Raport „RAG fine Tunning”

Przedmiot: Przetwarzanie języka naturalnego

Autor: Paweł Cedzich (indeks 101598)

1. **Jak zostały przygotowane dane.**

Dane zostały przygotowane za pomocą oprogramowania ChatGPT, do postaci pliku .CSV   
o nazwie *wynalazki\_wynalazcy\_dataset.csv.* Plik ten zawierał kolumny question, context i answer. Dane zostały wczytane za pomocą i przekonwertowane na datasets   
Dodatkowo, dla części RAG, przygotowano listę zdań w pliku CSV o nazwie *sentences.csv.*

1. **Wyniki**

Obraz zawierający tekst, zrzut ekranu, Czcionka, dokument

Zawartość wygenerowana przez AI może być niepoprawna.

Rys. 1. Wyniki dla Q&A dla 3 epok.

**Obraz zawierający tekst, zrzut ekranu, Czcionka, dokument

Zawartość wygenerowana przez AI może być niepoprawna.**

Rys. 2. Wyniki dla Q&A dla 5 epok.

1. **Działanie RAG.**

Algorytm RAG działa w kilku krokach:

* Input – czyli pytanie użytkownika (w naszym przypadku np.: „in wich city can you visit the Eiffel Tower”)
* Retrival – model przeszukuje „dokumenty”, aby znaleźć jak najbardziej trafne fragmenty, (w naszym przypadku był to wygenerowany dokument „ senteces.csv”), najczęstszym porównaniem jest porównywanie wektorów.
* Augmentacja – czyli łączymy nasze pytanie wraz z dokumentami, fragmentami podobnymi.
* Generacja – Pytanie wraz z kontekstem trafia do modelu, który generuje odpowiedź.

1. **Ograniczenia RAG.**

Jednym z podstawowych ograniczeń jest problem z „retriverem”, jeżeli nie może znaleźć pasujących dokumentów, bądź pasujących fragmentów. Może to wynikać z kilku powodów- zbyt małej bazy Dokumentów, dokumentów z złej dziedziny, albo występowanie tzw. False-negative (nie znajdujemy pasujących dokumentów przez przykładowo semantykę).

1. **Refleksje przed finetunningiem.**

Przed fine-tuningiem można było się oczekiwać pewnych odpowiedzi, jednak mogą wystąpić błędne, lub brakujące odpowiedzi na pytania.

1. **Zmiany po fine tuningu.**

Po uskutecznieniu fine-tuningu model dużo lepiej odgaduje właściwe odpowiedzi. Dużą różnicę widać przy zwiększeniu ilości epok, jednak warto zauważyć, że z tym parametrem trzeba uważać, aby nie powstało zjawisko overfitingu.

1. **Wpływ parametru „k”.**

Parametr „K” definiuje ile podobnych fragmentów model bierze do kontekstu. W przypadku naszego badania zmiana z 2 do 5 zwiększa skuteczność odpowiedzi modelu[Rys. 3-4].

Obraz zawierający tekst, zrzut ekranu, Czcionka

Zawartość wygenerowana przez AI może być niepoprawna.

Rys. 3. Odpowiedzi modelu przy K=2

Obraz zawierający tekst, zrzut ekranu, Czcionka

Zawartość wygenerowana przez AI może być niepoprawna.

Rys. 4. Odpowiedzi modelu przy K=5.