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*.

2. Wyniki

```
--- Testowanie 5 przykładowych danych z zestawu treningowego ---

--- Test #1 ---
Pytanie z zestawu: Who came up with the idea of the internet?
Kontekst z zestawu: Tim Berners-Lee invented the World Wide Web in 1989.
Oryginalna odpowiedź z zestawu: Tim Berners-Lee
Wygenerowana odpowiedź modelu: Tim Berners-Lee
--- Koniec Testu #1 ---

--- Test #2 ---
Pytanie z zestawu: Who created the battery?
Kontekst z zestawu: Alessandro Volta invented the battery in 1800.
Oryginalna odpowiedź z zestawu: Alessandro Volta
Wygenerowana odpowiedź modelu:
--- Koniec Testu #2 ---

--- Test #3 ---
Pytanie z zestawu: Who invented the printing press?
Kontekst z zestawu: Johannes Gutenberg invented the printing press around 1440.
Oryginalna odpowiedź z zestawu: Johannes Gutenberg
Wygenerowana odpowiedź modelu:
--- Koniec Testu #3 ---

--- Test #4 ---
Pytanie z zestawu: Who created the vaccination?
Kontekst z zestawu: Edward Jenner developed the smallpox vaccine in 1796.
Oryginalna odpowiedź z zestawu: Edward Jenner
Wygenerowana odpowiedź modelu: Edward Jenner
--- Koniec Testu #4 ---

--- Test #5 ---
Pytanie z zestawu: Who came up with the idea of the microscope?
Kontekst z zestawu: Anton van Leeuwenhoek improved the microscope in the 17th century.
Oryginalna odpowiedź z zestawu: Anton van Leeuwenhoek
Wygenerowana odpowiedź modelu:
--- Koniec Testu #5 ---
```

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

```
--- Testowanie 5 przykładowych danych z zestawu treningowego ---

--- Test #1 ---
Pytanie z zestawu: Who came up with the idea of the internet?
Kontekst z zestawu: Tim Berners-Lee invented the World Wide Web in 1989.
Oryginalna odpowiedź z zestawu: Tim Berners-Lee
Wygenerowana odpowiedź modelu: Tim Berners-Lee
--- Koniec Testu #1 ---

--- Test #2 ---
Pytanie z zestawu: Who created the battery?
Kontekst z zestawu: Alessandro Volta invented the battery in 1800.
Oryginalna odpowiedź z zestawu: Alessandro Volta
Wygenerowana odpowiedź modelu: Alessandro Volta
--- Koniec Testu #2 ---

--- Test #3 ---
Pytanie z zestawu: Who invented the printing press?
Kontekst z zestawu: Johannes Gutenberg invented the printing press around 1440.
Oryginalna odpowiedź z zestawu: Johannes Gutenberg
Wygenerowana odpowiedź modelu: Johannes Gutenberg
--- Koniec Testu #3 ---

--- Test #4 ---
Pytanie z zestawu: Who created the vaccination?
Kontekst z zestawu: Edward Jenner developed the smallpox vaccine in 1796.
Oryginalna odpowiedź z zestawu: Edward Jenner
Wygenerowana odpowiedź modelu: Edward Jenner
--- Koniec Testu #4 ---

--- Test #5 ---
Pytanie z zestawu: Who came up with the idea of the microscope?
Kontekst z zestawu: Anton van Leeuwenhoek improved the microscope in the 17th century.
Oryginalna odpowiedź z zestawu: Anton van Leeuwenhoek
Wygenerowana odpowiedź modelu: Anton van Leeuwenhoek
--- Koniec Testu #5 ---
```

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

3. 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.

4. 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).

5. Refleksje przed finetunningiem.

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

6. 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.

7. 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].

```
Q: In which city can you visit the Eiffel Tower?
A: Paris

Q: Where is the Colosseum located?
A: Rome

Q: Which city is home to the Sydney Opera House?
A:

Q: What famous landmark can be found in New York City?
A: Statue of Liberty

Q: Which European city features the Brandenburg Gate?
A: Berlin
```

Rys. 3. Odpowiedzi modelu przy K=2

```
Q: In which city can you visit the Eiffel Tower?
A: Paris

Q: Where is the Colosseum located?
A: Rome

Q: Which city is home to the Sydney Opera House?
A: Bondi Beach

Q: What famous landmark can be found in New York City?
A: The Statue of Liberty

Q: Which European city features the Brandenburg Gate?
A: Berlin
```

Rys. 4. Odpowiedzi modelu przy K=5.