

**Olatz ARREGI URIARTE**

Donostiako Informatika Fakultateko irakasle eta ikertzailea, UPV/EHU

RAG teknikak: elkarrizketa sistemen erantzunak hobetzen

Hain modan jarri diren txatbotak, ChatGPT-k, DeepSeek-ek edo beste edozeinek, arazoak eman ditzakete testuinguru jakin batean erantzun zehatzak bilatzen baditugu. Askotan, ematen dizkiguten erantzunak zehatzak ez izateaz gain, ez dira egiazkoak, eta hori larria izan daiteke. Hain zuzen ere, horri aurre egiteko garatu dira RAG (Retrieval Augmented Generation, edo euskaraz BAS, Berreskurapenak Aberastutako Sor-kuntza) teknikak, zeinak testuinguru espezifikoetan erantzunen kalitatea hobetuko duten. Horretarako, sistema horiek entrenamenduan erabili ez dituzten ezagutza-base jakinetatik hartuko dute informazioa, emaitzak nabarmen hobetuz.

RAG sistemen funtzionamendua, modu sinplean azalduta, honela labur daiteke:

1. Erabiltzaileak galdera egiten du.
2. Eskaera horri erantzuteko ezagutza-base edo dokumentu multzo espezifikoetan informazio esanguratsua bilatzen da informazioaren berreskuratze-teknikak erabilita.
3. Galdera aberastu hori (jatorrizko galdera + datu-baseetatik lortutako informazioa) hizkuntza-eredu handi sortzaileari (elkarrizketa-sistemek oinarrian dutenari) bidaltzen zaio.
4. Sistemaren erantzuna zuzenagoa eta zehatzagoa izango da, ereduak ikasitakoari berreskuratze-teknikak ere gehitu zaizkiolako. (Irudian ikus daiteke prozesua grafikoki).

Hasiera batean, eredu estatikoetan oinarritutako ChatGPT bezalako laguntzaileek soilik entrenamendu unean “ikasía” zuten informazioa erabiltzen zuten erantzuteko, eta ondorioz, informazio okerra, asmatutakoa edo zaharkitua eman zezaketen. Adibidez, Claude laguntzaileak 2024ko urrira arteko informazioa daukala dio, eta ChatGPT-k 2024ko ekainera artekoa. Hala ere, gaur egun laguntzaile horiek RAG tekniken antzekoak erabiltzen dituzte, eta, informazio eguneratua behar dutenean, webguneetan edo erabiltzaileak ematen dizkion dokumentuetan ere bilatzen dute. Beti, baina, ez dute hori egiten, eta gaur bertan (martxoaren 14an) ChatGPT-ri galdetzen bazaio zein den AEBetako presidentea, hauxe erantzuten du:

AEBetako presidentea Joe Biden da. 2021eko urtarrilaren 20an hartu zuen kargua, eta 2024ko azaroaren 5eko hauteskundeetan berriro hautatu zuten. Horrela, 2025eko urtarrilaren 20an hasi zuen bere bigarren agintaldia. Beraz, 2025eko martxoaren 14an, Joe Biden da AEBetako presidentea.

RAG sistema errealek, berriz, beti egiten dute aurreneko bilaketa ezagutza-baseetan eta, ondorioz, informazio eguneratua eta fidagarria erabiltzen duten heinean, erantzun zehatzagoak eta zuzenagoak ematen dituzte. Gainera, erantzuna nondik atera duten ere jakin daiteke, iturriak gehitu daitezkeelako emaitzetan. Abantailekin jarraituz, sistema horiek oso egokiak dira domeinu zehatzetan aplikatzeko, izan medikuntzan, izan industrian edo izan legegintzan, informazio propioa edo per-

sonalizatua erabil dezaketelako.

Idea bat egiteko, medikuntza arloko adibide sinplifikatu bat ekarri dugu hona. Demagun galdera hau egiten diogula laguntzaile arrunt bati: «Azaldu arnasa gutxiegitasun akutua-ren azken tratamendu-gidak paziente diabetikoentzat».

Estilo honetako erantzunak jasoko genituzke RAG tekniken erabileraren arabera:

BAS teknologia erabili gabe:

Tratamendu-gida orokorrekin oxigenoterapia, bentilazio mekanikoa eta posizio pronoren erabilera gomendatzen dute.

Paziente diabetikoen kasuan, glukosa-mailen kontrola funtsezkoa da, eta normalean 140-180 mg/dL artean mantentzea.

BAS teknologia erabiliz:

Tratamendu eguneratuak, 2024ko American Thoracic Society eta European Respiratory Society-ren gida berrietan oinarrituta:

1. Oxigenoterapia: SpO2 maila %92-96 artean mantentzea gomendatzen da diabetikoetan, hiperoxigenazio gorabeherak ekiditeko (Wang et al., 2024).
2. Glukosa kontrola: COVID-19 ikerketak oinarri hartuta...
3. Kortikoesteroideak: Meta-analisi berrienek (Martínez-García et al., 2023)...

RAG teknologia gabeko erantzunak oinarritzko ezagutza erakusten du, baina ez du informazio eguneraturik ematen eta ez du azken ikerketen erreferentzia zehatzik eskaintzen. Beste erantzunean, ordea, informazio eguneratua eta espezifikoa dago, eta gida zehatzen eta fidagarrien erreferentzia zuzenak egiten dira.

Etorkizun hurbilean, RAG sistemak gero eta integratuagoak, azkarragoak eta zehatzasun handiagoak izango dira, eta erabiltzaileen, enpresen edo erakundeen behar espezifikoetara egokitutako soluzio pertsonalizatuak eskainiko dituzte. •

