EXAMENSARBETE An On-Premise Diffusion Deep Research Model & Evaluation Framework for Resource Constrained Environments
STUDENTER Arvid Malm, Wolmar Boris-Möller
HANDLEDARE Patrik Edén (LU), Thomas Drakengren (Qrendo AB)
EXAMINATOR Pierre Nugues (LTH)

AI-baserad automatisk rapportskrivning med källhänvisning, helt utan internet

POPULÄRVETENSKAPLIG SAMMANFATTNING **Arvid Malm, Wolmar Boris-Möller**

Utvecklingen av rapportskrivande AI-agenter, så kallade Deep Research modeller, går i en rasande takt. Alla dessa förlitar sig däremot på moln-lösningar, vilket inte passar datakänsliga organisationer. Därför har vi tagit fram en Deep Research modell (DORA), samt ett ramverk för utvärdering av rapporter (TLDR), som kan köras helt lokalt.

Under de senaste åren har vi sett stor utveckling av AI, accelererad av nya språkmodeller, LLMer. Dessa LLMer kan också kombineras med kringliggande arkitekturer vilket ibland kallas agentisk AI. En sådant exempel är Deep Research, där en LLM-baserad agent skriver en lång rapport med källhänvisning för källor den själv sökt fram.

De lanserade modellerna är däremot uteslutande baserade på användning av molntjänster, vilket inte är acceptabelt för organisationer med höga krav på datasäkerhet. Ett viktigt exempel på ett sådant användingsområde är för myndigheter som på olika sätt vill analysera och syntetisera information som finns lokalt, tex för att göra systembeskrivningar inför offentliga upphandlingar av känsliga system. Här är det såklart otroligt viktigt att inte överlämna känslig information till datacenter eller molnleverantörer av beräkningskraft som inte ägs av myndigheterna.

För att möta detta behov har vi byggt en Deep Research model, kallad DORA, som kan köras helt lokalt. Arkitekturen är inspirerad av en modelltyp som föreslogs av forskare på Google Cloud i slutet av 2025. Idéen är i stort att härma det iterativa sättet forskning bedrivs på i verkligheten, vilket har visats fungera väldigt bra för AI genom Deep Research i tidigare jämförelser.

Bedömningen av vad en "bra" rapport är visade sig vara komplicerat. Helst har man en expert som bedömer text utifrån vissa kriterier, men skalbarheten för detta är låg. Därmed byggde vi ett automatiskt evalueringsramverk, kallat TLDR, som uttnyttjar LLMer som domare samtidigt som den kan verifiera fakta med hjälp av lokala dokument. TLDR bedömer rapporter på korrekthet samt fyra olika kriterier inom textkvalitet och levererar ett betyg mellan 0 och 100.

Resultaten visar att det fungerar att köra Deep Research lokalt, så länge lämpliga arkitekturval görs. DORA presterar också bättre än alla testade molnbaserade alternativ på evalueringsramverket TLDR med dataset inom kravhantering.

Figuren nedan visar hur våra föreslagna modeller DORA och TLDR strukturellt fungerar.