Projekt Retriever-Augmented Generation

1. Czy jest RAG

RAG (Retriever-Augmented Generation) jest systemem generatywnym mającym na celu generowanie odpowiedzi na podstawie kontekstu z bazy dokumentów dopasowanych do pytań zadanych przez użytkownika.

System ten można rozróżnić na :

1. bazę podzielonych dokumentów (kontekstów) zapisaną jako wektory,
2. retriever - model typu transformers którego zadaniem jest stworzenie grafu i osadzeń w przestrzeni wektorowej kontekstów jak i również przeszukiwanie bazy w celu znalezienia najbardziej odpowiedniego osadzenia,
3. generator, którego funkcją jest wygenerowanie treści na podstawie otrzymanego kontekstu z retriever'a. Zazwyczaj jest to duży model językowy.

Złożona struktura RAG'a ma zapewnić uniknięcia tak zwanych halucynacji modelu polegających generowaniu odpowiedzi niezwiązanych z pytaniami użytkownika albo całkowicie pozbawionych sensu. Struktura modelu zapewnia również elastyczność trenowania celu zwracania jak najdokładniejszych odpowiedzi przez część generującą nie polegając wyłącznie na "wiedzy" nabytej w trakcie procesu uczenia.

2. Realizacja Projektu

W trakcie realizacji projektu zdecydowałem się na użycie szkieletu programistycznego "llama index". Było to związane z łatwym tworzeniem struktury danych oraz względną bezkonfliktowością z paczkami już zainstalowanymi na system. System ostatecznie okazał się mieć mniejsz wsparcie społeczności oraz być bardziej hermetyczny niż początkowo zakładałem. Należy tutaj wspomnieć, że przy próbach optymalizacji retrivera nie współpracował on z biblioteką zewnętrzną "ONNX", a próby użycia jej z paczek llama index kończyły się niepowodzeniami instalacjami. Problemów tych pod koniec projektu upatrywałem w niekompletnym ustawieniu platformy CUDA na moim komputerze, która wspierała Pytorch ale nie Tensorflow. Spowodowało to ostatecznie konflikt sterowników (legacy and open kernel) na moim komputerze oraz przypadkowym wyczyszczeniem partycji z danymi, również tymi związanymi z projektem, przy próbie naprawy sterowników z bootloader'a na pendrive'ie. Bardzo interesującym aspektem była generacja pytań dopasowywanie opisu roli generatora do oczekiwanego efektu.

3. Dalsze rozwijanie projektu

W przyszłości projekt można rozwinąć o wykorzystanie następujących rozwiązań:

- proces fine tunning'u retrieval'a,
- odseparowanie tytułów i wykorzystanie ich jako jak metadata lub użycie ich w inny sposób