INSTITUTIONEN FÖR DATAVETENSKAP | LUNDS TEKNISKA HÖGSKOLA | PRESENTERAD 2025-02-19

EXAMENSARBETE Retrieval-Augmented Generation for Technical Question Answering
STUDENT Victor Tiet
HANDLEDARE Marcus Klang (LTH)
EXAMINATOR Jacek Malec (LTH)

Besvarning av tekniska frågor med hjälp av språkmodeller

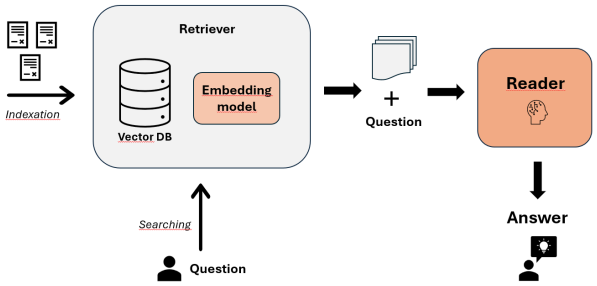
POPULÄRVETENSKAPLIG SAMMANFATTNING **Victor Tiet**

Många företag har idag samlat stora mängder intern dokumentation om sina tekniska system. Tänk om man kunde skapa en intern frågeassistent som, med tillgång till denna dokumentation, kan svara på anställdas frågor på ett automatisk sätt? Detta examensarbete kommer att utforska just den möjligheten.

I takt med att stora språkmodeller har blivit allt populärare har även tjänster kopplade till dem ökat. Ett exempel på ett populärt frågesvar system idag är ChatGPT. I många fall så har dem underliggande modellerna tränats på stora mängder offentlig data, vilket har gett dem breda kunskaper. Men hur får man en språkmodell att svara på frågor som handlar om specifik information? På information som kanske inte finns tillgängligt offentligt eller till och med innehåller känsliga företagshemligheter som inte kan delas med?

Många företag hanterar stora mängder intern data i form av dokumentation om sina tekniska system och produkter. För anställda kan det ibland vara både svårt och tidskrävande att hitta rätt information. Tänk om man kunde skapa en intern frågeassistent som automatiskt hämtar relevant information och sedan genererar ett svar?

I mitt examensarbete utforskar jag just detta. Vi visar att detta kan uppnås genom att implementera en *retriever*, som hämtar relevant information och sedan skickar den vidare till språkmodellen (här kallad *reader*). Enkelt uttryckt: om språkmodellen är en student som behöver information för att skriva en uppsats, fungerar retrievern som bibliotekarien som hjälper till att hitta rätt material.



Vårt arbete undersöker hur man bygger den optimala RAG-modellen för teknisk dokumentation genom att testa både olika språkmodeller och retrievers. Vi undersöker även hur datans domän kan påverka prestanda. För att göra detta genomför vi experiment på både teknisk dokumentation och populärvetenskapliga texter. Bland de modeller vi testar kommer vi fram till att Phi-3 är den bästa språkmodellen för den tekniska domänen, medan Jina-Embeddings-v3 är den bäst presterande retrievern inom samma område. Vi identifierar också vissa prestandaskillnader mellan domänerna, men drar slutsatsen att ytterligare forskning krävs för att ge ett tydligare svar på domänens inverkan på prestandan.