UNIVERZITET U NIŠU

ELEKTRONSKI FAKULTET

Katedra za računarstvo

**RETRIEVAL-AUGMENTED GENERATION (RAG) TEHNIKA ZA UNAPREĐENJE CHATBOT APLIKACIJE**

Diplomski rad

Studijski program: Elektrotehnika i računarstvo

Modul: Računarstvo i informatika

Kandidat:

Božidar Mitić, br. indeksa: 18282

Mentor:

Prof. dr Aleksandar Stanimirović

Niš, Septembar 2024. godina

Univerzitet u Nišu

Elektronski Fakultet

**RETRIEVAL-AUGMENTED GENERATION (RAG) TEHNIKA ZA UNAPREĐENJE CHATBOT APLIKACIJE**

**RETRIEVAL-AUGMENTED GENERATION (RAG) TECHNIQUE FOR ENHANCING CHATBOT APPLICATIONS**

Završni / Master / Diplomski rad

Studijski program: <naziv>

Modul: <naziv>

Student: <Ime Prezime>, br. ind. <broj>

Mentor: <Prof. / Doc dr Ime Prezime>

Zadatak: <*Ovde ide tekst za zadatak ispisan italic stilom.>*

Datum prijave rada: xx.xx.xxxx

Datum predaje rada: xx.xx.xxxx

Datum odbrane rada: xx.xx.xxxx

Komisija za ocenu i odbranu:

1. <Prof. / Doc dr Ime Prezime>, Predsednik Komisije

1. <Prof. / Doc dr Ime Prezime>, Član

1. <Prof. / Doc dr Ime Prezime>, Član

Sadržaj

[1 Uvod 4](#_Toc176985145)

[2 Generativni AI(GenAI) i Veliki Jezički Modeli(LLMs) 5](#_Toc176985146)

[2.1 Upoznavanje 5](#_Toc176985147)

[2.1.1 GenAI 5](#_Toc176985148)

[2.1.2 LLMs 5](#_Toc176985149)

[2.2 Primena 7](#_Toc176985150)

[3 Retrival-augmented generation (RAG) 8](#_Toc176985151)

[3.1 Uvod 8](#_Toc176985152)

[3.1.1 Poreklo 8](#_Toc176985153)

[3.2 Komponente 9](#_Toc176985154)

[3.2.1 Unapred obučen model (Pre-trained model) 9](#_Toc176985155)

[3.2.2 Skladištenje podataka (Data Storage) 9](#_Toc176985156)

[3.2.3 Mehanizam za pronalaženje (Retrieval mechanism) 9](#_Toc176985157)

[3.2.4 Popularne primene RAG-a u oblasti veštačke inteligencije 11](#_Toc176985158)

[3.2.5 Prednosti i nedostaci 12](#_Toc176985159)

[3.3 Osnovni RAG sistem 13](#_Toc176985160)

[3.3.1 Indeksiranje i embedding (Indexing and Embedding) 13](#_Toc176985161)

[3.3.2 Pretraga (Retrieval) 14](#_Toc176985162)

[3.3.3 Generisanje (Generation) 15](#_Toc176985163)

[3.4 Napredni RAG Sistem 16](#_Toc176985164)

[3.4.1 Prevođenje upita (Query Translation) 16](#_Toc176985165)

[3.4.2 Rutiranje (Routing) 21](#_Toc176985166)

[3.4.3 Kreiranje upita (Query Construction) 22](#_Toc176985167)

[3.4.4 Indeksiranje (Indexing) 23](#_Toc176985168)

[3.4.5 CRAG 26](#_Toc176985169)

[3.4.6 RAG vs Fine-tuning (Fino podešavanje) 28](#_Toc176985170)

[3.4.7 Budućnost RAG-a 29](#_Toc176985171)

# Uvod

Sa brzim razvojem tehnologije, količina informacija kojima smo svakodnevno izloženi postaje sve veća i kompleksnija. Ovaj eksponencijalni rast informacija predstavlja značajan izazov za obradu i ekstrakciju korisnog sadržaja iz velikih skupova podataka. Tradicionalne metode pretrage i obrade često nisu dovoljne da pruže relevantne i precizne informacije u realnom vremenu, što otvara prostor za primenu naprednijih tehnika kao što je Retrieval-Augmented Generation (RAG).

RAG predstavlja kombinaciju tehnika tradicionalnih sistema za pretragu informacija (baze podataka) sa mogućnostima generativnih velikih jezičkih modela (LLMs). Zbog svoje fleksibilnosti i moći da obrađuje raznovrsne informacije, RAG se primenjuje u mnogim oblastima, uključujući, ali ne ograničavajući se na, chatbot aplikacije, automatizovanu analizu dokumenata, generisanje odgovora na složena pitanja, pa čak i personalizovane preporuke.

Cilj ovog rada je istražiti i prikazati potencijal RAG tehnike, s posebnim osvrtom na njenu primenu u unapređenju funkcionalnosti chatbot aplikacija. Posebna pažnja biće posvećena kako teoretskom, tako i praktičnom aspektu integracije RAG sistema, uključujući analizu njegovog uticaja na efikasnost, tačnost, i skalabilnost chatbot aplikacija. Pored toga, istražiće se i mogućnosti daljeg razvoja i prilagođavanja ove tehnike za specifične industrijske primene.

Drugo poglavlje obuvata upoznavanje sa Generative AI-em i velikim jezičkim modelima koji spadaju u najvažnije komponente svakog RAG sistema, kao njihovim primenama.

U trećem poglavlju rada, akcenat će biti stavljen na komponente RAG sistema, uključujući unapred obučene modele, skladištenje podataka, i mehanizme za pronalaženje relevantnih informacija. Posebno će biti obrađene popularne primene RAG tehnike u veštačkoj inteligenciji, s naglaskom na prednosti koje ona donosi u poređenju sa tradicionalnim pristupima. Takođe će biti analizirani i nedostaci.

Četvrto poglavlje biće posvećeno implementaciji praktičnog dela, gde će biti detaljno objašnjena integracija RAG sistema u chatbot aplikaciju. Prikazaće se koraci u indeksiranju podataka, generisanju odgovora, i načinima optimizacije performansi sistema za različite primene. Cilj ovog dela je da prikaže kako se RAG može koristiti za poboljšanje funkcionalnosti chatbot sistema kroz konkretne primere.

# Generativni AI(GenAI) i Veliki Jezički Modeli(LLMs)

## Upoznavanje

Da bismo razumeli način na koji funkcionise RAG najpre moramo u najmanju ruku razumeti šta je to generative AI (GenAI) i veliki jezički modeli (LLMs), osnovu njihove kompleksne arhitekture i njihovu ulogu u modernoj tehnologiji.

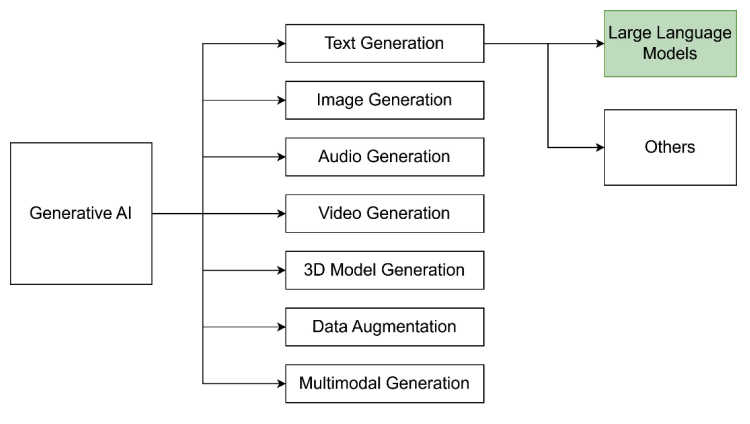
### GenAI

GenAI predstavlja sistem koji ima mogućnost generisanja novog sadržaja kao što su tekst, slike, audio i video. Za razliku od specijalizovanih AI sistema koji su dizajnirani za specifične zadatke poput klasifikacije slika ili prepoznavanja govora, generativni AI modeli mogu kreirati potpuno nove sadržaje koje je često vrlo teško, a možda i nemoguće razlikovati od sadržaja koji je stvorio čovek.

Ovi sistemi koriste tehnike mašinskog učenja kao što su neuronske mreže koje se obučavaju na ogromnim količinama podataka. Generativni modeli mogu modelovati verovatnoću distribucije podataka i semplovati tu distribuciju za generisanje novih primera. To rade tako što uče obrasce i strukturu u okviru podataka za obuku.

### LLMs

Jedna od najbrže rastućih grana u oblasti GenAI-a je generisanje prirodnih jezika pomoću velikih jezičkih modela. Ovde spadaju tehnike kao što su Natural Language Processing (NLP) i Natural Language Understanding (NLU)



Slika 1. Podela GenAI-a

Veliki jezički modeli predstavljaju deo generativne veštačke inteligencije (GenAI), koji su specifično dizajnirani za razumevanje i generisanje ljudskog jezika. Oni funkcionišu kao neuronske mreže obučene na ogromnim količinama tekstualnih podataka (članci, knjige, transkripti video zapisa, itd.), što je razlog za reč "veliki" u njihovom imenu. Za razliku od sistema za generisanje teksta zasnovanih na pravilima, ono što ove modele čini posebnim je njihova sposobnost da stvore potpuno nov, originalan tekst koji deluje prirodno.

Obučavanjem velikih jezičkih modela na velikim količinama podataka nekog jezika, ovi modeli stiču duboko razumevanje strukture i karakteristika određenog jezika, kao što su sintaksa i kompleksna gramatika, razumevanje idioma, semantika, kulturne reference i slično. Nakon toga, oni su u stanju da vrše rezimiranje tekstova, a čak i generisanje potpuno novih tekstova pretpostavljajući sledeću reč u rečenici na osnovu sadržaja sa kojim su se upoznali.

U jednom trenutku uvodi se pojam **obraćanje pažnje.** Ideja je inspirisana ljudskom sposobnošću selektivnog fokusiranja na određene delove informacija, dok se ostale zanemaruju. Cilj je da se na neki način ova jako specifična ljudska osobina izimitira i integriše u domenu velikih jezičkih modela, omogućavajući im mnogo viši nivo razumevanja i generisanja jezika. Na ovaj način su nastali **transformeri**, koji se sastoje od kodera i dekodera i korsite koncept obraćanja pažnje i na taj način uče zavisnostni na velikim udaljenostima u tekstu, a ne samo zavisnosti reči koje su blizu jedna do druge. Zbog toga modeli transformera imaju sposobnost razumevanja, ne samo pojedinačnih reči, već i celokupnog konteksta u kojem se te reči nalaze. Transformer model obrađuje podatke tako što prvo tokenizuje ulazne podatke, a zatim istovremeno izvodi matematičke operacije kako bi otkrio odnose između tokena. Ovo omogućava računaru prepoznavanje obrasca koje bi i čovek uočio da mu je postavljeno isto pitanje.

A diagram of a transformer

Description automatically generated

Slika 2. Arhitektura transformera

## Primena

Veštačka inteligencija postaje sve prisutnija u ljudskim životima, donoseći sa sobom promene u različitim aspektima društva i industrije. Kako se njene mogućnosti šire, generativna veštačka inteligencija i veliki jezički modeli pronalaze primenu u mnogim oblastima. Ovi modeli omogućavaju automatizaciju složenih zadataka, stvaranje novog sadržaja, kao i bolje razumevanje i interpretaciju ljudskog jezika. Štaviše, oni doprinose pristupačnosti pomažući osobama sa invaliditetom, putem aplikacija za pretvaranje teksta u govor i generisanje sadržaja u pristupačnim formatima. Od zdravstva do finansija, veliki jezički modeli transformišu industrije ubrzavajući procese, poboljšavajući korisničko iskustvo i omogućavajući efikasnije i na podacima zasnovano donošenje odluka.

Primene velikih jezičkih modela inteligencije:

* **Generisanje teksta**: Sposobnost generisanja jezika, poput pisanja e-mailova, blogova ili drugog sadržaja kao odgovor na zadate upite, koji se zatim mogu dodatno usavršavati. Odličan primer je retrieval-augmented generation (RAG), o čemu će biti reči kasnije.
* **Sažimanje sadržaja**: Sažimanje dugih članaka, priča u novinama, istraživačkih izveštaja, korporativne dokumentacije, u sažete tekstove prilagođene dužini izlaznog formata.
* **AI asistenti**: Chatbot-ovi koji odgovaraju na korisnička pitanja, obavljaju zadatke u pozadini i pružaju detaljne informacije na prirodnom jeziku kao deo integrisanog rešenja za korisničku podršku.
* **Generisanje koda**: Pomaže programerima u izradi aplikacija, pronalaženju grešaka u kodu i otkrivanju bezbednosnih problema u više programskih jezika, pa čak i „prevođenju“ između njih.
* **Analiza osećanja**: Analiza teksta kako bi se utvrdio ton korisnika u cilju razumevanja povratnih informacija.
* **Prevođenje jezika**: Pruža širu pokrivenost organizacijama širom sveta sa tečnim prevodima i višejezičnim mogućnostima.

Uskoro će veliki jezički modeli značajno uticati na razne industrije, od finansija i osiguranja, preko ljudskih resursa, pa sve do zdravstva i šire, automatizujući korisničku samouslugu, ubrzavajući vreme odgovora na sve veći broj zadataka, kao i pružajući veću preciznost, poboljšano usmeravanje i inteligentno prikupljanje konteksta.

# Retrival-augmented generation (RAG)

## Uvod

Kada govorimo o problemima sa velikim jezičkim modelima, važno je istaći njihove nedostatke, kao što su nemogućnost dinamičkog ažuriranja ili pristupanja spoljnim informacijama, rizik od generisanja netačnih informacija ili "halucinacija" i nedostatak transparentnosti u načinu na koji dolaze do određenih zaključaka. Ovi modeli, iako moćni, inherentno su statiči, oslanjaju se isključivo na podatke na kojima su obučeni, a oni mogu postati zastareli ili nedovoljno relevantni za određene zadatke. Obučavanje velikih jezičkih modela je veoma zahtevno, kako u pogledu vremena, tako i resursa, što onemogućava njihovo konstantno ažuriranje. U međuvremenu, kako tehnologija napreduje, količina informacija na internetu eksponencijalno raste, što dodatno otežava zadatak održavanja modela aktuelnim. Ovi izazovi su podstakli potragu za inovativnim rešenjem koje može prevazići ova ograničenja.

Tu na scenu stupa Retrieval-Augmented Generation (RAG). RAG rešava ove probleme kombinovanjem generativnih sposobnosti LLM-ova sa spoljnim sistemom za pretragu koji može pristupiti i uključiti najrelevantnije, najnovije informacije iz velike baze podataka ili izvora znanja. Na taj način, RAG poboljšava tačnost, relevantnost i činjeničnost generisanog teksta, što ga čini posebno korisnim za zadatke koji zahtevaju specifične, aktuelne ili detaljne informacije. Ovaj hibridni pristup ne samo da poboljšava kvalitet generisanog sadržaja, već omogućava i transparentnije i prilagodljivije AI sisteme.

A diagram of a process

Description automatically generated

Slika 3. Osnovni RAG pristup

### Poreklo

RAG se prvi put javlja u radu "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," koji su predstavili Patrick Lewis i njegovi saradnici 2020. godine. Ovaj rad predstavlja prekretnicu u razvoju tehnika za kombinovanje pretrage i generisanja teksta. U njemu autori uvode RAG kao pristup koji kombinuje moć velikih jezičkih modela sa informacijama iz spoljnjih izvora, poput Vikipedije, kako bi se poboljšala tačnost i relevantnost generisanih odgovora u zadacima koji zahtevaju obimno znanje.

## Komponente

RAG sadrži nekoliko glavnih komponenti: unapred obučen model (pre-trained model), mehanizam za pronalaženje (retrieval mechanism) i skladištenje podataka (data storage). U pozadini, arhitektura je nešto kompleksnija, ali je neophodno prvo razumeti funkciju glavnih komponenti pre nego što se pređe na detaljnije procese i dodatne elemente.

### Unapred obučen model (Pre-trained model)

Unapred obučen model predstavlja osnovni deo RAG arhitekture. Ovaj model je sekvencijalno-sekvencijalni model (seq2seq) treniran na ogromnim količinama tekstualnih podataka. Njegova glavna funkcija je generisanje teksta na osnovu ulaznih podataka i informacija koje dobija iz mehanizma za pronalaženje. Ovaj model koristi naučene parametre za kreiranje smislenog i kontekstualno relevantnog izlaza. U kontekstu RAG-a, unapred obučeni model kombinuje informacije iz više izvora, generišući odgovore koji su precizniji i bogatiji informacijama nego kada bi se oslanjao samo na svoje interne parametre.

### Skladištenje podataka (Data Storage)

Skladištenje podataka predstavlja ključnu komponentu koja omogućava čuvanje i brz pristup velikim količinama informacija koje RAG model koristi za generisanje odgovora. Baze podataka za dokumenta (Document Databases) služe kao veliki i pristupačni sistemi za skladištenje eksternih skupova podataka ili dokumenata iz kojih RAG model preuzima informacije. Ove baze podataka moraju biti dovoljno velike i dobro organizovane da bi omogućile efikasan pristup potrebnim informacijama.

Skalabilna rešenja za skladištenje su takođe od suštinskog značaja, posebno kada se radi o ogromnim skupovima podataka. Cloud storage i distribuirani fajl sistemi često se koriste za ovu svrhu, omogućavajući sistemu lako skaliranje u skladu sa rastućim zahtevima za skladištenjem i obradom podataka. Pored toga, ovakva rešenja omogućavaju visok nivo dostupnosti i otpornosti na kvarove, čime se osigurava kontinuitet u radu modela.

### Mehanizam za pronalaženje (Retrieval mechanism)

Mehanizam za pronalaženje je komponenta koja se koristi za pretragu i dohvat informacija iz raznih vrsta baza podataka, bilo da je reč o vektorskim, grafičkim, SQL ili dokumentnim bazama. Kada se postavi upit, ovaj mehanizam ga obrađuje i koristi odgovarajuću tehniku pretrage, bilo da je to pretraga po ključnim rečima, relacijama među podacima, ili vektorska pretraga u slučaju numeričkih reprezentacija. Odabrani rezultati predstavljaju delove podataka ili dokumente koji su najrelevantniji za postavljeni upit. Mehanizam za pronalaženje omogućava pronalazak najaktuelnijih i najpreciznijih informacija, koje se zatim kombinuju sa znanjem unapred obučenog modela, čime se poboljšava celokupni izlaz generisan od strane RAG-a.

#### Klasične tehnike pronalaženja informacija

* **Frekvencija termina-inverzna frekvencija dokumenta (Term Frequency-Inverse Document Frequency (TF-IDF)):** Ova tehnika se široko koristi za procenu relevantnosti dokumenta u odnosu na postavljeni upit. TF-IDF izračunava značaj svakog termina u dokumentu na osnovu njegove učestalosti u samom dokumentu i retkosti kroz sve dokumente. Termin sa najvišim TF-IDF skorom smatra se najrelevantnijim.
* **Torba reči (Bag of Words - BoW):** Ova tehnika predstavlja dokument kao torbu, odnosno skup pojedinačnih reči. Za svaku reč se pamti broj ponavljanja i nije bitan njihov redosled. BoW je korisna za obradu upita sa greškama u pisanju i smanjenje uticaja reči poput "i", "ali", "the", itd.
* **Dodeljivanje težina terminima (Term weighting):** Ova tehnika dodeljuje težine svakom terminu u dokumentu na osnovu njegove relevantnosti za upit. Težine se koriste za izračunavanje ukupnog skora dokumenta, koji se zatim poredi sa pragom relevantnosti kako bi se odredila njegova važnost.

#### Savremene tehnike pronalaženja informacija

* **BM25:** Ova tehnika je proširenje TF-IDF metode koja uzima u obzir dužinu dokumenta i broj termina koje on sadrži. BM25 dodeljuje više skorove dokumentima sa većom dužinom i više termina, jer se smatra da su oni informativniji.
* **Pristupi zasnovani na neuronskim mrežama:** Ove tehnike koriste neuronske mreže za učenje složenih obrazaca u matrici upit-dokument. Neuronske mreže mogu naučiti da izvlače relevantne karakteristike iz vektora dokumenata i upita, poput dužine rečenica, oznaka vrsta reči i imenovanih entiteta.
* **Modeli zasnovani na dubokom učenju:** Ovi modeli koriste tehnike dubokog učenja, poput konvolucijskih neuronskih mreža (CNN) i rekurentnih neuronskih mreža (RNN), za učenje složenih obrazaca u matrici upit-dokument. Modeli dubokog učenja mogu naučiti da izvlače relevantne karakteristike iz vektora dokumenata i upita, kao što su vektorski prikaz reči (word embeddings), vektorski prikaz rečenica (sentence embeddings) i latentne semantičke strukture.

### Popularne primene RAG-a u oblasti veštačke inteligencije

Pre nego što se detaljno objasni kako funkcioniše RAG sistem, biće reči o njegovim popularnim primenama u oblasti veštačke inteligencije. Sa sve širim razvojem veštačke inteligencije, naročito velikih jezičkih modela, u različitim industrijama i oblastima, RAG se pojavio kao ključna inovacija koja dodatno unapređuje mnoge od tih primena. U nastavku će biti predstavljene neke od tih najpopularnijih oblasti u kojima je RAG doneo značajne koristi [29].

1. **Kreativno pisanje**: RAG može biti korišćen za generisanje kreativnog sadržaja poput priča, pesama i članaka. Kroz analizu postojećih tekstova, jezički model uči obrasce i strukture jezika, omogućavajući mu da kreira originalan i zanimljiv sadržaj, često neprimetno različit od onog koji pišu ljudi.
2. **Dizajn proizvoda**: RAG se može primeniti u dizajnu inovativnih proizvoda generisanjem novih ideja na osnovu postojećih dizajna proizvoda ili stvaranjem potpuno novih koncepata. Na primer, jezički model može analizirati uspešne dizajne i generisati nove sa unapređenom funkcionalnošću ili estetikom.
3. **Medicinska dijagnoza**: RAG može pomagati lekarima u postavljanju medicinskih dijagnoza generisanjem mogućih dijagnoza na osnovu simptoma i medicinske istorije pacijenata. Jezički model analizira velike količine medicinskih podataka i identifikuje obrasce i odnose koje ljudski lekar može prevideti, čime doprinosi tačnijim dijagnozama i tretmanima.
4. **Prevođenje jezika**: RAG može biti korišćen za generisanje prevoda tekstova, govora ili čak celih knjiga. Kroz analizu postojećih prevoda, jezički model uči obrasce i nijanse jezika, omogućavajući mu da proizvodi tačne i prirodne prevode.
5. **Virtuelni asistenti**: RAG se može koristiti za kreiranje virtuelnih asistenata koji mogu komunicirati s korisnicima, odgovarati na pitanja ili čak izvršavati zadatke. Jezički model analizira velike količine podataka i uči kako da odgovara korisnicima na način sličan ljudima, omogućavajući mu da pruži personalizovana iskustva i poboljša interakcije sa korisnicima.
6. **Razvoj igara**: RAG se može primeniti na generisanje nivoa u igrama, likova ili čak celih igara. Jezički model analizira postojeće igre i uči obrasce i strukture dizajna igara, omogućavajući mu da kreira nove i inovativne igre koje često ne zaostaju za onima koje kreiraju ljudi.
7. **Arhitektonski dizajn**: RAG se može koristiti za generisanje arhitektonskih dizajna na osnovu postojećih zgrada ili čitavih gradova. Jezički model analizira velike količine arhitektonskih podataka i uči obrasce i odnose između različitih dizajnerskih elemenata, omogućavajući mu da kreira inovativne i funkcionalne dizajne.
8. **Dizajn mode**: RAG se može primeniti na generisanje modnih dizajna na osnovu postojećih stilova ili čak čitavih kolekcija. Jezički model analizira velike količine modnih podataka i uči obrasce i odnose između različitih dizajnerskih elemenata, omogućavajući mu da kreira inovativne i stilizovane dizajne.
9. **Edukativni sadržaj**: RAG može biti korišćen za generisanje edukativnog sadržaja kao što su planovi lekcija, kvizovi ili čak celi kursevi. Jezički model analizira postojeće obrazovne materijale i uči obrasce i strukture učenja, omogućavajući mu da kreira visokokvalitetan edukativni sadržaj prilagođen potrebama pojedinih učenika.

### Prednosti i nedostaci

RAG prevazilazi ograničenja tradicionalnih jezičkih modela koji se oslanjaju isključivo na podatke na kojima su obučeni i time omogućava korišćenje aktuelnih informacija i šireg spektra izvora podataka, što donosi niz prednosti [30]:

1. **Osiguravanje relevantnosti uz ažurirane informacije**: RAG omogućava velikim jezičkim modelima pristup najnovijim dostupnim informacijama, što je od ključne važnosti za teme koje zavise od vremenski osetljivih podataka. Ovo osigurava da generisani sadržaj ostane relevantan i odražava najnovije događaje i podatke.
2. **Povećanje preciznosti odgovora kroz poboljšanu tačnost**: Integrisanjem spoljnih izvora informacija, RAG značajno poboljšava tačnost i nivo detalja u odgovorima generisanim od strane velikih jezičkih modela. Ovo poboljšanje je naročito uočljivo kod odgovora na činjenična pitanja, gde su preciznost i tačnost od najveće važnosti.
3. **Smanjenje pristrasnosti kroz eksterno preuzimanje podataka**: Jedna od inovativnih primena RAG-a je njegova sposobnost da se suoči sa pristrasnostima koje mogu postojati u obučavajućim podacima jezičkih modela. Preuzimanjem i integrisanjem informacija iz različitih eksternih izvora, RAG uvodi širu perspektivu koja pomaže u smanjenju pristrasnosti u rezultatima modela.
4. **Poboljšano kontekstualno razumevanje**: RAG-ova komponenta za pretragu ističe se u prepoznavanju i pristupanju informacijama specifičnim za korisnički upit. Ova sposobnost omogućava generisanje odgovora koji su ne samo tačni, već i kontekstualno prilagođeni potrebama korisnika, pružajući personalizovanije i relevantnije iskustvo.
5. **Prevazilaženje ograničenja pretreniranja modela**: Tradicionalni jezički modeli često se suočavaju sa izazovom ograničenog znanja zasnovanog samo na podacima na kojima su obučeni. RAG rešava ovaj problem pristupanjem i integrisanjem dodatnih eksternih informacija, proširujući bazu znanja modela i njihovu primenljivost.

RAG donosi brojne prednosti, ali ima i svoje nedostatke. Jedan od glavnih problema je povećana latencija, jer sistem mora da pretraži spoljne izvore informacija pre nego što generiše odgovor, što može usporiti proces. Pored toga, RAG zavisi od kvaliteta i ažurnosti tih eksternih izvora, pa ako su podaci zastareli ili neadekvatni, to može negativno uticati na tačnost rezultata. Takođe, implementacija RAG sistema zahteva dodatne resurse, jer je potrebno održavati bazu podataka i razviti efikasne strategije pretrage, što sve može povećati kompleksnost u odnosu na tradicionalne modele koji ne koriste pretragu.

## Osnovni RAG sistem

A diagram of a diagram of a document

Description automatically generated

Slika 5. Osnovni RAG sistem

### Indeksiranje i embedding (Indexing and Embedding)

Zamišlja se osoba u ogromnoj biblioteci sa milionima knjiga raspoređenih po nebrojenim policama. Njen zadatak je da pronađe informacije o vrlo specifičnoj temi. Kako bi to uradila? Mogla bi lutati kroz hodnike, knjigu po knjigu, ali to bi bilo neefikasno i oduzelo bi mnogo vremena. Ovde na scenu nastupa indeksiranje. Indeksiranje u veštačkoj inteligenciji omogućava organizaciju i brzo preuzimanje informacija iz velikih skupova digitalnih podataka i koristi se kada se radi sa vektorskim bazama podataka. Proces započinje učitavanjem podataka iz različitih izvora, što omogućava kreiranje pretražive baze podataka.

A diagram of a diagram

Description automatically generated

Slika 6. Indeksiranje i embedding

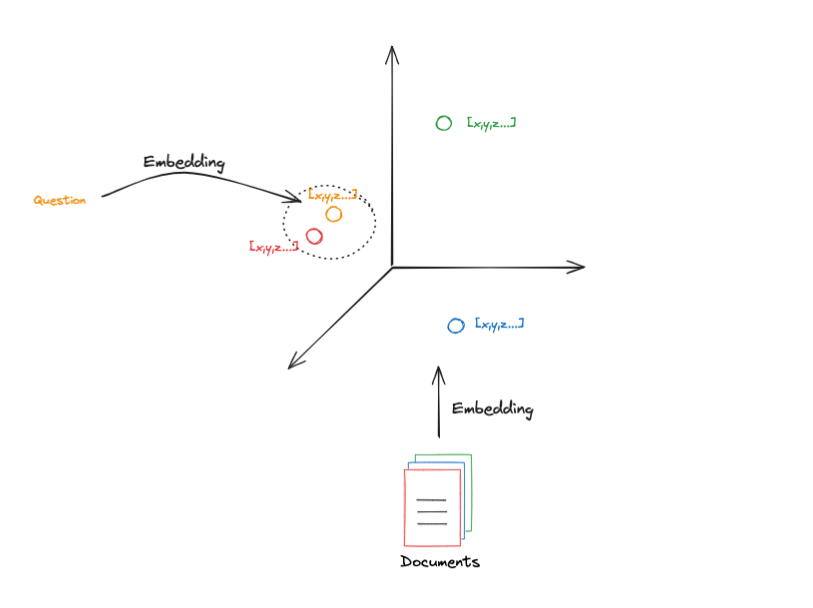
Embedding predstavlja prevođenje ovih dokumenata u format koji omogućava računaru brzu pretragu. Ovo najčešće obavljaju veoma kompleksni embedding modeli. Dokumenti se transformišu u numeričke reprezentacije, odnosno vektore. Rane metode, poput Bag of Words, koristile su retke vektore zasnovane na čuvanju broja ponavljanja reči u dokumentu, dok moderni modeli mašinskog učenja stvaraju guste vektore koji hvataju suštinu teksta. Indeksiranje uključuje deljenje dokumenata na manje delove, njihovu konverziju u vektore(embedding), a zatim organizaciju tih dokumenata na način koji omogućava brz pristup.

### Pretraga (Retrieval)

Kada se govori o pretrazi, misli se na proces upotrebe mehanizma za pretragu u cilju pronalaska odgovarajućih dokumenata ili delova dokumenata koji odgovaraju korisničkom upitu. Kao što je već rečeno, pretragu je moguće obaviti nad raznim vrstama baza podataka, ali ovde će fokus biti na najčešći slučaj kada je u pitanju RAG, a to su vektorske baze podataka.

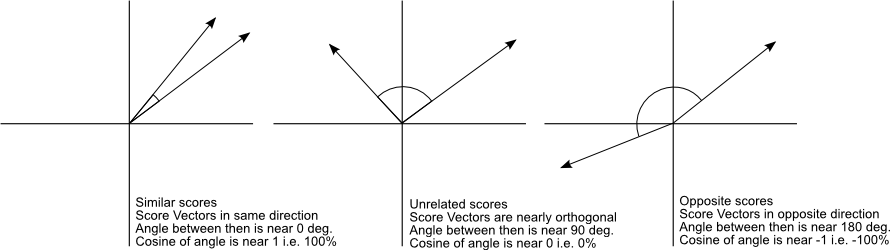
Kod vektorskih baza podataka, kao što se dokumenti prevode u vektore, isto tako je neophodno i korisnički **upit(query)** prevesti u vektor da bi preraga bila moguća. Vektori se upoređuju pomoću procesa koji se zove **provera sličnosti (similarity check).** Provera sličnosti omogućava pribavljanje sličnih dokumenata na osnovu semantičkog značenja.

Ukoliko je svaki dokument predstavljen preko trodimenzionalnog vektora, onda oni zauzimaju neku poziciju u trodimenzionalnom prostoru (koordinatnom sistemu). Dokumenti koji su slični na osnovu semantičkog značenja nalaze se međusobno blizu u ovom prostoru. Vrši se pribavljanje jednog ili više dokumenata čiji su vektori bilski sa vektorom upita.



Slika 7. Pretraga (Retrieval) pomoću provere sličnosti

Jedna od najčešće korišćenih metoda za poređenje vektora je **kosinusna sličnost (cosine similarity)**. U najprostijem obliku, meri se kosinus ugla između dva vektora, pri čemu vrednost može da se kreće od -1 do 1. Vrednost 1 označava da su vektori potpuno slični i da imaju identične smerove, dok vrednost 0 označava da su vektori ortogonalni, što znači da između njih ne postoji sličnost. Vrednost -1 bi značila da su vektori potpuno suprotni u smislu smerova.



Slika 8. Kosinusna sličnost

### Generisanje (Generation)

**Generisanje** je poslednji korak osnovnog RAG sistema. U ovom koraku se formira **prompt**, koji uglavnom sadrži dva ključna elementa: **kontekst** i **pitanje**. Kao kontekst prosleđuju se relevantne informacije koje su preuzete iz baze podataka u prethodnom koraku, dok se kao pitanje prosleđuje originalni korisnički upit. Prompt se pažljivo oblikuje da bi model bolje razumeo šta korisnik traži i u kom formatu bi odgovor trebalo da bude. Na osnovu ovog prompta, veliki jezički model generiše konačan odgovor. **Model koristi kako informacije iz prompta, tako i svoje prethodno naučeno znanje**, kako bi kreirao najrelevantniji i tačan odgovor. Ovaj proces može da proizvede različite formate odgovora, od kratkih, direktnih rešenja, do opširnih sažetaka, u zavisnosti od korisničkog upita.

A diagram of a diagram

Description automatically generated

Slika 9. Slanje prompt-a velikom jezičkom modelu

## Napredni RAG Sistem

A diagram of a process

Description automatically generated

Slika 10. Napredni RAG sistem

### Prevođenje upita (Query Translation)

Prevođenje upita se odnosi na proces transformisanja inicijalnog upita u različite forme ili pod-upite u cilju poboljšanja procesa pronalaženja informacija. Ova transformacija može uključivati razbijanje kompleksnog upita na jednostavnije delove, preformulisanje upita radi boljeg razumevanja konteksta, ili prilagođavanje upita tako da bolje odgovara podacima u bazi. Cilj prevođenja upita je da se maksimizira relevantnost informacija koje će sistem preuzeti. U nekim slučajevima, upit može biti preformulisan kako bi se bolje podudarao sa sličnim terminima i konceptima unutar dokumenata. Ovaj korak je posebno važan kada se koristi više izvora podataka ili višejezični sistemi.

#### Multi-query

**Multi-query** koristi veliki jezički model za generisanje više upita, na osnovnu prvobitnog korisničkog upita . Ovi upiti se mogu izvršavati paralelno, a dobijeni rezultati se zatim zajedno prosleđuju modelu. Ova tehnika je naročito korisna kada jedno pitanje zavisi od više pod-pitanja, što omogućava potpunije prikupljanje informacija.

A diagram of a diagram

Description automatically generated

Slika 11. Embedding i provera sličnosti kod multi-query strategije

Pretrage baza na osnovu upita se mogu izvršavati paralelno, što znači da nije neophodno pristupati istoj bazi podataka za svaki od upita, već se relevantni dokumenti mogu pribaviti iz većeg broja različitih baza podataka koji će se na kraju koristiti za formiranje jednog odgovora.

A diagram of a diagram of a document

Description automatically generated with medium confidence

Slika 12. Paralelna pretraga baza kod multi-query strategije

#### RAG-Fusion

**RAG-Fusion** ima nekih silčnosti u poređenju sa **multi-query** strategijom. Ovde se takođe koristi veliki jezički model za generisanje većeg broja upita na osnovu prvobitnog upita. Kod RAG-Fusion strategije, postoje više različitih retrivera koji mogu koristiti različite načine pretrage, gde svaki od njih radi pretragu za jedan od generisanih upita. Na primer, jedan retriver može raditi vektorsku pretragu, dok drugi može koristiti ključne reči. Ono što izdvaja RAG-Fusion je proces **rangiranja** dokumenata**.** Kada retriveri preuzmu relevantne informacije, RAG-Fusion vrši **evaluaciju i rangiranje** tih dokumenata na osnovu njihove relevantnosti i na taj način omogućava da se najvažnije informacije organizuju tako da budu u prvom planu prilikom generisanja odgovora, osiguravajući da konačan rezultat bude što precizniji i relevantniji.

A diagram of a diagram of a computer

Description automatically generated with medium confidence

Slika 13. RAG-Fusion, pribavljanje i rangiranje dokumenata

#### Decomposition

**Decomposition** je strategija koja podrazumeva razlaganje složenog pitanja ili problema na manje, lakše savladive pod-probleme, koji se zatim rešavaju nezavisno, a njihova rešenja se na kraju spajaju kako bi se dobio sveobuhvatan odgovor.

Postoje dva pristupa:

1. **pristup rekurzivnog odgovaranja (Recursive Answering Approach)**. Ovde se pitanja postsavljaju jedno po jedno, a za njihovo odgovaranje koriste se zajedno odgovori na prethodna pitanja i kontekst odnosno dokumenti koji su pribavljeni za trenutno pitanje. Na taj način zadržava se stara perspektiva i sinhronizuje rešenje sa novom perspektivom, davajući detaljnije rešenje. Ovaj pristup se pokazao efikasnim kod veoma složenih upita.

A diagram of a diagram

Description automatically generated with medium confidence

Slika 14. Pristup rekurzivnog odgovaranja

1. **Pristup paralelnog odgovaranja (Parallel Answering Approach)**. U ovom pristupu, korisnički upit se razlaže na vise delove, kao i ranije. Razlika je u tome što se ti delovi rešavaju paralelno, gde se svako pitanje se obrađuje individualno, a zatim se svi odgovori kombinuju kako bi se dobio mnogo detaljniji kontekst, koji se zatim koristi za odgovaranje na korisnički upit. Ovaj pristup predstavlja efikasno rešenje za većinu slučajeva.

A diagram of a diagram of a brain

Description automatically generated with medium confidence

Slika 15. Pristup paralelnog odgovaranja

#### Step-back prompting

**Step-back prompting** je tehnika koja omogućava modelu da se fokusira na dublje razmišljanje pre nego što reši zadatak. Ona podrazumeva kreiranje apstraktnog pitanja na osnovu korisničkog upita kako bi se dobila šira perspektiva zadatka i prikupio dodatni kontekst. Nakon toga, direktni kontekst se generiše na osnovu korisničkog upita, a oba konteksta se kombinuju za dobijanje preciznijeg odgovora. Ova tehnika je posebno efikasna za pitanja koja zahtevaju razmišljanje, jer omogućava modelu da istraži širu sliku korisničkog upita i uporedi je sa specifičnim kontekstom.

A diagram of steps to a step-by-step process

Description automatically generated

Slika 16. Primer korišćenja step-back prompting-a

#### HyDE (Hypothetical Document Embeddings)

Ideja iza ovog pristupa je da su dokumenti veliki delovi podataka koji sadrže informacije iz dobro formulisanih, gustih tekstova, dok korisnički upit nije tako dobro konstruisan. Iz tog razloga, konstruišemo hipotetički dokument ili odgovor na osnovu korisničkog upita, koji kada se pretvori u vektor, u teoriji je bliži delovima dokumenta nego sam korisnički upit u visoko-dimenzionalnom vektorskom prostoru.

A diagram of a diagram

Description automatically generated

Slika 17. Ilustracija HyDe modela

### Rutiranje (Routing)

Ponekad kod kompleksnijih RAG aplikacija postoje više različitih baza podataka za podatke iz različitih oblasti. **Rutiranje** se najčešće odnosi na proces odlučivanja koja baza podataka ili mehanizam za pretragu će biti korišćeni za pribavljanje relevantnih dokumenata na osnovu korisničkog upita, ali se ne odnosi nužno na izbor baze podataka, već je moguće vršiti i odlučivanje koji prompt koristiti. U suštini, cilj **rutiranja** je da se na osnovu korisničkog upita odredi gde ga treba proslediti dalje. Dva najčešće korišćenih tipova **rutiranja** su:

1. **Logičko rutiranje (Logical routing):**

Kod ovog tipa rutiranja, velikom jezičkom modelu se prosleđuje znanje o tome koje sve vrste baza postoje i šta svaka od njih čuva. Ime je nastalo na osnovu činjenice da veliki jezički model koristi **logiku** da bi odlučio kojoj bazi podataka pritupiti ili koji retriever koristiti.

A diagram of a brain

Description automatically generated

Slika 18. Logičko rutiranje

1. **Semantičko rutiranje (Semantic routing):**

**Semantičko rutiranje** se odvija u dva koraka. U prvom koraku se vrši embedding nad korisničkim upitom i na primer, pormptovima. Nakon toga se njihove numeričke reprezentacije porede u vektorskom prostoru i bira se prompt koji najviše odgovara datom upitu. Odatle i ime “semantičko”, jer vrši poređenje na osnovu semantičkog značenja.

A diagram of a notepad and a note

Description automatically generated

Slika 19. Semantičko rutiranje

### Kreiranje upita (Query Construction)

Postoji veliko interesovanje o povezivanju ljudskog jezika sa različitim tipovima podataka, bili to struktuirani, nestruktuirani ili čak polustruktuirani podaci. Ovaj korak naprednog RAG sistema radi baš to, prevodi korisnički upit sa ljudskog jezika u odgovarajući format koji mogu izvršiti i razumeti ciljani tipovi baza podataka[8].

Od koristi je pomenuti sledeće vrste:

1. **Text-to-metadata-filter:**

Koristi se za prevođenje upita u format koji razumeju vektorske baze podataka sa filterima za metapodatke, omogućavajući struktuirane upite nad embedd-ovanim nestruktuiranim dokumentima [11]. Ova vrsta je uglavnom od najvećeg interesa kada su u pitanju moderni RAG sistemi.

1. **Text-to-SQL:**

Kada su u pitanju struktuirani podaci, veliki izvor predstavljaju relacione baze podataka, pa odatle potreba za prevođenje ljudkog jezika u SQL naredbe. Važno je naglasiti da je u cilju generisanja što tačnijih naredbi potrebno proslediti sve CREATE TABLE naredbe u bazi, nakon čega sledi nekoliko primera SELECT naredbi [11].

1. **Text-to-SQL+semantic:**

U savremenom radu sa podacima, kombinovano skladištenje struktuiranih i nestruktuiranih podataka postaje sve češće. Dodavanje podrške za vektore u relacionim bazama podataka igra ključnu ulogu u omogućavanju hibridne pretrage. Moguće je vršiti proveru sličnosti nad vektorima unutar relacione baze podataka i na taj način poboljšava text-to-SQL pretragu uz pomoć semantičkih operatora [11].

1. **Text-to-Cypher:**

Cypher predstavlja upitni jezik definisan za rad sa graf bazama podataka. Graf baze se koriste zato što rešavaju neke probleme i nedostatke koji se javljaju kod drugih tipova baza podataka. Vektorske baze podataka ne razumeju međusobne odnose između vektora, dok relacione baze, iako dobro modeluju vezu među entitetima, promene šeme mogu da budu mnogo skupe. Graf baze podataka su idealne za odnose tipa više-prema-više i za hijerarhije koje je teško predstaviti u tabelarnoj formi [11].

### Indeksiranje (Indexing)

Osnove indeksiranja i embedding-a su obrađene u poglavlju posvećenom osnovnama RAG sistema. U ovom poglavlju će fokus biti na različitim metodama indeksiranja.

#### Chunk Optimization

Chunking, kao što sam naziv sugeriše, podrazumeva razbijanje velikog dokumenta na manje, kontekstualne delove (chunk-ove). Time što se dokument deli na lakše upravljive delove, omogućava se modelu da ga efikasnije obradi [23].

Najpre treba razmisliti o tome kakav je tekst nad kojim se vrši embedding. Vršenje embeddinga nad prevelikim ili premalim segmentima može dovesti to manje optimalnih rezultata. Iz tog razloga je vrlo važno odabrati optimalnu veličinu segmenata kako bi se osigurala tačnost dobijenih rezultata [22].

Izbor odgovarajuće strategije segmentacije zahteva pažljivo razmatranje faktora kao što su priroda sadržaja, embedding model, optimalna veličina bloka, očekivana dužina i složenost korisničkih upita, kao i način na koji će rezultati biti korišćeni u konkretnoj aplikaciji [22].

Mogu se izdvojiti 5 novoa:

* **Fixed Size Chunking**: Tekst se deli na segmente fiksne dužine, bez obzira na sadržaj ili strukturu [22], [24].
* **Recursive Chunking**: Tekst se deli na manje segmente na hijerarhijski način koristeći skup separatora, i rekurzivno se deli dok se ne postigne željena veličina segmenta [22], [24].
* **Document-Based Chunking**: Dokument se deli prema njegovoj strukturi, uzimajući u obzir tok i format sadržaja [24].
* **Semantic Chunking**: Tekst se deli na osnovu semantičkog značenja, grupišući delove teksta koji su semantički slični [23], [24].
* **Agentic Chunking**: Koristi velike jezičke modele da odluči koliko i koji deo teksta treba uključiti u segment na osnovu konteksta [23], [24].

#### Multi-representation indexing/ Multi vector representation

Inspiracija i ideja potiču iz rada [25] koji uvodi metodu korišćenja velikih jezičkih model a za kreiranje propozicija, odnosno sažetaka, koji su optimizovani za pretragu. Zaključak je da se mogu dobiti mnogo bolji rezultati ukoliko se koriste propozicije umesto čistih dokumenata.

Multi-representation indexing fukncioniše tako što se najpre kreiraju propozicije (sažeci), koji se potom embedduju i na osnovu njih se vrši provera sličnosti sa korisničkim upitom. Celi dokumenti se čuvaju zasebno u odgovarajućem skladištu dokumenata. Kada se pronađu najsličniji sažeci, vraćaju se celi dokumenti kojima oni odgovaraju [21].

A diagram of a multi-representation

Description automatically generated

Slika 20. Multi-representation indexing

#### Specialized embeddings (ColBERT)

**ColBERT**(**Co**ntextualized **l**ate interaction over **BERT**) se prvi put javlja u radu [26] kao ideja za efikasnije pretraživanje i rangiranje dokumenata korišćenjem dubokih jezičkih modela, konkretno BERT-a. Ovaj model uvodi inovativnu arhitekturu kasne interakcije, koja omogućava nezavisno kodiranje upita i dokumenata, a zatim koristi brzu, ali preciznu interakciju za procenu sličnosti. Time se smanjuju računarski troškovi, omogućava pretraga velikih kolekcija dokumenata, i značajno ubrzava procesiranje upita u poređenju sa tradicionalnim BERT modelima.

Umesto da koristi klasičnu metodu gde se svaki upit-dokument par mora obraditi zajedno, ColBERT uvodi kasnu interakciju (late interaction), gde se upiti i dokumenti nezavisno enkodiraju koristeći BERT [26].

Upit se najpre tokenizuje i priprema, dodaje se specijalni token [Q] kako bi se označilo da je to upit. BERT model potom obrađuje upit i generiše vreću embeddinga (bag of embeddings), što su vektorske reprezentacije reči iz upita [26].

Dokumenti se sličnim procesom enkodiraju. Pre dokumenta se dodaje specijalni token [D] kako bi se označilo da je to dokument, a nakon BERT obrade generiše se vreća embeddinga za dokument. Ponekad se uklanjaju tokeni interpunkcije kako bi se smanjio broj embeddinga [26]..

Kada se upit i dokument enkodiraju u vreće embeddinga, ColBERT izračunava sličnost između embeddinga iz upita i dokumenta koristeći kosinusnu sličnost koja je ranije detaljnije objašnjena. Zatim računa relevantnost dokumenta prema upitu na osnovu maksimuma sličnosti između embeddinga [26].

A diagram of a document

Description automatically generated

Slika 21. Opšta arhiktektura ColBERT-a [26]

#### Hierarchical indexing (RAPTOR)

**RAPTOR** (**R**ecursive **A**pproach for **P**arsing **T**exts for **O**ptimal **R**etrieval) je inovativna metoda, prvi put viđena u [27], dizajnirana za unapređenje performansi RAG sistema, posebno kada se radi sa dugim i složenim tekstovima koji često imaju hijerarhijsku strukturu i više podtema. Dok tradicionalni RAG sistemi obično preuzimaju kratke, neprekidne delove teksta iz skupa dokummenata, što ograničava njihovu sposobnost da obuhvate širi kontekst dokumenta. RAPTOR rešava ovaj problem kreiranjem rekurzivnog stabla .

RAPTOR segmentira dokumente u kratke delove (oko 100 tokena). Ovi delovi se enkodiraju koristeći BERT-bazirani enkoder (SBERT) i formiraju listove stabla. Konačno stablo se dobija rekurzivnim enkodiranjem, klasterovanjem i sažimanjem delova teksta, stvarajući više slojeva apstrakcije.

RAPTOR koristi algoritam za klasterovanje kako bi grupisao slične delove teksta. Nakon klasterovanja, vrši se sažimanje sadržaja grupisanih tekstova pomoću jezičkog modela. Sažeti tekst se ponovo enkodira i dalje klasteriše, pri čemu se proces nastavlja dok dalja klasterizacija ne postane neizvodljiva. Rezultat je višeslojno stablo gde svaki nivo predstavlja različit stepen sažimanja, od detaljnih informacija na dnu do apstraktnih sažetaka na vrhu. Koristi se meko klasterovanje, omogućavajući da delovi teksta pripadaju više klastera. Ova fleksibilnost je važna jer pojedinačni segmenti teksta često sadrže informacije relevantne za više tema, što omogućava modelu da bolje obuhvati složenost stvarnih dokumenata.

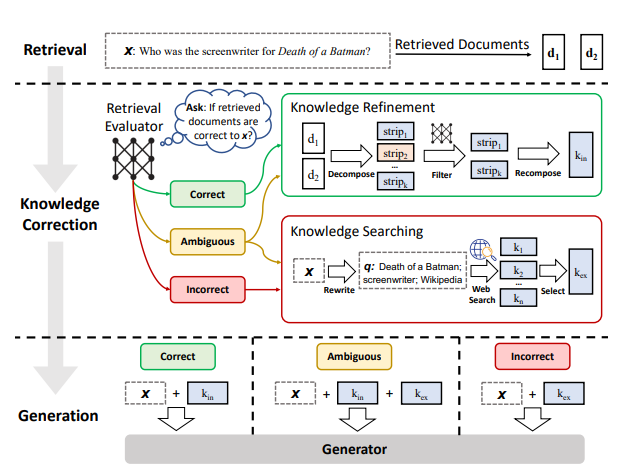
A diagram of a diagram

Description automatically generated

Slika 22. RAPTOR stablo

### CRAG

**CRAG (C**orrective **R**etrieval**-A**ugmented **G**eneration**)** se prvi put javlja u [28], kao rešenje za problem netačnih ili nerelevantnih dokumenata prikupljenih tokom pretrage kod RAG-a. Ključni problem koji CRAG rešava je kako model postupa kada retrieval (pretraga) ne uspe da pronađe tačne ili relevantne informacije. CRAG je dizajniran tako da bude **lagan i skalabilan**, što znači da se može lako integrisati u različite RAG sisteme bez velikog povećanja resursa. U eksperimentima se pokazalo da CRAG značajno poboljšava performanse na zadacima koji zahtevaju generisanje teksta, kako u kratkoj, tako i u dugoj formi [28].



Slika 23. Pretraga i generisanje kod CRAG-a [28]

CRAG uvodi sloj evaluacije preuzetih dokumenata kako bi osigurao njihovu relevantnost za postavljeno pitanje. Kada model preuzme dokumente, CRAG koristi poseban **evaluacioni model** da proceni koliko su ti dokumenti relevantni za upit korisnika. Ova procena se izražava kao nivo poverenja (confidence degree), koji zatim aktivira jednu od tri moguće akcije [28]:

1. **Correct (Tačno)** - Ako su dokumenti ocenjeni kao relevantni, oni prolaze kroz dodatni korak obrade. CRAG koristi tehniku koja rastavlja informacije iz dokumenata, filtrira nebitne delove, a zatim ponovo sklapa preostale informacije u precizniji oblik. Ova faza se naziva **knowledge refinement** (pročišćavanje znanja).
2. **Incorrect (Netačno)** - Ako dokumenti nisu relevantni, odbacuju se i preusmerava se pretraga na šire izvore, kao što je pretraga putem interneta, u cilju pronalaska tačnijih informacija. Ova metoda proširuje tradicionalne RAG pristupe tako što koristi veći opseg informacija kada su lokalni izvori neadekvatni.
3. **Ambiguous (Neodređeno)** – Ukoliko se ne može jasno utvrditi da li su dokumenti tačni ili netačni, koristi se "neodređena" akcija, koja kombinuje oba izvora (lokalnu pretragu i pretragu interneta) kako bi postigao uravnotežen pristup.

### RAG vs Fine-tuning (Fino podešavanje)

RAG nije konvencionalna metoda za fino podešavanje u mašinskom učenju. Umesto toga, RAG je arhitektura koja kombinuje jezičke modele sa sistemom za pretragu kako bi poboljšala generaciju odgovora korišćenjem spoljašnjih podataka. Dok tradicionalno fino podešavanje prilagođava unapred treniran model određenom zadatku kroz dodatnu obuku, RAG dinamički uključuje relevantne informacije iz eksternih izvora tokom generacije, čime poboljšava kvalitet i relevantnost odgovora. Međutim, komponente RAG sistema, poput jezičkog modela i sistema za pretragu, mogu se dodatno fino podešavati kako bi se unapredila njihova preciznost i performanse.

Evo kratkog poređenja između RAG-a i finog podešavanja [29]:

|  |  |  |
| --- | --- | --- |
|  | RAG | Fino podešavanje |
| Ažuriranje Znanja | Direktno ažurira bazu podataka za pretragu, što je pogodno za dinamična okruženja podataka i zahteva ređe ponovno obučavanje | Oslanja se na statičke podatke, što zahteva ponovno obučavanje za ažuriranja i manje je pogodno za informacije koje se često menjaju |
| Spoljno Znanje | Efikasno koristi spoljne resurse, posebno sa strukturiranim i nestrukturiranim bazama podataka | Usaglašava spoljno znanje sa predobukom velikih jezičkih modela, ali je manje efikasno za dinamične izvore podataka |
| Obrada Podataka | Zahteva minimalnu obradu podataka | Ograničenja u datasetovima mogu uticati na performanse |
| Prilagođavanje Modela | okusira se na integraciju spoljnog znanja, ali nudi manje prilagodljivosti u ponašanju ili stilu modela | Omogućava veću kontrolu nad ponašanjem modela, stilom i domenom znanja, prilagodljiv je specifičnim potrebama. |
| Interpretabilnost | Pruža veću interpretabilnost zbog mogućnosti praćenja odgovora do izvora podataka | Često funkcioniše kao crna kutija, što otežava razumevanje odluka modela, rezultirajući manjom interpretabilnošću |
| Zahtevi za Latenciju | Uključuje pretragu podataka, što može povećati latenciju | Obično nudi manju latenciju jer funkcioniše bez potrebe za pretragom podataka nakon obuke |
| Smanjenje Halucinacija | Manje je sklono halucinacijama zbog oslanjanja na pretražene dokaze | Može smanjiti halucinacije sa obukom specifičnom za domen, ali je i dalje podložno njima kod nepoznatih ulaza |
| Etika i Privatnost | Pitanja se javljaju zbog skladištenja i pretrage tekstova iz spoljnih baza podataka | Etička i privatnosna pitanja mogu nastati zbog osetljivog sadržaja u podacima za obuku. |
| Računarski Resursi | Zahteva resurse za strategije pretrage i održavanje baze podataka | Zahteva resurse za pripremu datasetova, definisanje ciljeva i računarsku snagu za obučavanje |

### Budućnost RAG-a

Budućnost RAG modela u oblasti veštačke inteligencije i obrade prirodnog jezika donosi niz uzbudljivih unapređenja koja će poboljšati njihovu efikasnost i proširiti primenu u različitim domenima [30]. Jedno od ključnih unapređenja je **poboljšana integracija pretrage i generisanja** informacija. Ovo će omogućiti RAG modelima da se fokusiraju na bolje povezivanje ovih komponenti kako bi omogućili nesmetano funkcionisanje, odnosno obezbeđivanje da preuzete informacije budu visoko relevantne i tačne za zadatak koji je u toku.

RAG modeli će se takođe širiti u specifične domene poput **zdravstva, prava i finansija**, što će omogućiti da modeli efikasno koriste znanja iz određenih industrija, pružajući precizne i stručne odgovore na složene upite specifične za te oblasti.

Još jedna bitna oblast unapređenja je **povećanje efikasnosti i skalabilnosti** RAG sistema. Optimizacija računarskih procesa omogućavaće modelima da obrađuju veće skupove podataka i sve složenije upite, čime će biti otvoren put za širu komercijalnu i industrijsku primenu. Takođe, unaprediće se **razumevanje konteksta**, što će omogućiti modelima da bolje interpretiraju nijanse, dvosmislenosti i specifične namere korisnika.

U skladu sa brzim promenama u svetu, očekuje se razvoj naprednijih mogućnosti za **preuzimanje informacija u realnom vremenu**. To će podrazumevati preuzimanje aktuelnih podataka iz dinamičnih izvora, poput interneta ili ažurnih baza podataka, što će ove sisteme učiniti posebno korisnim za obradu novosti, trendova i aktuelnih događaja.

Budućnost donosi i **multimodalne RAG sisteme** koji će ići dalje od teksta, uključujući slike, video zapise i audio sadržaje. Ova unapređenja će stvoriti sveobuhvatnije sisteme sposobne da razumeju i generišu odgovore u različitim tipovima medija.

**Interaktivno učenje** i **integracija povratnih informacija** dodatno će poboljšati RAG sisteme, omogućavajući im da se unapređuju kroz vreme zahvaljujući učenju iz korisničkih interakcija i kontinuiranom poboljšanju performansi na osnovu povratnih informacija u realnom vremenu. Etička pitanja takođe će postati sve važnija, sa naporima da se obezbedi **odgovorna upotreba** RAG modela, rešavajući probleme kao što su pristrasnost, dezinformacije i transparentnost u preuzimanju i obradi podataka.

Očekuje se i unapređenje u **višelingvističkim sposobnostima** RAG modela, omogućavajući efikasniju pretragu i generisanje sadržaja na više jezika. Pored toga, potencijal za veću **saradnju sa ljudskom inteligencijom** omogućiće da RAG sistemi rade zajedno sa ljudskim ekspertima, kombinujući efikasnost i skalabilnost veštačke inteligencije sa kreativnošću i prosuđivanjem ljudi.

Kako se tehnologija RAG-a nastavlja razvijati, ovi modeli postaće sve sofisticiraniji, prilagodljiviji i ključni za razne primene, pomerajući granice sposobnosti veštačke inteligencije u razumevanju i generisanju ljudskog jezika.

# Implementacija RAG sistema sa Q&A Chatbotom za pravila društvenih igara

U ovom poglavlju biće opisana implementacija Q&A chatbota baziranog na RAG tehnici za pravila društvenih igara. Sistem je razvijen korišćenjem Flask-a za API, LangChain-a za integraciju sa velikim jezičkim modelima, i Pinecone-a za upravljanje vektorskom bazom podataka. Cilj ovog projekta je da se korisnicima omogući da postavljaju pitanja o pravilima različitih društvenih igara i dobijaju tačne i relevantne odgovore.

Korisnici mogu uneti svoja pitanja putem korisničkog interfejsa, koji komunicira sa backend-om preko Flask API-ja. Backend koristi LangChain za generisanje odgovora na osnovu pretraživanja vektorske baze podataka koja sadrži pravila igara. Sistem implementira mehanizme za unapređeno pretraživanje i generisanje odgovora koristeći RAG tehniku, što omogućava pružanje preciznih informacija i uvida u pravila društvenih igara.

Pored objašnjenja osnovnih funkcionalnosti sistema, u ovom poglavlju će biti prikazani izazovi projekta, kao i delovi koda koji ilustruju kako se integracija između različitih komponenti ostvaruje i kako sistem obrađuje korisničke upite.

## Izazovi projekta

Implementacija Q&A chatbota za pravila društvenih igara nosi sa sobom mnoge izazove koji čine tradicionalne metode pretrage i generisanja odgovora neadekvatnim. Pravilnici društvenih igara su često napisani u različitim formatima i stilovima, a informacije su često rasute po velikim i kompleksnim dokumentima. Tradicionalni sistemi za pretragu se uglavnom oslanjaju na jednostavno pretraživanje ključnih reči, što može rezultirati netačnim ili nerelevantnim odgovorima kada je potrebno precizno odgovoriti na specifična pitanja.

Jedan od ključnih izazova je varijabilnost pravila u različitim društvenim igrama. Svaki pravilnik može biti organizovan na jedinstven način, što otežava jednostavno pretraživanje ključnih informacija. Neki pravilnici su sažeti i strukturisani, dok su drugi opširni i obuhvataju specifične situacije i izuzetke. Tradicionalni Q&A sistemi često ne uspevaju da prepoznaju kontekst i precizno pronađu relevantan deo pravilnika.

### Zašto RAG?

U ranijim oblastima je detaljno objašnjeno kako RAG sistem funkcioniše kao i njegove prednosti i primene. U kontekstu ovog projekta, RAG sistem se izdvaja kao posebno koristan zbog svoje sposobnosti da efikasno pretražuje i generiše odgovore iz kompleksnih i nestrukturisanih izvora kao što su pravilnici društvenih igara.

Pravilnici društvenih igara često sadrže raznolike i specifične informacije, a jednostavna pretraga ključnih reči često nije dovoljna da obuhvati sve nijanse i kontekste pravila. RAG nudi rešenje ovih problema omogućavajući dubinsku pretragu kontekstualnih informacija i generisanje tačnih odgovora prilagođenih korisničkom upitu. Na ovaj način, RAG sistem može prepoznati specifične delove pravilnika relevantne za pitanje korisnika, što značajno unapređuje tačnost i korisnost odgovora.

### Potreba za tačnim informacijama

Za chatbot aplikaciju koja se bavi pravilima društvenih igara, ključno je da informacije budu tačne i relevantne, posebno kada su pravila igre bitna za donošenje odluka tokom igre gde pogrešne odluke mogu dovesti do frustracije igrača. RAG pomaže u rešavanju ovog problema koristeći pretragu vektorskih reprezentacija podataka, što osigurava da su povučeni podaci zaista relevantni i u skladu sa postavljenim pitanjem.

## Alati i tehnologije

Za implementaciju ovog projekta, za logiku i komunikaciju sa velikim jezičkim modelom na serverskoj strani korišćen je Python, dok je za korisnički interfejs na klijentskoj strani korišćen React.

### Flask

Flask je popularan mikro-framework za Python, razvijen sa ciljem da omogući jednostavno i brzo kreiranje web aplikacija i API-ja. Flask predstavlja skup biblioteka i modula koji omogućavaju programerima da razvijaju web aplikacije bez potrebe da se bave detaljima nižeg nivoa poput protokola, upravljanja nitima i sličnim tehničkim aspektima [31]. U ovom projektu, Flask je korišćen za kreiranje API-ja koji omogućava komunikaciju između front-enda i back-enda. Zbog svoje jednostavne konfiguracije, Flask je bio idealan izbor za implementaciju, jer omogućava brzu i efikasnu razmenu podataka između korisničkog interfejsa i jezičkog modela koji se koristi za generisanje odgovora.

A computer code with text

Description automatically generated with medium confidence

Slika 24. Flask API implementacija

### LangChain

LangChain predstavlja framework koji služi za razvoj aplikacija koje koriste velike jezičke modele. Nudi alate za integraciju velikih jezičkih modela sa spoljnim izvorima podataka, API-jima i korisničkim interakcijama, omogućavajući složene procese poput odgovaranja na pitanja, sažimanja i pretraživanja dokumenata [32].

Framework se sastoji od dve open-source biblioteke, čije konkretne primene će biti opisane kasnije :

* + 1. langchain**-**core- Osnovne apstrakcije i LangChain jezik izraza.
    2. langchain-community - 3rd-party integracije

### Pinecone

Kao što je već pomenuto, vektorske baze podataka igraju ključnu ulogu u procesu pretrage informacija iz baza podataka. U tu svrhu koristi se Pinecone, upravljana vektorska baza podataka sa jednostavnim API-jem, koja omogućava dugoročno skladištenje vektora za AI aplikacije [33].

Svaka Pinecone baza podataka se sastoji od jednog ili većeg broja indeksa. Indeks predstavlja strukturu podataka koja omogućava čuvanje, pretragu i pristup vektorima koji se nalaze u njemu. Svaki indeks ima i svoju dimenziju koja označava dimenzionalnost vektora koji se mogu čuvati u njemu. Dimenzionalnost vektora zavisi od embedding modela koji se koristi (u projektu je ta vrednost 1536).

## Opis funkcionalnosti i koda

Nakon što smo detaljno proučili ključne alate i tehnologije, sledeći korak je razumevanje celokupne funkcionalnosti projekta. Ovaj deo će prikazati kako različite komponente sistema međusobno funkcionišu i doprinose ostvarivanju ciljeva projekta. Analiziraćemo kako se informacije obrađuju, skladište i koriste za generisanje odgovora, što je ključno za efikasno rešavanje problema vezanih za pravila društvenih igara.

### Učitavanje dokumenata

Dokumenti se, u zavisnosti od obima projekta i potreba za skalabilnošću, mogu čuvati u posebnim bazama podataka. Međutim, s obzirom na to da se ovaj projekat ne bavi velikim količinama dokumenata, odlučeno je da se koriste lokalni fajl sistem kao za čuvanje dokumenata. Ovaj pristup omogućava jednostavniju implementaciju i manje resursno zahtevan sistem, jer lokalni fajl sistem pruža dovoljno prostora i fleksibilnosti za trenutne potrebe projekta. Ukoliko se obim dokumentacije poveća, postoji mogućnost prelaska na drugi vid čuvanja dokumenata.

LangChain nudi širok spektar document loadera (TextLoader, PyPDFLoader, YouTubeLoader, HTMLLoader), koji predstavljaju specijalizovane alate za učitavanje podataka iz različitih izvora kao što su tekstualni fajlovi, PDF-ovi, web stranice, baze podataka ili čak transkripti video zapisa. Njihova glavna funkcija je da automatizuju proces prikupljanja podataka iz tih izvora, formatiraju ih i pripreme za dalje obrade.

Document loaderi u LangChain-u pružaju metodu **load**, koja je ključna za učitavanje podataka iz podešenog izvora. Kao izlaz ove metode dobijaju se Document objekti, koji sadrže:

* Tekstualni sadržaj - Glavni deo dokumenta, npr. sadržaj stranica iz PDF-a ili transkript videa.
* Metapodaci - Dodatne informacije o dokumentu, poput izvora, stranice, autora, URL-a, itd.

Neki document loaderi podržavaju i opcionu metodu **lazy\_load**, koja omogućava "lazy loading" podataka. To znači da se podaci učitavaju postepeno, samo kada su potrebni, što smanjuje potrošnju memorije kod većih setova podataka.

U ovom projektu koristi se PyPDFLoader jer su pravila društvenih igara obično dostupna u obliku zasebnih PDF fajlova koji se jednostavno mogu preuzeti. Ova metoda omogućava efikasno i precizno učitavanje pravila direktno iz tih fajlova, čime se olakšava njihov dalji rad u sistemu.

A screenshot of a computer program

Description automatically generated

Slika 25. Učitavanje fajlova

DATA\_PATH – putanja do direktorijuma u kome se nalaze PDF fajlovi

clean\_document - Ova funkcija prima tekst kao ulaz i čisti ga tako što uklanja višak belih prostora (tabulator, novi red) i trimuje krajeve. Ovaj korak je važan za pripremu teksta pre obrade

load\_pdf - Ova funkcija koristi PyPDFLoader za učitavanje PDF fajla sa zadate putanje (file\_path) koja se prosleđuje kao parametar. Zatim koristi metodu PyPDFLoader-a load\_and\_split da podeli PDF sadržaj na stranice. Rezultat je lista Document objekata koji sadrže tekst i metapodatke.

load\_documents – Ova funkcija najpre pretražuje direktorijum DATA\_PATH i generiše listu (pdf\_files) putanja do svih PDF fajlova koji se nalaze u njemu. Zatim koristi ThreadPoolExecutor za paralelno učitavanje PDF fajlova koristeći funkciju load\_pdf. Ovo omogućava efikasnije procesiranje više fajlova istovremeno. Učitani dokumenti se spajaju u jednu listu (flat\_documents), zatim se tekst u svakom Document objektu čisti koristeći funkciju clean\_document. Čisti dokumenti se zatim dodaju u listu cleaned\_documents

### Inicijalizacija baze

Sada kada imamo učitane dokumente, spremni su za njihovu dalju obradu i čuvanje u vektorsku bazu podataka, koju najpre moramo inicijalizovati.

A screenshot of a computer program

Description automatically generated

Slika 26. Pinecone inicijalizacija i kreiranje indeksa

create\_new\_index – funkcija koja u našoj Pinecone bazi kreira novi indeks sa zadatim imeno, dimenzijom 1536 (mora da se poklapa sa dimenzijom vektora koje vraća embedding model) i metrikom “cosine”, što označava da će se u ovom indeksu pretraga vršiti na osnovu kosinusne sličnosti koja je već detaljnije objašnjena. Funkcija takođe podešava i specifikacije indeksa

initialize\_pinecone – ova funkcija učitava API ključeve pomoću load\_dotenv, proverava da li indeks već postoji i, ako nije prisutan, kreira ga korišćenjem funkcije create\_new\_index, a zatim vraća instancu tog indeksa za dalju upotrebu u Pinecone operacijama.

Ukratko, ovaj deo koda omogućava efikasno upravljanje i pristupanje Pinecone indeksima kako bi se velktori dokumentata mogli sačuvati za brzo pretraživanje.

### Kreiranje i čuvanje vektora

Sledeći korak u obradi dokumenata odnosi se na odluku o načinu njihove podele i indeksiranja. U ovom projektu, odluka je bila da se koriste LangChain-ovi Text Splitteri. Kada se radi sa dugim tekstovima, neophodno je podeliti ih na manje delove ili "chunkove", kako bi odgovarali kontekstualnom prozoru modela. LangChain nudi nekoliko različitih tipova Text Splittera koji olakšavaju ovu obradu [34], neki od njih su:

* Recursive
* HTML
* Code
* Token
* Character
* Semantic

U ovom projektu, odluka je bila da se koristi **RecursiveCharacterTextSplitter** zbog njegove sposobnosti da zadrži semantički povezane delove teksta zajedno. Pravila društvenih igara često sadrže informacije koje su tematski povezane unutar različitih sekcija, a ovaj tip chunkinga omogućava da se tekst podeli na manje delove na način koji čuva tu povezanost. RecursiveCharacterTextSplitter pokušava da inteligentno raspodeli tekst kako bi se obezbedila relevantnost unutar svakog dela, što olakšava preciznije pretrage i poboljšava kvalitet odgovora na upite korisnika o pravilima igre.

A screen shot of a computer code

Description automatically generated

Slika 27. Chunking funkcija

A computer code with text

Description automatically generated with medium confidence

Slika 28. Glavna funkcija za indeksiranje i embedding

embed\_and\_index\_documents – Ova funkcija obavlja čitav proces embeddinga i indeksiranja ranije učitanih dokumenata. Prvo se dokumenti dele na manje delove (chunkove) pomoću funkcije **split\_documents**, a zatim se za svaki chunk izračunava jedinstveni identifikator (Slika 30). Ovi identifikatori se koriste za proveru postojanja duplikata, sa ciljem da se u indeks dodaju samo novi vektori. Na taj način se izbegava dodavanje već postojećih podataka, čime se štede vreme i resursi. U toku izračunavanja identifikatora, poziva se funkcija get\_embeddings, koja kreira embedding za svaki chunk pomoću određenog embedding modela (u projektu se koristi OpenAI text-embedding-3-small, što pokazuje slika 29), čime se dobija njihova numerička reprezentacija. Na samom kraju, novi vektori se dodaju u bazu podataka pozivom funkcije add\_vectors\_to\_database.

A computer code with text

Description automatically generated

Slika 29. Implementacija OpenAI modela za kreiranje embeddinga

A screenshot of a computer code

Description automatically generated

Slika 30. Računanje identifikatora za svaki chunk

A close-up of a computer screen

Description automatically generated

Slika 31. Vektor u bazi sa svojim metapodacima

### Pretraga i generisanje (Retrieval and generation)

Cela funkcionalnost LangChain-a zasniva se na konceptu lanaca (chains), koji predstavljaju osnovni način organizacije i upravljanja radnim tokovima. Svaka operacija u LangChain-u, bilo da se radi o interakciji sa jezičkim modelima, pretrazi u bazama podataka, ili primeni različitih alata, strukturisana je u sekvence koraka unutar lanca. Ovaj pristup omogućava modularnost i fleksibilnost, jer se svaki korak može nezavisno kontrolisati i prilagoditi specifičnim potrebama aplikacije. Na taj način, lanci obezbeđuju jednostavno povezivanje različitih komponenti sistema i optimizuju ceo proces rada [32].

Postoje dva tipa lanaca koje LangChain podržava:

1. LCEL (LangChain Execution Language) chains - Ovi lanci nude fleksibilnost i omogućavaju korisnicima da lako modifikuju svaki deo lanca.
2. Legacy Chains - kreirani pomoću starijih metoda nasleđivanja klasa.

Prednost LCEL lanaca leži u mogućnosti jednostavnog modifikovanja i praćenja svakog koraka lanaca. Pored toga, ovi lanci podržavaju strimovanje (streaming), obavljanje asinhronih operacija, kao i obradu više upita odjednom (batch processing). Iz tih razloga, u ovom projektu koriste se LCEL lanci, kod kojih se za nadovezivanje internih koraka koristi operator | (Slika 32). Ovaj operator povezuje dva koraka tako da izlaz prvog automatski postaje ulaz drugog, čime se pojednostavljuje radni tok i omogućava lakše kreiranje složenih procesa.



Slika 32. Primer LCEL lanca

Prvi korak u procesu pretrage, koji omogućava davanje tačnih i relevantnih odgovora, jeste prevođenje korisničkog upita. U ovom projektu, za tu svrhu je korišćen **RAG-Fusion** pristup. RAG-Fusion je odličan izbor kada je u pitanju RAG sistem za pravila o društvenim igrama iz nekoliko razloga:

* **Višestruki upiti**: Društvene igre često imaju složena i specifična pravila koja mogu biti opisana na različite načine u dokumentaciji. RAG-Fusion omogućava generisanje više upita na osnovu jednog korisničkog pitanja, čime se pokrivaju različiti načini na koje pravila mogu biti formulisana.
* **Kombinovanje relevantnih informacija**: Pravila društvenih igara često imaju povezane koncepte koji su raspoređeni u različitim delovima dokumenta. RAG-Fusion omogućava sistemu da prikupi informacije iz više izvora, što pomaže da se svi važni delovi pravila objedine i predstave korisniku na jasan i konzistentan način.
* **Preciznost pretrage**: Korišćenjem tehnike Reciprocal Rank Fusion, ovaj pristup olakšava rangiranje najbitnijih i najrelevantnijih dokumenata, što je posebno korisno kada je potrebno pronaći specifičan deo pravila u dokumentima koji sadrže puno informacija.
* **Kontekstualna relevantnost**: RAG-Fusion omogućava da se ključne informacije efikasno sažmu u finalni odgovor, kako bi korisnik dobio tačnu i potpunu informaciju koja je direktno vezana za njegovo pitanje o pravilima igre.

A white background with red and blue text

Description automatically generated