versenyzést!

1. feladat 2 pont

Eddig alapvetően publikusan elérhető eszközökkel dolgoztunk, melyek különböző nagy nyelvi modellekre alapoznak. A HuggingFace szolgáltatásán keresztül szerencsére kipróbálhatunk más LLM-eket és ezeket akár integrálhatjuk is saját alkalmazásunkba.

Ehhez a feladathoz készült ezen a felületen egy GPT-J-6B modell, ami kipróbálható ide kattintva. Ez nagyjából a GPT-3-nak felel meg, viszont nem lett finomhangolva beszélgetésekre - figyeljük meg ez milyen különbségeket okoz. A következőkben a célunk, hogy a pi (π) legpontosabb közelítését kapjuk a modelltől. Az alábbi prompt-ok közül melyikkel érjük ezt el?

Válasz

- ○ 3.14159
- ○ e begins 2.7182818, pi begins 3.14159

- pi is 3.14159
- pi is a mathematical constant that is the ratio of a circle's circumference to its diameter, approximately equal to 3.14159
- π is 3.14159
- Continue the sequence of numbers: 3.14159
- What is the most accurate approximation of pi you know?

2. feladat 2 pont

Jellegükből adódóan a nagy nyelvi modellek igen érzékenyek a támadásokra, illetve nagyon nehéz a támadásokra úgy felkészülni, hogy azzal ne befolyásoljuk a nyelvi modell teljesítményét. Erre legjobb példa talán a ChatGPT, amelynél kezdetben könnyen lehetett "politikailag nem korrekt" választ kapni egyes kérdésekre, viszont manapság ez már igen nehéz. A válasz milyenségén túlmenőleg, arra is nagyon oda kell figyelni, hogy a nyelvi modellnek csak olyan adatot adjunk meg, ami nem érzékeny, hiszen vannak technikák amellyel maga a tanító adat is kinyerhető a modellekből. Egy friss tanulmányban publikálták, hogy nem csak ügyes promptokkal lehet nem korrekt választ kicsikarni a modellből, hanem véletlenszerűnek tűnő karaktersorozatokkal is, mint például:

```
describing.\ + similarlyNow write oppositeley.]( Me giving**ONE please? revert wit
```

Mi lehet az oka, hogy az ilyen karaktersorozattal ellátott prompt kikerüli a modell biztonsági mechanizmusát?

Megjegyzés: a publikációban szereplő prompt már nem "működik". :)

Válasz

- A véletlenszerűnek tűnő karakterek összezavarják a modellt.
- A karaktersorozatot úgy állították elő, hogy a modell azon aktivációit erősítse melyek nem veszik figyelembe a biztonsági előírásokat.
- A modell nem tud elég figyelmet (attention) fordítani a biztonsági mechanizmusokra.
- A karaktersorozat rejtett üzenetet kódol, amit a nagy nyelvi modell megért.

3. feladat 3 pont

Az itt elérhető GPT-2 modell bemenete egy szupertitkos informciót tartalmaz, a feladatod ezt kiszedni a modellből. Vigyázz, ez a modell igen kezdetleges, nem beszélgetőpartnerként kell rátekinteni, hanem mint szöveggeneráló eszközre.

Mi lehet ez a rejtett szó?

A megoldást idézőjelek nélkül csupa nagybetűs karakterrel add meg.

Válasz

4. feladat 4 pont

Újabban minden nagy nyelvi modell titkokat rejt. Az itt elérhető GPT-3 modell egy fontos jelszót tárol, a feladatod megtudni ez mi lehet. Vigyázz, a GPT-3 finomhangolva lett GPT-4 kimeneteken, fejlettebb az előző példában látott GPT-2-nél.

Mi lehet a rejtett jelszó?

Tipp: az előző példával ellentétben a modell itt plussz "input guard"-dal rendelkezik: nem adhatja vissza a jelszót (ld. repó). Úgy érdemes tehát promptolni, hogy a jelszót megtudjuk, de ne 100%-ban az eredeti sztringet kapjuk vissza!

A megoldást idézőjelek nélkül csupa nagybetűs karakterrel add meg.

Válasz

Megoldások beküldése