

EXAMENSARBETE Automating Knowledge Graph Construction with LLMs and Comparing RAG against GraphRAG for Enterprise Question Answering

STUDENTER Ebba Rakuljic Feldt, Elin Hellström

HANDLEDARE Marcus Klang (LTH)

EXAMINATOR Jacek Malec (LTH)

Grafer som hjälper AI förstå intern data

POPULÄRVETENSKAPLIG SAMMANFATTNING **Ebba Rakuljic Feldt, Elin Hellström**

Många företag sitter idag på stora mängder intern information och data som kan vara svår att överblicka. AI-assistenter och chatbotar används i allt större utsträckning för att besvara frågor kring intern företagsdata. Detta ställer högre krav på systemen att förstå relationer i datan, utan att hitta på information eller förvilla användaren. I detta examensarbete undersöker vi om AI kan ge bättre och mer korrekta svar genom att använda kunskap som organiserats i kunskapsgrafer.

Stora språkmodeller, såsom ChatGPT, är generalister som tränas på stora mängder offentlig textdata från nätet. Modellen har bara kunskap som fanns i dess träningsdata, och riskerar därmed att snabbt bli inaktuell. Vill man att modellerna även ska använda extern data, som privat företagsdata eller domänspecifik information, måste man utöka modellens kunskapsbas. Traditionellt görs detta genom att träna om modellen på ny data, men detta är en resurskrävande process. De mest framstående modellerna idag är dessutom stängda modeller som inte går att bygga vidare på. När en språkmodell saknar kunskap riskerar den dessutom att hitta på svar, ett fenomen som kallas hallucinationer. En lösning på dessa problem är RAG (Retrieval-Augmented Generation). Modellen får relevant kontext innan den svarar, vilket utökar dess kunskapsbas och minskar hallucinationer. RAG kan dock ha svårt för mer komplexa eller sammanfattande frågor som kräver resonemang och en djupare förståelse för datan.

I detta arbete undersöker vi därför en variant av RAG, där data organiseras i en graf med noder

(t.ex. personer eller koncept) och relationer mellan dessa. Vidare jämför vi grafbaserad RAG mot traditionell RAG. Vi undersöker också tillförlitligheten i att syntetiskt generera testdataset och i att använda språkmodeller för utvärdering, tekniker vi tror kan påskynda införandet av RAG-system inom industrin, där referensdata sällan existerar på förhand.

Våra resultat visar att traditionell RAG fortfarande presterar bättre än grafbaserad RAG på fråge-svarsdata, samtidigt som den är enklare och mer kostnadseffektiv att sätta i drift. Samtidigt utesluter resultaten inte att grafbaserad RAG kan vara användbar för vissa typer av frågor eller att grafer har andra fördelar, till exempel för att visualisera data. Vi bedömer att kunskapsgrafer är lovande för RAG, men att det krävs mer forskning och innovation för att se om dessa system kan överträffa traditionell RAG. Vi konstaterar också att både syntetisk datagenerering och språkmodeller är lovande vid utvärdering av AI-system där annoterad valideringsdata saknas.