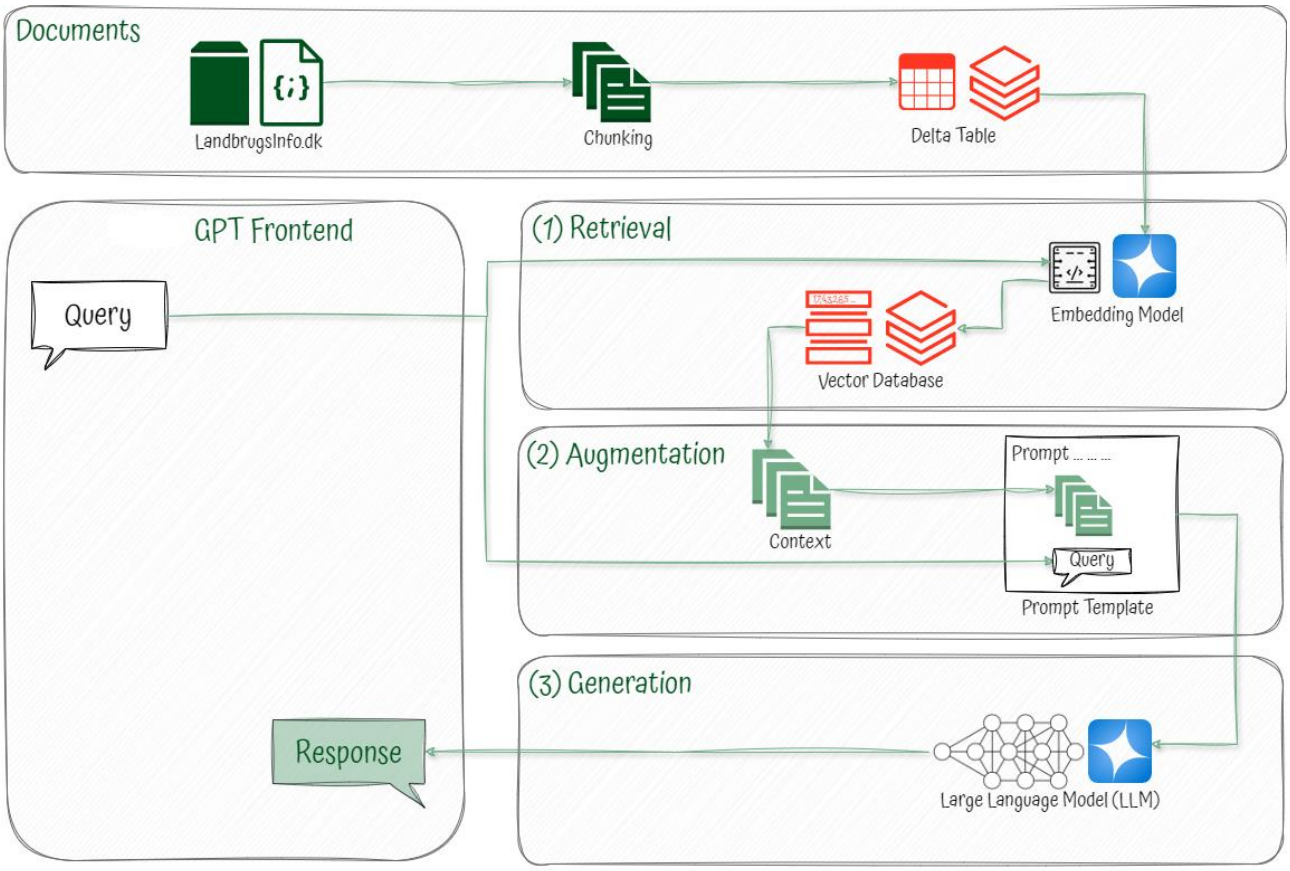
Arbejde med generativ kunstig intelligens

Arbejdet i 2024 projektet *104208 Så ny viden i marken* har fokuseret på udviklingen af en avanceret GenAI-arkitektur, der kan udnytte alt fra nyhedsindhold til Landsforsøgene på Landbrugsinfo.dk med henblik på at levere præcise og relevante svar på brugerforespørgsler. Arkitekturen er baseret på Retrieval Augmentation Generation (RAG) og er bygget op omkring de tre nøgleprocesser om at genfinde dokumenter, berige en prompt skabelon og genere et tilfredsstillende svar. Disse RAG-processer skal medføre genereringen af en mere ideelt svar som en rå GenAI model ikke ville kunne genere uden adgang til dataene bag en betalingsmur, som eksempelvis Landbrugsinfo.dk, der skulle indeholde svaret på en konkret faglig forespørgelse.



Figur 1 Løsningsarkitektur for inddragelse af data fra landbrugsinfo.dk

Løsningsarkitekturen kombinerer således moderne AI-teknologi med domænespecifik viden fra landbruget og fungerer som et værktøj, der kan hjælpe professionelle med at finde og anvende relevant information hurtigt og effektivt.

Først behandles dataene fra Landbrugsinfo.dk gennem en proces, hvor større dokumenter opdeles i mindre, håndterbare stykker ved hjælp af *chunking*. Disse mindre datastykker gemmes derefter i en organiseret datakilde kaldet en Delta Table. Dette gør det muligt at opbevare og genfinde data struktureret, hvilket er fundamentalt for resten af løsningens processer.

Når en bruger indsender en forespørgsel via systemets frontend, aktiveres genfindingsprocessen. Her anvender løsningen en AI embedding-model til at repræsentere både forespørgslen og de tilgængelige dokumenter som numeriske vektorer. Disse vektorer lagres i en specialiseret vektordatabase, som gør det muligt hurtigt og præcist at identificere dokumenter med høj similaritet til og relevans for forespørgslen.

Efter genfindingsfasen går løsningen videre til at berige forespørgslen. De identificerede dokumenter udtrækkes og bruges til at skabe en kontekst, som er essentiel for at sikre, at systemet forstår og svarer korrekt på brugerens spørgsmål ud fra denne. Denne kontekst integreres i en såkaldt Prompt Template, eller prompt-skabelon, som er en struktureret skabelon, der anvendes til at formulere en fuldstændig forespørgsel til en Large Language Model (LLM). På denne måde sikrer løsningen, at den relevante information bliver korrekt indarbejdet i den efterfølgende genereringsproces.

Den sidste fase indebærer, at en LLM (fx gpt-4o eller gpt-4o mini) genererer et svar baseret på både brugerens oprindelige spørgsmål og den tilføjede kontekst fra de relevante dokumenter. Svaret leveres derefter tilbage til brugeren via frontenden, hvilket afslutter interaktionen.

Effektvurdering

Selvom løsningen viser styrke i sin evne til at finde korrekte og relevante kilder, er der rapporteret udfordringer med kvaliteten af de genererede svar. Udover at løsningsarkitekturen er blevet præsenteret og demonstreret over tre dage til interesserede på Agromek 2024, så har vi haft omkring 20 eksterne og interne konsulenter og fagpersoner til at teste løsningen igennem november og december. De indledende testresultater herfra viser**, at løsningen næsten altid returnerer de rigtige kildehenvisninger**, men at **de genererede svar kun er fuldt tilfredsstillende i omtrent 50 % af forespørgslerne**. Dette peger på, at genfindings- og berigelsesprocesserne fungerer godt, mens svargenereringen har behov for yderligere optimering. Alligevel har løsningen et stort potentiale som søgeværktøj og yderligere som værktøj der kan forbedre beslutningsstøtte inden for landbruget, og med yderligere udvikling kan den blive endnu mere pålidelig og effektiv.

En anden efterspurgt effekt som løsningen på nuværende tidspunkt mangler, er egenskaben til at søge efter informationer på tværs af flere dokumenter (fx dokumenter gennem en

årrække) og sammenslutte disse informationer i et tilfredsstillende svar. Denne evne er teknisk set mulig ved en udvidelse af løsningen, så denne manglende effekt er en konsekvens af ikke at være indeholdt i en standard RAG arkitektur, og kræver en mere advanceret *agentisk arkitektur* med en central *planning agent,* som bl.a. koordinerer genfindingsfasen til at f.eks. søge efter dokumenter på tværs af en årrække.

Løsningen er desuden designet med en lognings- og gemmefunktion, der sikrer, at alle brugerforespørgsler, genererede svar og tilhørende kontekster samt kildehenvisninger logges og gemmes. Disse data er tilgængelige for yderligere kvalitativ og kvantitativ evaluering, hvilket gør det muligt at analysere løsningens ydeevne og identificere potentielle forbedringsområder. Ved at logge hver forespørgsel kan vi få indsigt i, hvilke typer spørgsmål brugerne stiller, og hvordan løsningen håndterer forskellige forespørgsler. Denne logning bidrager ikke kun til løbende forbedring af løsningens funktionalitet, men sikrer også gennemsigtighed og sporbarhed i svarprocessen. Det understøtter en iterativ udviklingsproces, der løbende hæver kvaliteten og robustheden af de genererede resultater.