Projekat 2 Duboko učenje

Nemanja Vujić

January 2025

1 Uvod

Cilj ovog projekta jeste kreiranje sistema LLM (Velikih jezičnih modela) koji će kombinacijom tih (LLMova) rešiti problem prikupljanja, validacije i obrade informacija. Uz samo rešenje ovog projekta potrebna je i detaljna analiza. Problem ćemo u narednoj sekciji raščlaniti na manje probleme odnosno po jedan problem koji će jedan adekvatan LLM agent.

Definicije pojmova:
LLM Agent - je LLM model koji je optimizovan za resavanje specificnog problema uz pomoc spoljnih alatki. Optimizacija se vrsi na osnovu zadatog problema i moze biti ili finetuning postojeceg modela ili pomocu dobro napisane rutine u promptu.
Rutina - je skup koraka koji se izvrsavaju u odredjenom redosledu, i koji imaju odreenu svrhu. Skup koraka rutine, u nasem kontekstu, je skup napisanih koraka sa spiskom alatki koji se koriste da bi se doslo do odredjenog cilja. Denisanjem rutina u Agentu, mooemo jednostavno i precizno denisati sta agent treba da radi.
Alatka - je spoljna alatka koja se koristi u rutini, i koja ima odredjenu svrhu. Alatka se, uglavnom, definise kao funkcija koja se poziva sa odredjenim parametrima, i koja vraca odredjene rezultate. Unutar same funkcije, moze se koristiti bilo koja funkcionalnost koja je dostupna u programskom jeziku kojim se pise (npr. API). Pošto LLM-ovi rade sa tekstom, takve funkcije treba pretvoriti u tekstualne šeme koje će LLM razumeti.

2 Analiza problema

Cilj analize. Problem koji rešavamo u ovom projektu je kreiranje sistema baziranog na LLMovima koji omogućavaju asistenciju u istraživanju. Da bismo to postigli, problem je potrebno podeliti na manje podprobleme gde će svaki od njih biti rešen pomoću specifično definisanog agenta.

2.1 Glavni zadaci sistema:

1. Pretraživanje informacija
2. Generisanje teksta od informacija
3. Validacija
4. Interaktivna komunikacija sa korisnikom.

Sistem će biti napravljen tako po jedan agent rešava jedan od podproblema i ciljea projekta. Uz jasno definisane I/O streamove i primopredaju agenata.

2.2 Podproblemi i podela agenta

2.2.1 Agent za pretrazivanje informacija

Agent je odgovoran za pretrazivanje informacija uz pomoc alatki. Alatke koje koristi su Wikipedia Search i BeautifulSoup scraping alatke. Ne koristi ih direktno kao "tool" vec pozivaju funkcije koje izvrsavaju zadatke. Agent od liste svih linkova sa Wiki stranice za odredjeni topic bira 5 koji su najbitniji za isti.
Ulaz: Tema kao "keyword" i lista od 100 linkova.
Izlaz: Lista izabranih linkova koji se pretvaraju u dokumente.
Model: Za ovo je koriscen LLama 3.2-1b zbog sposobnosti izbora a adekvatne lakoce za implementaciju.

2.2.2 Agent za generisanje sazetaka

Agent sluzi za generisanje jednog sazetog teksta od 6 dokumenata koje je prethodni izabrao. Ovaj agent preuzima informacije iz agenta za pretrazivanje i generise sazeti tekst na osnovu njih.
Ulaz: Dokumenti.
Izlaz: Tekstualni sazetak.
Model: Koriscen Llama 3.2-1b zbog decision making sposobnosti.

2.2.3 Agent za validaciju

Ovaj agent proverava tacnost informacija i poredi ih sa spoljnim izvorima. Ovaj agent koristi direktan tool za internet search koji poziva API. Uz pomoc toga proverava informacije generisane od prethodnih agenata za savremenim resursima.

Ulaz: Sazeti generisani tekst vezan za temu.
Izlaz: lista dokumenata, cinjenica i podataka vezanih za temu.
Model: Koriscen je model 3.3-70b kao jedan od najvecih i kompleksnijih modela zato sto ovaj zadatak zahteva up to date podatke. Model 3.3-70b je najnoviji i izasao je u decembru 2024 godine.

2.2.4 Agent za interaktivnu komunikaciju

Alat je odgovoran za pretrazivanje informacija uz pomoc alatki. Alatke koje koristi su Azure Cognitive Search i slicne scraping alatke.
Ulaz: kljucne reci i fraze vezane za temu.
Izlaz: lista dokumenata, cinjenica i podataka vezanih za temu.

3 Dijagram rada sistema

Sistem funkcionise tako sto od korisnika prvo uzima kljucnu rec ili vise kljucnih reci koje cine zadati topic. To se prosledjuje prvom agentu koji uz pomoc tzv. toolova pronalazi adekvatne informacije na internetu i ingestuje ih. Sa obzirom na to da wikipedija po stranici ima dosta linkova model bira 5 najrelevantnijih linkova koje prosledjuje scraping toolu za izvlacenje cinjenica i kreiranje dokumenata.
Nakon toga posao preuzima agent za sumerizaciju koji od tih 6 dokumenata pravi jedan konkretan dokument sa adekvatnim podacima.
Ovaj dokument preuzima agent za proveru cinjenica koji proverava da li se cinjenice poklapaju sa trenutnim cinjenicama na internetu i dodaje ili menja po potrebi dokument.
Nakon svega toga ovaj dokument se ubacuje u chat model kao context koji se koristi za QnA sa korisnikom.

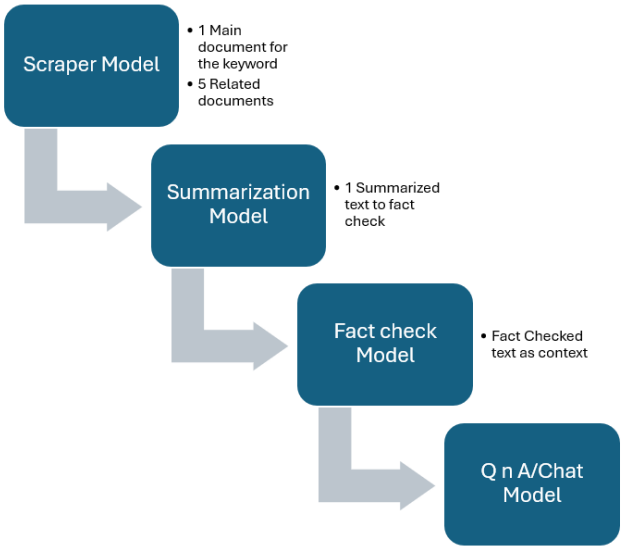

Figure 1: Dijagram modela

4 Tok rada

Prvi model pocinje sa radom:
Enter the desired research topic:
Powerlifting
Best match for 'Powerlifting': Powerlifting - Pronadjen uz tool koji koristi wikipedia API
Nakon cega wikipedia API uzima sve linkove sa stranice za powerlifting i samu tu stranicu i prosledjuje ih nazad modelu da on odluci koje linkove ce da koristi.
Model uzvraca odgovor:
Generated Topics: Barbell, Weightliftg, International_Powerlifting_Federation, Drug_test, Boxing - Ovi topici se sada pretrazuju i pretvaraju u dokumente sa wikipedie.
Drugi model preuzima dokumente:
Drugi model kao rezultat izbacuje sledece:
{*'role': 'assistant', 'content': 'The text discusses the sport of powerlifting, including its history, rules, and techniques. Here are the main points summarized: n n**History** n n* Powerlifting originated in ancient Greece and was later developed in the United Kingdom and the United States. n* The sport was initially... ...*}
Treci model proverava drugi:
Fact checking in progress...
Fact checking completed
Cetvrti model ingestuje tekst kao kontekst i komunicira sa korisnikom:
Start QnA. Type '/exit' to end.
You: What are the 3 excersises in powerlifting
Assistant: The 3 excersises in powerlifting are squat, bench press, and deadlift.