INSTITUTIONEN FÖR DATAVETENSKAP | LUNDS TEKNISKA HÖGSKOLA | PRESENTERAD 2026-06-28

EXAMENSARBETE A Comparative Evaluation of Hallucination Detection Methods for RAG Systems
STUDENTER David Larsson, Yousif Kutaiba
HANDLEDARE Marcus Klang (LTH), Samir Jasarevic (Robert Bosch AB)
EXAMINATOR Xuan-Son Vu (LTH)

Hur granskar man språkmodellens svar?

POPULÄRVETENSKAPLIG SAMMANFATTNING **David Larsson, Yousif Kutaiba**

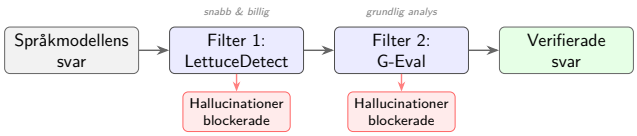
Stora språkmodeller kan ge övertygande men ogrundade svar. Detta arbete jämför fem metoder för att fånga sådana fel och visar att en kombinerad lösning överträffar varje metod på egen hand.

Verktyg som ChatGPT, som bygger på språkmodeller, kan framstå som pålitliga men riskerar att producera felaktiga eller påhittade uppgifter i sina svar. Detta kallas inom forskningen för hallucinationer. En vanlig lösning är Retrieval-Augmented Generation (RAG), där modellen får tillgång till externa dokument att basera sina svar på. Men även med rätt information kan modellen producera svar som inte stöds av dokumenten, vilket gör det viktigt att kunna upptäcka sådana fel automatiskt.

I detta examensarbete jämförs fem metoder för att upptäcka hallucinationer i RAG-system. Två av dem, RAGAS och G-Eval, använder en stor språkmodell som domare för att bedöma om ett svar är troget sitt källmaterial, och testas med fyra olika underliggande modeller. De övriga tre, LettuceDetect, HHEM-2.1 och Lynx, är specialbyggda klassificerare tränade för att identifiera hallucinationer.

Samtliga metoder utvärderades på RAGTruth, ett etablerat referensdataset, över 404 noggrant balanserade exempel. Metoderna jämfördes i träffsäkerhet, robusthet över uppgiftstyper och svarslängder, samt körtid och kostnad. G-Eval med GPT-5.2 uppnådde högst övergripande prestanda och var mest stabil över olika uppgiftstyper. Den specialiserade klassificeraren LettuceDetect följde nära efter, och var samtidigt avsevärt snabbare och billigare. RAGAS visade sig vara långsammast och dyrast utan motsvarande prestandavinst. Ett intressant fynd var att valet av metod spelade minst lika stor roll som valet av underliggande modell.

De två bäst presterande metoderna kombinerades till en tvåstegspipeline. Först filtrerar LettuceDetect bort tydliga hallucinationer till försumbar kostnad, varefter återstående svar skickas till G-Eval med GPT-5.2 för en grundligare analys. Pipelinen utvärderades på ett egenproducerat dataset med 91 exempel inom cybersäkerhetsreglering, framtaget i samarbete med Robert Bosch AB. Pipelinen fångade fler hallucinationer än någon av metoderna var för sig, vilket tyder på att de kompletterar varandra: LettuceDetect upptäcker tillagd information som saknar stöd i källtexten, medan G-Eval bättre fångar subtila förvrängningar.



Arbetet visar att automatisk hallucinationsdetektering är möjlig, men att inget enskilt verktyg löser problemet helt. Genom att kombinera en snabb och billig klassificerare med en mer kapabel men dyrare bedömare kan man uppnå en rimlig balans mellan kostnad och tillförlitlighet.