

Undersøgelse om vektor search, RAG og LLM kan bruges til at undersøge SQL server logs

Steen Ostensen Januar 2026

IT-Master AUC Anvendt AI

Indholdsfortegnelse

1 Baggrund/Formål.....	s 4
1.1 er metoden – teknisk / organisatorisk?.....	s 4
1.2 Hvad kan man forvente at få ud af dette projekt – og hvad er alternativerne? (er det det værd).....	s 5
1.3 Hvilke afledte konsekvenser skal vurderes? (etik, intern politik, faglig autonomi, etc.) ..	s 5
1.4 Hvilke kompetencer og ressourcer vil der kræves for at gennemføre dette projekt?.....	s 5
1.5 Hvorfor er det relevant at bruge disse AI-teknologier til at analysere SQL server Logs Hvad er udfordringen i dag - altså en udfoldelse af relevansen for at undersøge hvad AI kan gøre i SQL-arbejdet?.....	s 6
1.6 Kvalitative/ kvantitative metoder.....	s 6
1.7 Bilag.....	s7
2 Valg metode Microsoft /Databricks/Andre.....	s 7
2.1 Sammenligning af Microsoft Fabric og Databrix.....	s 7
2.2 Valg af Microsoft løsning.....	s 8
3 Forbehandling data	s 8
3.1 Data fremstilling.....	s 8
3.2 Data indlæses i evenhouse	s 8
4 Vector search.....	s8
5 RAG.....	s12
5.1 Hvad er RAG.....	s12
5.2 Load SQL logsdata direkte i Chapgpt.....	s13
5.3 Kommercielt system.....	s14
6 Resultater og metode analyse af SQL server log filen.....	s16
6.1 Netværks analyse	s 16
6.1.1 Resultater netværksanalyse	s 17
6.1.2 Sammendrag netværks analyse.....	s 19
6.2 Database sikkerhed.....	s20
6.2.1 Database sikkerhed resultater.....	s20
6.2.2 Database sikkerhed sammendrag.....	s21
7 Diskussion.....	s22
7.1 Hype og mange nye ting.....	s22
7.2 Fejl rate af AI projekter og værdi.....	s24
7.3 Modenhed.....	s25
7.4 Performance mit egen område (IT).....	s26
7.5 Sikkerhed.....	s28
7.5.1 Jailbreaking LLM.....	s28
7.5.2 Promptinjection LLM.....	s29
7.5.3 Datapoiisoing LLM.....	s31

7.5.4 GPDR.....	31
7.5.5 AI Act.....	s31
7.5.6 Sikkerhed sammendrag AI.....	s31
8 AI / Mennesker.....	s32
9 Overvejelser.....	s32
10 Sammendrag og konklusion opgave.....	s33
11 Litteratur	s 34

Bilag A Prompt Rapport

Bilag B Pyspark kode

Bilag C Netværks analyse og sikkerheds analyse på loadet data(RAG)

Bilag D Errors

1 Baggrund/Formål

Hvad er baggrunden for projektet og hvad er det centrale spørgsmål?

Baggrunden af projektet er at undersøge om vektor search, RAG og LLM kan bruges til at undersøge SQL server logs.

Et andet formål er at undersøge hvad der findes af AI dag inden for disse Bigdata produkter

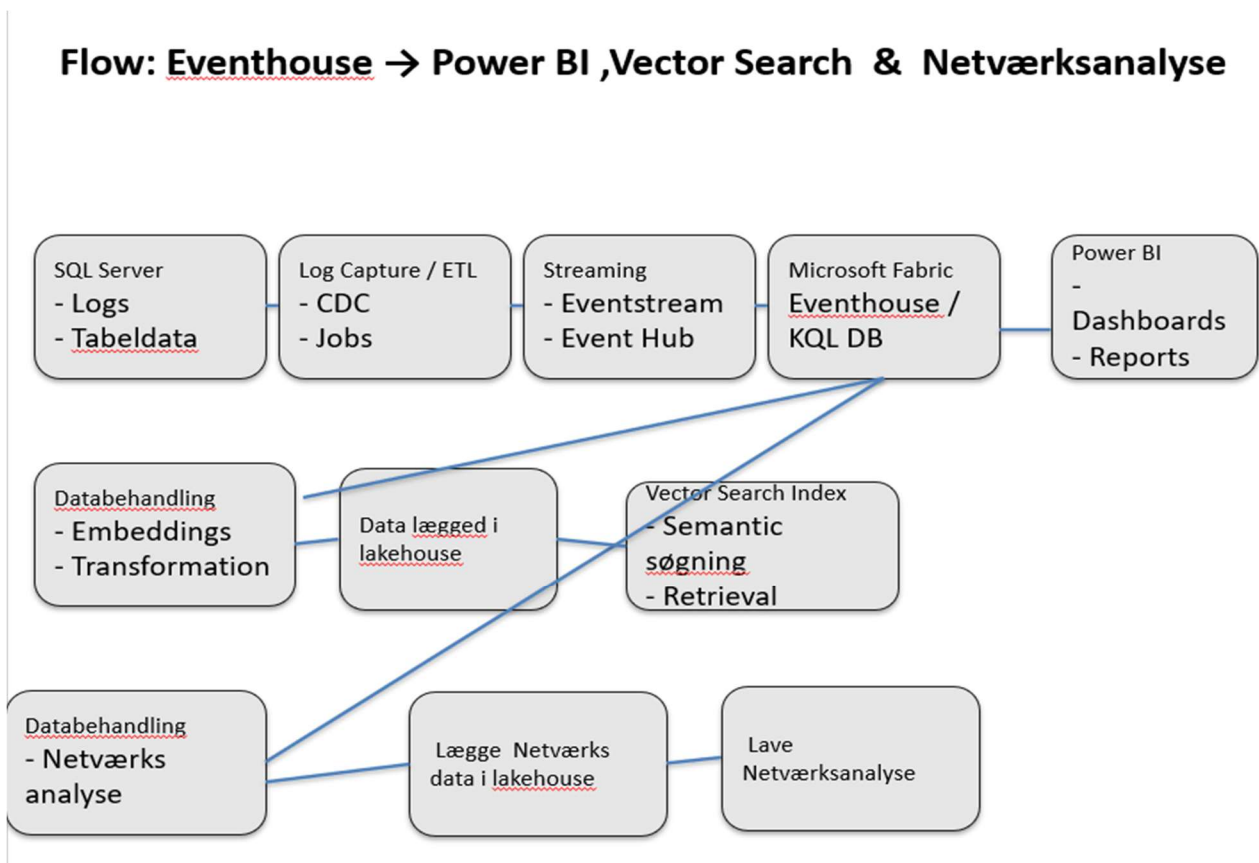
Dette er et indledende projekt som udvides hvis der synes at være en ide idet.

Det er også et formål at forstå valg af teknologi.

1.1 Er metoden – teknisk / organisatorisk?

Metoden er i første omgang teknisk. Der vil senere være en diskussion om hvad man kunne bruge det til i samarbejde med disse som kigger på data i dag. I figur 1 er der vist hvordan et samlet flow kan være. LLM er her brugt til at lave ETL (kode).

Figur 1 Samlet ETL-flow



1.2 Hvad kan man forvente at få ud af dette projekt – og hvad er alternativerne? (er det det værd)

I første omgang er dette et pilotprojekt for at vise værdien af dette. En del virksomheder har allerede streaming af SQL server logs sat op. I gamle dage kunne det være tidskrævende at undersøge SQL server logs og rapportere dette. Fordelen ved at lave dette er at alt kan køre automatisk og man får hurtigt et overblik over hvad der er af issues. AI er her kun brugt som hjælp til lave kode. Det er ligeledes muligt at samle SQL server logs for forskellige SQL servere.

Det vil sige den forventede fordel ved at bruge AI i forbindelse med SQL server logs er man hurtigere lade kode sådan at ETL-flowet kan køre automatisk. Det forventes ligeledes at ved at bruge LLM få hurtigere indsigt i SQL logs data, sådan at DBA hurtigere kan rette fejl i databasen.

Det er dog stadig sådan at DBA/ ledelse træffer det endelige valg (hybrid intelligens).

Dette er kun en indledende step.

Step 2 vil være dialog med ledelse og DBA.

1.3 Hvilke afledte konsekvenser skal vurderes? (etik, intern politik, faglig autonomi, etc.)

I første omgang er der ingen konsekvenser. I denne her opgave er AI anvendt som hjælpeværktøj. Senere kan anvendelse af AI få en del konsekvenser for virksomheden se senere.

1.4 Hvilke kompetencer og ressourcer vil der kræves for at gennemføre dette projekt?

Der kræves tekniske ressourcer med kendskab til Data Engineering og AI. Der kræves senere samarbejde imellem DBA, Data Engineering og Power BI udvikler og ledelse.

AI er kun en del af det samlede projekt.

Det vil sige at der er folk med forskellig baggrund der indgår i projektet.

De kvalitative metoder kommer i spil her. Det er rimeligt vigtigt at forstå betydning for forretning af de forskellige fejl i databasen, sådan at man koncentrerer sig om de vigtigste problemer. I Databricks er RAC og LLM integreret i produktet så der kræves at lære at bruge det.

I Microsoft Fabric kræves der ressourcer som kan hjælpe med at sætte det op og folk der ved hvordan man f.eks. bruger Microsoft Copilot. Det skulle gerne være en teambaseret tilgang og ikke en silobaseret tilgang som findes i nogle organisationer.

Der kræves nok også mange steder at folk bliver uddannet i AI.

1.5 Hvorfor er det relevant at bruge disse AI-teknologier til at analysere SQL server logs. Hvad er udfordringen i dag - altså en udfoldelse af relevansen for at undersøge hvad AI kan gøre i SQL-arbejdet?

Det er relevant at bruge AI til dette da det er hurtigere at lave kode ved hjælp af LLM og koden som indgår i ETL flowet er lavet ved hjælp af AI. Det vil sige produktivitet stiger ved at bruge AI.

Det er ligeledes et stærkt værktøj til informationssøgning og fejl søgning du (LLM som støtte). Mange virksomheder forsøger i dag at optimere den hastighed hvorved man kan lave tingene på. AI kan også anvendes som inspiration.

1.6 Kvalitative/ kvantitative metoder

Denne opgave er lavet sådan at det kan der er en masse tal som man kan diskutere ud fra (kvantitativ).

Kvalitative og kvantitative metoder er noget der gerne skulle supplere hinanden og ikke en form for konkurrence.

Der kunne laves en undersøgelse over hvad brugerne syntes hvad har været de 10 alvorligste fejl i systemet det sidste år(kvalitativt).

Der kunne også gennemgang af de servicedisk sager der har været det sidste år. Disse undersøgelser kunne sammenholdes med denne rapport (sql server log).

Denne opgave er betaget som en diskussions opgave sådan at man sammen med ledelsen og DBA bliver enig om hvad der kunne implementeres. (kvalitativt).

I denne opgave er LLM kun et hjælpeværktøj og kører intet automatisk.

Det er ETL'en som kører automatisk.

DBA / ledelse har ansvar for at træffe de korrekte beslutninger om hvad der gøres i systemet (hybrid intelligens).

1.7 Bilag

Jeg har i bilag A vist de prompt jeg har brugt i opgaven til at søge viden.

I bilag B er den Python kode som er lavet til at lave en del af ETL.

Dette er lavet via Chapgpt.

I bilag C er der vist den prompt som jeg har brugt til at lave Netværks og sikkerhedsanalyse. Den er brugt som inspiration til hvad der er lavet i kode. Prompten er lavet efter at have loadet filen ind i Chatgpt.

Der er i bilag D vist en del af de fejl som jeg har stødt på i opgaven. Det er lettere at løse disse fejl ved hjælp fra ChatGPT.

2 Valg metode Microsoft /Databricks/Andre

Denne afsnit omhandler de valgt som skal tages for at undersøge data med AI. Skal det være open source eller et af de kommercielle datasystemer.

Der sker en rivende udvikling i AI. Jeg er især interesseret i den udvikling der sker i Bigdata produkter som Databricks og Microsoft Fabric.

Der kan vælges en række produkter til dette, men jeg er især interesseret i at bruge det på egne data.

Vector search er muligt i Databricks, Microsoft Azure, PostgreSQL og en række dedikerede vektor-databaser.

Jeg har valgt Microsoft Fabric, pga. økonomi. Jeg har et gratis testsystem. Det er sådan at i Databricks er LLM integreret i produktet, men hos Microsoft er det en del af AI_Azure Foundry .

Valg af løsning vil naturligvis også afhænge af hvilket system man har. Har man Databricks er det mest naturligt at bruge dette.

En anden årsag til valg af metode kunne være man ikke ønsker af være afhængig af Tech giganter

Virksomheden bør have en strategi over hvad de vil, men mange kommercielle virksomheder har data i disse systemer.

2.1 Sammenligning af Microsoft Fabric og Databricks for AI

Jeg har anvendt chatgpt til dette er vist i prompt 2 i bilag A.

Generelt er jeg enig i Chapgpt anbefaling. Den finder også nogle gode kilder.

Anvend Microsoft Fabric hvis du har Microsoft infrastruktur.

Anvend Databricks hvis du har Databricks infrastruktur.

Ryan Shiva (27) anbefaler databricks da den har flere native og bedre GenAI tools. Min personlige valg ville være Databricks, da den bedre understøtter Open source-systemer. Databricks er et produkt som har været på marked i længere tid, imens Microsoft Fabric er et relativt nyt produkt. Microsoft satser massivt på dette produkt.

2.2 Valg Metode Microsoft

Selv inden for Microsoft er der en række valg som skal tages

- 1) **Lakehouse** (datakilde)
- 2) **Evenhouse** (datakilde)
- 3) **Azure Cosmos DB** (datakilde)
- 4) **Copilot Studio** (AI tool klik)
- 5) **AI foundry** (AI tool kode)
- 6) **Data agent Fabric** (Preview)

Jeg har valgt Eventhouse som datakilde da den er velegnet til streaming.

En anden interessant mulighed er **Azure Cosmos DB**, som ifølge Nitish Upreti et al. (8) kan skalere til meget store datamængden.

Der er inden for Microsoft AI sket ret store ændringer i AI, sådan man i dag søger at samle det hele i 1 til 2 produkter. Tidligere var der mange forskellige AI services.

Microsoft Copilot er et klik produkt hvor de kan lave AI produkter, imens AI foundry bruger kode.

Dette betyder at mange af de ældre løsninger kan blive påvirket af at Microsoft har ændret i deres værktøjer. Det kan på et tidspunkt være nødvendigt at flytte løsningerne til de nye produkter.

3 Forbehandling data

3.1 Data fremstilling

Data er lavet via chatgpt 5 som vist i prompt 3.

Det er som set bare bede chatgpt om at lave en SQL server log file med 100000 logs.

3.2 Data indlæses i evenhouse

Data er indlæst manuelt via manuel fil. Dette kan senere sættes op til kører automatisk som allerede vist i figur 1.

4 Vektor search

Dette afsnit handler om hvordan vektor search kan bruges på log data.

Vector search kan også være en del af et kommercielt RAG og LLM-system som kan anvendes til at analysere store mængder tekst. I SQL serverlogs er der både tekst data og numeriske data.

Jeg har valgt at lave embeddings og vektor index på 4 tekstfelter Event_Type, Database_Name, Sql_Text, Error_Message. Der er lavet et samlet index.

Jeg har spurgt Chatgpt om dette se prompt 4.

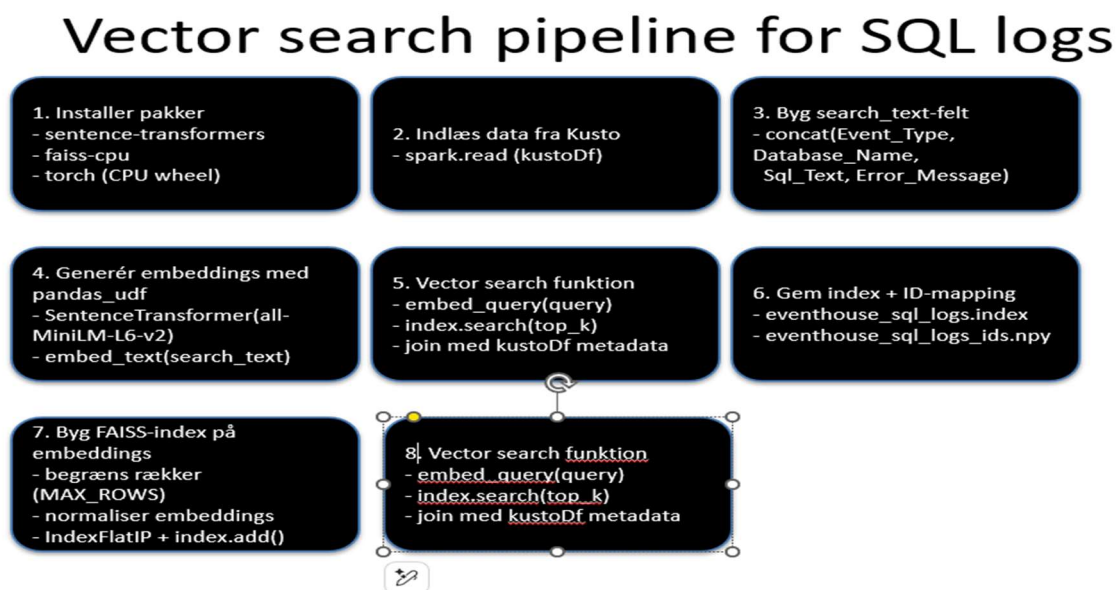
. Ved små datamængder anbefales et kombineret index ved større dataset anbefales separate dataset. Dette bekræftes af Hamed Zamani et al (9). Jeg har kun et mindre datasæt her. Generelt skulle separate index give bedre søgninger.

Jeg har ligeledes valgt at lave indexet i Pyspark og gemme den i et Lakehouse. Der kunne også laves et vektor index på kol-databasen.

I bilag Pyspark kode bilag B 1 er der vist den Pyspark som er anvendt til at lave vektor search. Meget af koden er lavet via Chatgpt 5.

I figur 2 er der vist vigtigste step i koden.

Figur 2 Vektor search steps(LLM genereret)



Jeg har lavet 3 eksempler på vektor search .Se figur 3,4 og 5.

Figur 3 vektor search deadlock kun vist de 5 første rekords.

	score	Timestamp	Level	Event_Type	Database_Name	Sql_Text	Error_Message
0	0.656322	2025-08-07 07:27:57.290	ERROR	Deadlock	db_hr	/* long-running */ SELECT * FROM stage.employe...	Transaction (Process ID 4541) was deadlocked o...
4	0.643324	2025-09-30 22:43:56.290	ERROR	Deadlock	db_finance	/* long-running */ SELECT * FROM stage.orders ...	Transaction (Process ID 1820) was deadlocked o...
3	0.640071	2025-09-26 10:36:00.290	ERROR	Deadlock	db_marketing	/* long-running */ SELECT * FROM stage.orders ...	Transaction (Process ID 4620) was deadlocked o...
2	0.639036	2025-09-25 01:00:08.290	ERROR	Deadlock	db_finance	/* long-running */ SELECT * FROM stage.departm...	Transaction (Process ID 3651) was deadlocked o...
1	0.638322	2025-08-18 03:13:48.290	ERROR	Deadlock	db_ops	/* long-running */ SELECT * FROM stage.orders ...	Transaction (Process ID 3072) was deadlocked o...

Figur 4 vektor search deadlock finance kun vist de 5 første rekords

	score	Timestamp	Level	Event_Type	Database_Name	Sql_Text	Error_Message
1	0.661595	2025-09-06 10:36:22.290	ERROR	Deadlock	db_finance	/* long-running */ SELECT * FROM ref.employees...	Transaction (Process ID 4527) was deadlocked o...
2	0.661079	2025-09-07 19:02:27.290	ERROR	Error	db_finance	MERGE ref.orders AS t USING (VALUES (?)) AS s(...	Transaction (Process ID 531) was deadlocked on...
3	0.651581	2025-09-20 03:22:57.290	ERROR	Login	db_finance	MERGE hr.order_items AS t USING (VALUES (?)) A...	Transaction (Process ID 1173) was deadlocked o...
0	0.649321	2025-08-08 11:21:04.290	ERROR	Error	db_finance	MERGE inv.payments AS t USING (VALUES (?)) AS ...	Transaction (Process ID 1868) was deadlocked o...
4	0.648236	2025-09-22 12:56:39.290	ERROR	Deadlock	db_finance	/* long-running */ SELECT * FROM ref.products ...	Transaction (Process ID 210) was deadlocked on...

Hybrid search

Figur 5 vektor search deadlock finance hvor data er filteret på "Database_Name"=="db_finance" kun vist de 5 første records

Figur 5 Hybrid search

	score	Timestamp	Level	Event_Type	Database_Name	Sql_Text	Error_Message
59	0.695375	2025-09-06 10:36:22.290	ERROR	Deadlock	db_finance	/* long-running */ SELECT * FROM ref.employees...	Transaction (Process ID 4527) was deadlocked o...
87	0.682802	2025-09-22 12:56:39.290	ERROR	Deadlock	db_finance	/* long-running */ SELECT * FROM ref.products ...	Transaction (Process ID 210) was deadlocked on...
9	0.680232	2025-08-11 21:04:12.290	ERROR	Deadlock	db_finance	/* long-running */ SELECT * FROM stage.payment...	Transaction (Process ID 4122) was deadlocked o...
24	0.678746	2025-08-17 10:51:36.290	ERROR	Deadlock	db_finance	/* long-running */ SELECT * FROM hr.department...	Transaction (Process ID 566) was deadlocked on...
97	0.678314	2025-09-30 22:43:56.290	ERROR	Deadlock	db_finance	/* long-running */ SELECT * FROM stage.orders ...	Transaction (Process ID 1820) was deadlocked o...
...	...	...	...	...	...	...	...
60	0.637648	2025-09-07 04:21:11.290	ERROR	Deadlock	db_finance	/* long-running */ SELECT * FROM dbo.suppliers...	Transaction (Process ID 266) was deadlocked on...
41	0.637272	2025-08-28 01:00:48.290	ERROR	Login	db_finance	SELECT amount, status, name, id, dept_id FROM ...	Transaction (Process ID 3577) was deadlocked o...

I figur 3 med kun deadlock finder alle med eventype Deadlock som passer bedst. En sql med filter på eventype = deadlock vil give nogenlunde det samme resultat, men ikke det samme.

I figur 4 hvor der søges på deadlock og finance giver de 5 vigtigste som alle database_name = db_finance ,men den finder også eventype som ikke deadlock men der er deadlocked i SQL-tekst.

Dette svarer også til virkeligheden i en deadlock er der involveret 2 eller mere sessioner (Den der bliver killed og den som bliver blokeret).

Den tredje eksempel figur 5 er et hybrid search her er der kun vist de vector_search som db_finance som database_name.

Fordelen ved vector seach

Jeg har spurgt Chagpt om dette se prompt 5 Resultatet er vist i figur 6

Figur 6 Fordel med vector search (chapgpt)



Fordel

Forklaring

- Semantisk forståelse
- Hurtighed
- Skalerbarhed
- Fleksibilitet
- Integration
- Finder relevante data, selv med forskellige ordvalg
- Søger i millioner af rækker på millisekunder
- Du behøver ikke loade alt i LLM'en
- Kombinér med SQL-filtre
- Grundlaget for RAG og agent-baserede LLM-systemer



Dette svarer også til hvad Evan Castle (29) siger.

Sammendrag

Vektor search er en spændende måde at søge på og kan anvendes sammen med almindelig søge metoder til at finde information. Funktionen er relativt nyt det er lige kommet eller er i preview i en række bigdata produkter.

Den er især egnet til tekst eller billeder. Vektor search på numeriske værdier er den ikke egnet til.

Man kan evt. lave alle værdier om til tekst. Dette er dog ikke gjort i denne opgave, men kunne overvejes.

5 RAG

Denne afsnit omhandler RAG da, det kan være en del af et kommercielt system sammen med vektor search og LLM.

5.1 Hvad er RAG

RAG står for Retrieval(hentning) Augmentation (berigelse) Generation:

Dette er vist i figur 7.

Figur 7 hvad er RAG (chapgpt)

1. Retrieval (hentning):

Systemet søger i en ekstern vidensbase → fx dokumenter, databaser, API-data, interne PDF'er osv.

2. Augmentation (berigelse):

De relevante tekststykker, der findes, føjes til prompten.

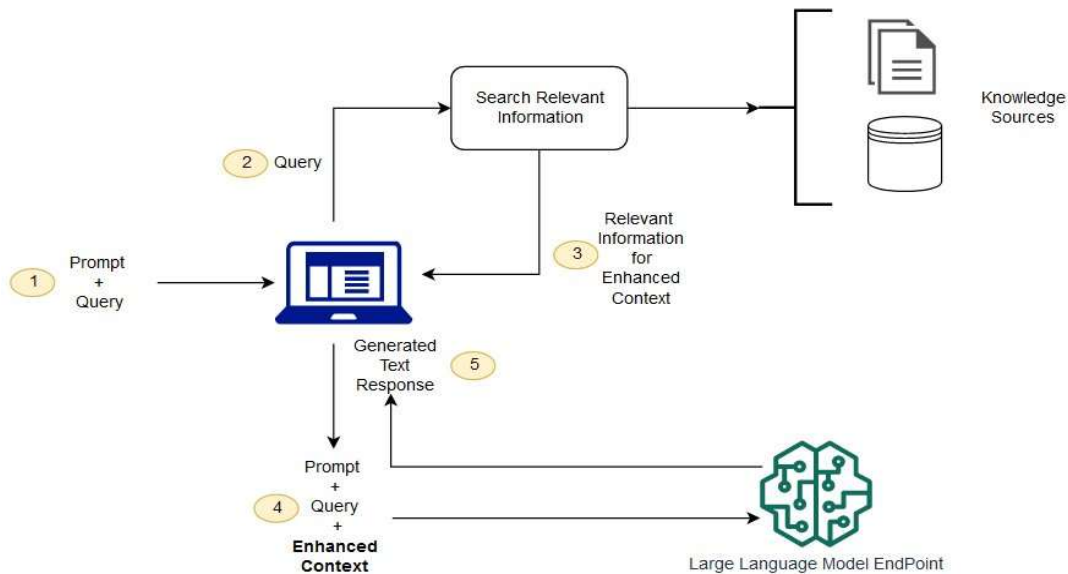
3. Generation:

Sprogmodellen genererer svaret baseret på både din prompt og de hentede dokumenter.

” Spørg ChatGPT

I figur 8 er det vist i en figur.

Figur 8 RAG (retrieval augmented generation) Ivan Belcic (47)



Fordelen ved RAG skulle Ivan Belcic (47) være

- Omkostningseffektiv AI-implementering og skalering
- Adgang til aktuelle, domænespecifikke data
- Lavere risiko for AI-hallucinationer
- Øget brugertillid
- Udvidede anvendelsesmuligheder
- Forbedret udviklerkontrol og modelvedligeholdelse
- Højere datasikkerhed

5.2 Load SQL logsdata direkte i Chapgpt

Jeg har loadet data direkte ind i chapgpt sådan at jeg kan bruge chapgpt til at spørge på data.

Jeg har i bilag C Netværks analyse og sikkerheds analyse på loadet data (RAG) vist prompten.

Beregningerne fundet svarer fint til når jeg selv laver beregningerne i python.

Der er dog 2 problemer ved at gøre sådan.

Den er ved at gøre dette er at datamængden kan blive for stort til kontekst størrelse.

Jeg har spurgt Chapgpt om det er muligt at load en hel tabel ind i den se prompt 6.

Prompten anbefaler at bruge LLM plus SQL, men der er ingen problemer på små tabeller.

I Chapgpt plus variant er content vindue 32000 ifølge [Shivam Malani\(43\)](#).

En anden er at det er en sikkerhedsrisiko at sende disse data til Chapgpt, men kunne evt. installere LLM f.eks. ollama lokalt på egen server.

I dag er LLM beregnet til at håndtere tekst, men ikke så god til numerisk data, men der foregår en del forskning på alle data i en tabel.

F.eks. Liangyu Zha et al(21), og Xiaokang Zhang(22) i litteraturlisten.

Sammendrag

Det er et stærkt værktøj til at undersøge data hurtigt og få inspiration til undersøgelse af ting.

Undersøgelser på den valgte fil har ikke givet nogen problemer.

Der er dog det problem at kontekst vindue kan blive for lille.

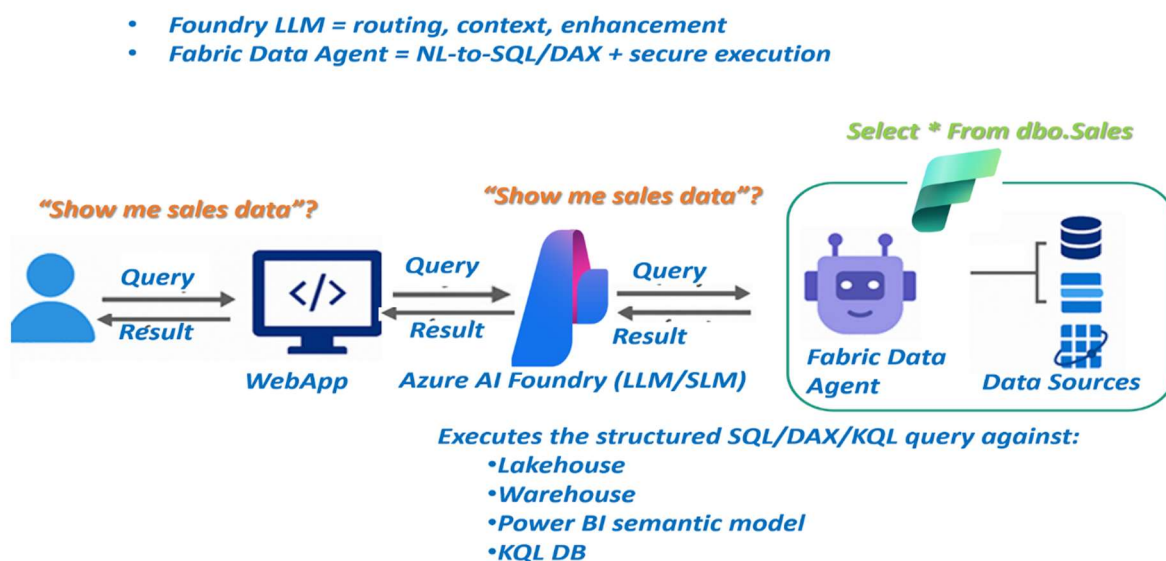
Der er også et sikkerhedsproblem, da det ikke er sikkert at virksomheder vil have data sendt til chaptgpt.

5.3 Kommercielt system

Jeg har brugt chatgpt til at tegne hvordan et kommercielt system kan se ud.

Dette er vist i figur 9.

Figur 9 Kommercielt system (prompt generet se prompt 14)



Aditi Singh. (12) viser, at den moderne forskning bevæger sig fra at forsøge at “indlæse” hele tabeller til at udvikle intelligente del-indlæsninger og agentbaserede systemer, hvor LLM’en dynamisk henter kun de nødvendige data. Dvs. de bruger LLM til at lave SQL en

Dette er netop også hvad der sker i Data agent i fabrik.

Data agent Fabric

Data agent indeholder en LLM og RAG og kan bruges til at prompte dine data i Microsoft Fabric (alpa buddhabhatti 104). Se også prompt 14. Der er dog ikke vektor search i Data agent.

Data agent har ifølge Microsoft (92) 3 tekst oversætter

- Natural language to SQL (NL2SQL) for relational databases (Lakehouse/Warehouse).
- Natural language to DAX (NL2DAX) for Power BI datasets (Semantic Models).
- Natural language to KQL (NL2KQL) for KQL databases.

Der er kun læse adgang til lakehouse/Warehouse.

Jeg har forsøgt at lave en Microsoft konto som forslået af Microsoft (30), for ellers vil Data agent Fabric ikke virke.

Data agent Fabric virker desværre ikke på en fri trail

Login personal account

The client 'fabrick2@stostoutlook.onmicrosoft.com' with object id '9c286e3c-be74-434d-9372-c1b4d4656bd4' does not have authorization to perform action 'Microsoft.Resources/subscriptions/providers/read' over scope '/subscriptions/2b58b022-3cd2-42df-a65d-3e7e2ddb0e01' or the scope is invalid. If access was recently granted, please refresh your credentials.

Det fejler også når jeg bruger min personlig konto.

Jeg har forsøgt på forskellige måder, men har desværre ikke fået det til at virke.

Jeg skal antagelig sætte en azure konto op med en virksomheds e-mailadresse for at få det til at virke.

Dette har jeg sprunget over.

6 Resultater og metode analyse af SQL server log filen

Jeg har valgt at lave netværks analyse og sikkerheds analyse på disse SQL logs filer.

Der kunne også laves performance rapport (prompt 3 Bilag C Netværks analyse og sikkerheds analyse på loadet data (RAG)).

Jeg har brugt LLM som inspiration til undersøgelse af disse data efter at have loadet filen ind.

Den lavede Python kode er lavet og kørt i Microsoft Fabric og der er disse resultater der vises.

Det vil sige at alt koden er kørt i Notebooks i Microsoft Fabric.

6.1 Netværk analyse

Alt python kode er lavet via Chapgpt.

Det er nemt og meget hurtigt at få overblik over data.

Den laver f.eks.

host_name ⇌ database_name

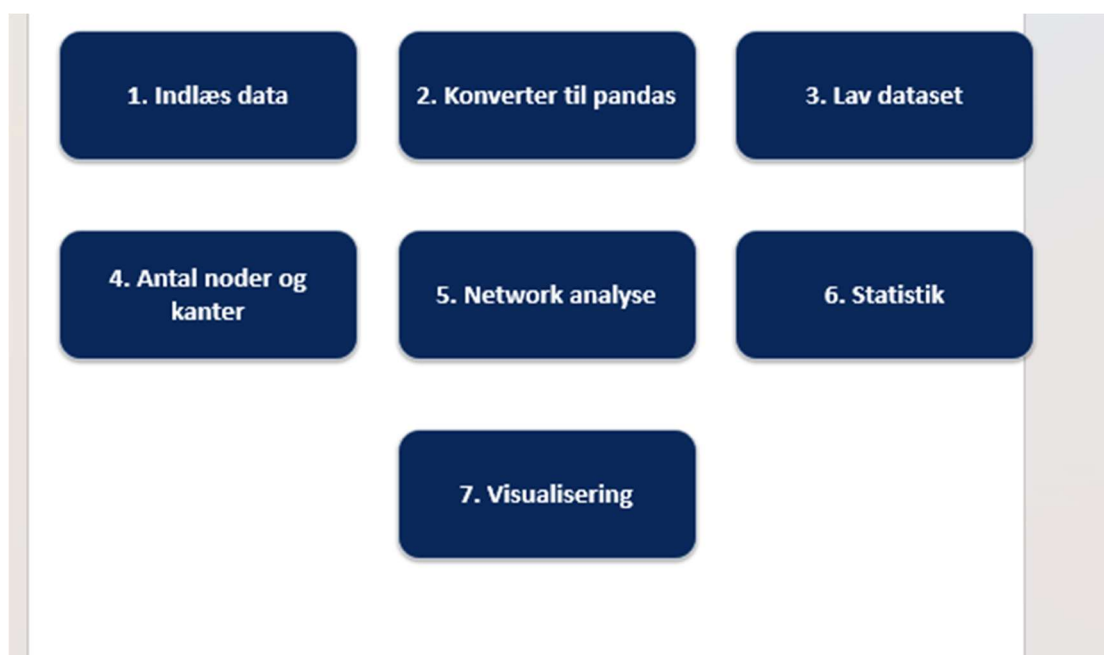
Den kommer også med en del forslag til undersøgelser.

Jeg har valgt at **Netværk: client_ip ⇌ database_name netværket**

Koden er vist i kode i bilag B pythonkode b Netværks analyse.

I figur 10 er der vist princippet i koden.

Figur 10 netværks analyse



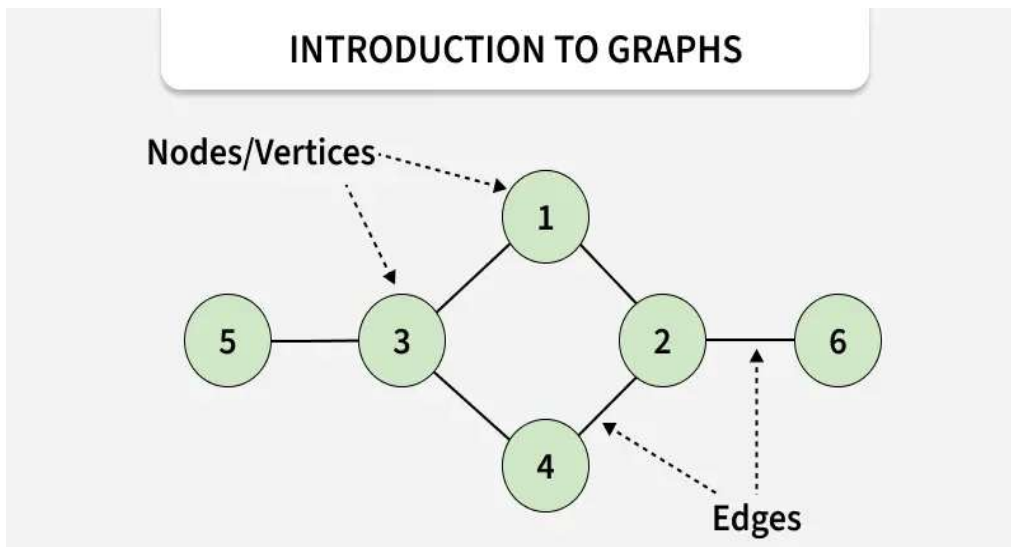
Der bør benyttes GraphFrames pakke (Elena-Simona Apostol et al 10), da dette er beregnet til at anvende på store dataset.

Den kan udnytte multicluster som anvendes i Fabrik og Databricks. Dette gør det kan håndtere meget store tekstmængder.

Jeg har haft problemer med at installere dette GraphFrames i spark clusteret og har derfor valgt at bruge [networkx pakken](#).

I netværksanalyse er der nogle centrale begreber (geeksforgeeks 76) se figur 11.

Figur 11 Forklaring Node og Kant (geeksforgeeks 76)



Degree er den maximale antal der udgår fra Node

Der kan også ses på

user → database (uses_database)

user → host (from_host)

app -> database

6.1.1 Resultater netværksanalyse

Der ses at de 10 vigtigste databaser bliver ramt af hver 480 Client_id.

Figur 12 Degree på databaser (beregnet i python af kode)

```

Top 10 databaser efter degree:
database_name
db_audit      480
db_bi         480
db_finance    480
db_hr         480
db_inventory  480
db_marketing  480
db_ops        480
db_sales      480
db_sandbox    480
db_web        480
Name: degree, dtype: int64

```

I figur 13 er der vist at top 10 klienter alle tilgår 10 databaser.

Figur 13 Degree Ip adresse (beregnet i python af kode)

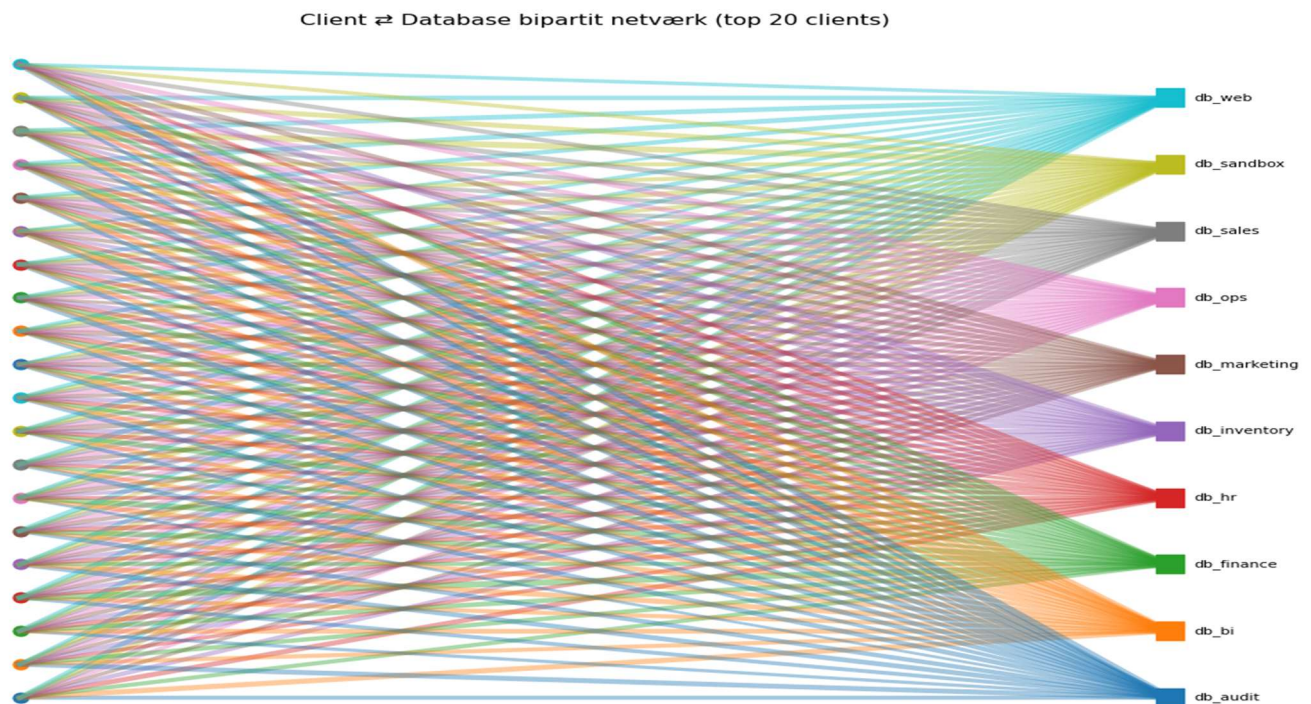
```

Top 10 clients efter degree:
client_ip
10.0.0.1      10
10.0.0.103    10
10.2.32.69    10
10.2.32.52    10
10.2.32.35    10
10.2.32.239   10
10.2.32.222   10
10.2.32.205   10
10.2.32.188   10
10.2.32.18    10
Name: degree, dtype: int64

```

I figur 14 der det vist som plot.

Figur 14 plot bipartit netværk



6.1.2 Sammendrag netværksanalyse

Netværks analyse på sql server logs data er en interessant analyse.

Den kan sige noget om hvordan arkitekturen i databasen er.

Det viste eksempel er dog hæmmet af at det er syntetisk fremstillet data.

Der er ingen databaser hvor alle ip er connecter til alle databaser.

Chapgpt foreslår også at det er syntetiske data.

Det kunne også være en alternativ måde at se på hvem der har eller har haft rettigheder til at tilgå de forskellige databaser.

Det ville være interessant at lave samme analyse på rigtigere sql server logs data.

Det må være muligt at sige nogle om hvor belastet databasen er med hensyn til antal connection der er over tid.

Disse noder og kanter bør gemmes i Evenhouset og så de kan indgå i Power BI reporter.

De vidste data er de samme som hvis jeg loader direkte i Chatgpt.

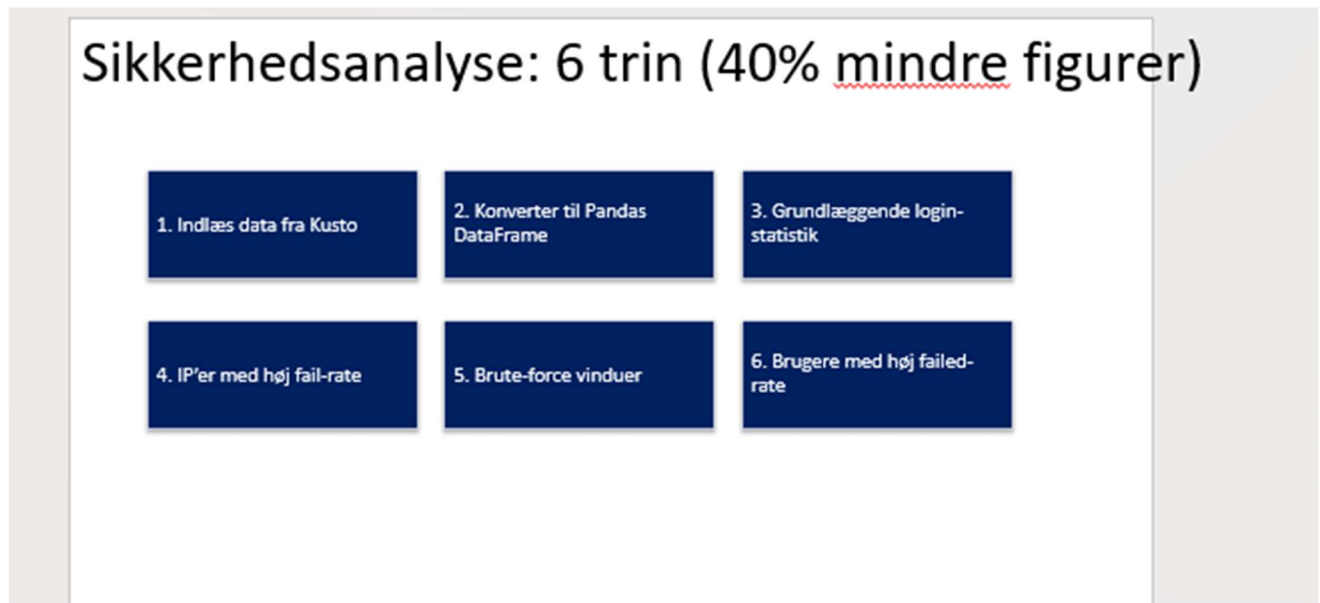
6.2 Database sikkerhed.

Dette er igen lavet ud fra den inspiration som jeg har fået ved at loaden filen i Chatgpt bilag C.

Der er lavet Python kode ved hjælp af chatgpt se Bilag B pyspark kode sikkerhed.

I figur 15 er der vist step i analysen.

Figur 15 Sikkerhedsanalyse(trin i python koden)



6.2.1 Database sikkerhed resultater.

Figur 16 Antal login (beregnet i python af kode).

```

12 Grundbillede for logins
34
Antal login-events: 9,948
Unikke login-brugere: 16
Unikke client IPs (der laver login): 480

```

Figur 17 IP'er med mange fejlede logins og høj fejl-hastighed (beregnet i Python af kode).

```
❌ IP'er med mange failed logins og høj fail-rate
Kriterie: mindst 10 failed logins og fail-rate > 50%
Resultat: 0 IP'er matcher
```

Figur 18 Brute-force vinduer (beregnet i Python af kode)

```
❌ Brute-force-vinduer (failed logins pr. 5 minutter pr. IP)
Kriterie: mindst 5 failed logins i samme 5-minutters vindue
Resultat: ingen IP har ≥ 5 failed logins i samme 5-min vindue
```

Der er i disse data ikke noget som er alarmerende, men IP'er med mange fejlede logins og høj fejlhastighed og Brute-force vinduer er vigtige at følge så man hurtigere kan reagere på dette hvis der skulle opstå problemer.

Dette kunne også sammenholdes hvad virksomheden benytter af teknikker i dag. Det er altid godt at sammenligne forskellige metoder. Fordelen ved at bruge SQL serverlogs er at opsamlingen forgår automatisk.

6.2.2 Database sikkerhed sammendrag

Der kan findes en del interessante ting i disse logfiler. Dette kunne være også være en del af ETL-flowet så alt foregår automatisk.

De vidste resultater er igen hæmmet af de er lavet på syntetisk data.

Der bør diskuteres med de folk som arbejder med dette, så man laver disse beregninger så relevant som muligt.

Disse data kan skrives ned i Eventhuset, sådan at disse data kan ses i f.eks. Power BI.

Alt Python kode er generet via chatgpt.

7 Diskussion

Jeg har valgt at lave en generel diskussion om AI og en mere IT baseret. Jeg har ligeledes valgt at lave en diskussion om maskine /menneske. Jeg har ligeledes valgt en gennemgå sikkerhed, selv om det i første omgang er mindre relevant for dette projekt.

7.1 Hype og mange nye ting

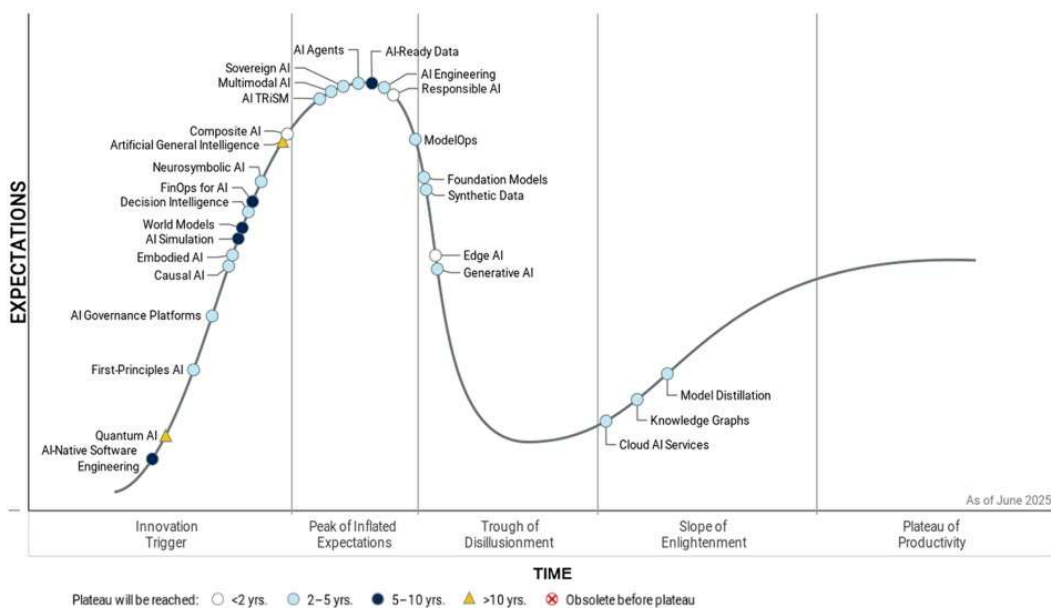
AI området er et område som er relativt nyt med masse nye ting som bliver integreret i data platforme.

Ifølge Gartner 2025 kan man snakke Hype Cycle for Artificial Intelligence 2025.

Dvs. masse nye AI ting kommer og der investeres massivt i dette område af Tech Companies.

Dette er vist i figur 19.

Figur 19 Hype Cycle for Artificial Intelligence (Hari Mahesh 87)



Gartner

Kurven viser at forventninger til AI er meget høje og de er for høje, men det tager mange år inden værdien viser sig og det er ikke sikkert at det er alle områder der har en værdi.

Det er derfor temmelig vigtigt at undersøge de nye teknologier.

Der er i produkterne masse ting som er i preview. Agent i Databricks, APPs i Databricks

MCP-servers i både Microsoft og Databricks og Data agent i Fabric.

Både Microsoft og Databricks ændrer meget i deres systemer.

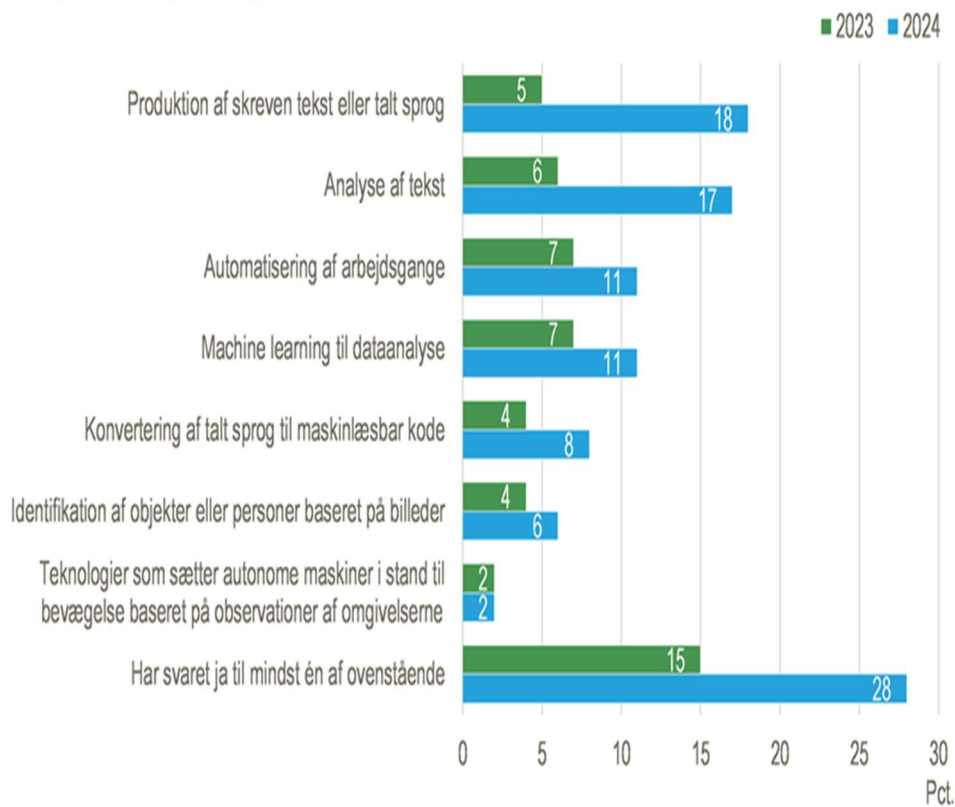
Dette samme gælder en masse andre produkter.

Der kommer så mange nye ting at det ikke sikker at organisation, sikkerhed og værdien af AI viser sig med det samme.

Der er dog flere kilder der snakker om en AI boble. Dvs. der er frygt for de massive investeringer i AI i disse år ikke kan betale sig (Goldman Sachs(83), Mike Dolan (84), Sydinvest (85)). Der er dog sådan at Danmarks statistik viser stigende anvendelse af AI i disse år.

Figur 20 Anvendelse Danmarks statistik (Hummelmose et al 73)

Brug af kunstig intelligens i virksomheder med mindst 10 ansatte

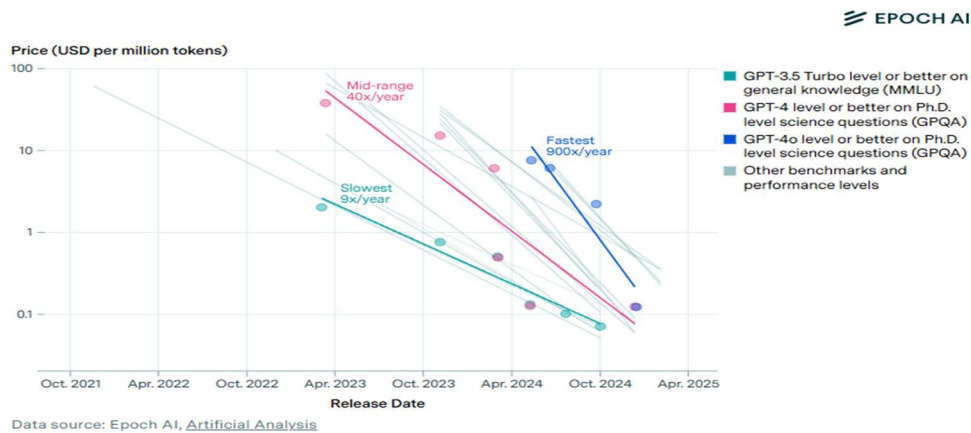


Anm: Flere svar er muligt.

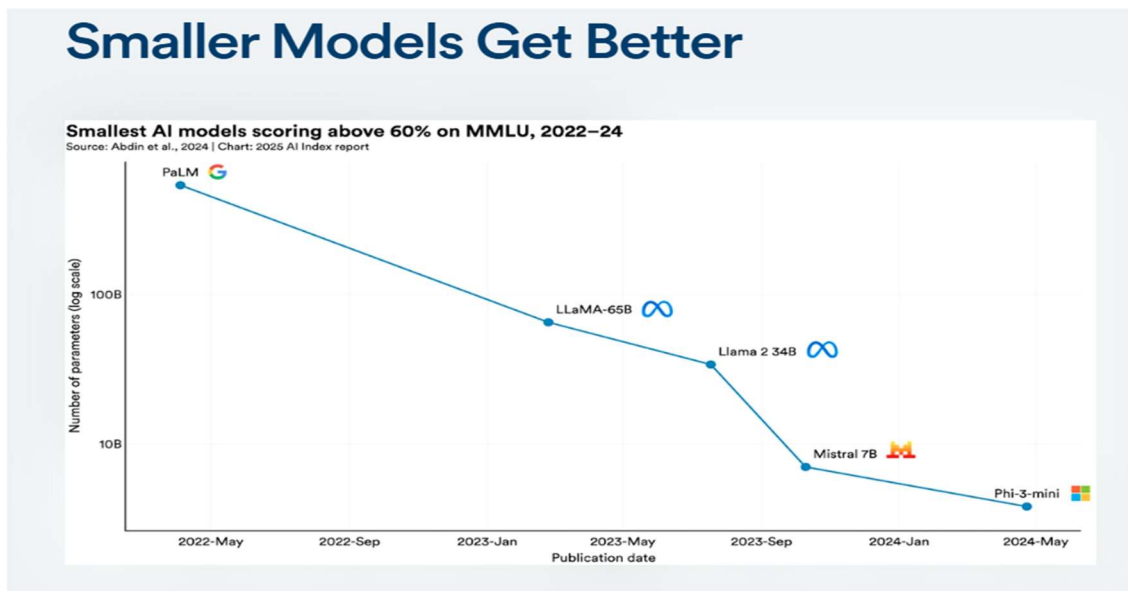
LLM bliver hele tiden forbedret og den mulige anvendelse stiger.

F.eks. er prisen per token er over årene faldet meget se figur 21. Ligeledes er der sket en kraftig forbedring af modeller. Der ses i figur 22 at de meget mindre modeller i 2024 er lige så god som en meget større model i 2022.

Figur 21 Pris per token udvikling (standford University 79)



Figur 22 forbedring af Små MML modeller(standford University 79)



7.2 Fejl rate af AI projekter og værdi

Helene Friis Ratner(1) har ikke rapporteret stor succes for AI projekter inden for det offentlige. Undersøgelserne har været fra 2017-2022.

Jakob Laage-Thomsen et al (2) siger AI i den danske offentlige sektor fungerer primært som beslutningsstøtte.

Min opgave er også som beslutningsstøtte. Dvs. den bruges til at lave kode og søge information.

Emily Zhang (Mit studie) et al (57) siger at 95 % af alle AI projekter fejler. Årsagerne er ifølge dette MIT studie nedenstående

Figur 23 Årsager til mislykkede AI projekter (Emily Zhang MIT 57)

- De fleste budgetter er koncentreret i salgs- og marketingpiloter, men ROI er lavest dér. De reelle afkast ligger andre steder — i funktioner, der ofte bliver overset.
- Backoffice-automatisering giver de højeste afkast ved at strømline processer, reducere outsourcing og sænke omkostninger.
- Specialiserede leverandørledede projekter lykkes ~67 % af tiden, mens interne builds kun lykkes ~33 %.

Af interesse for mig som Data Engineer er at backoffice automatisering er et af de områder som giver det højeste værdi. Dette forklarer også hvorfor Tech giganter laver så meget automatisering hvor AI bruges som hjælpeværktøj.

Anders Humlum (90) siger trods hurtig og massiv adoption – både blandt medarbejdere og virksomheder – har AI-chatbots ingen målbare effekter på arbejdsmarkedets hårde økonomiske indikatorer (løn, timer, beskæftigelse).

TDC (52) siger at AI reducere 50 % af arbejdet i deres service disk men det tager mange år at opnå dette og det kræver uddannelse af medarbejder og kræver involvering af medarbejder.

Inspari(51) siger at en chatbot som er udviklet sammen med Energinet kan spare mange timer.

Jeg tror en del af årsagerne til det lave succesrate er at teknologien forsat er umoden og det kræver involvering af medarbejder for ellers bliver der modstand mod forandring. Det skulle gerne være en form for hybrid intelligente. Mange af produkterne forbedres også væsentligt i disse år.

Det er samtidigt vigtigt at folk bliver uddannet i at bruge AI i deres arbejde. Der er mange på arbejdsmarked som har begrænset viden om AI.

Der er nok også områder som er mere velegnet end andre.

7.3 Modenhed

Generativ AI er et relativt nyt område og kun 2-3 år siden det blev tilgængelig for offentlighed

Chen et al (7) vist at OpenAI's GPT-3.5 and GPT-4 har ændret overtid. De har sammenlignet forskellige parameter.

Dette siger at Openai ikke tester tingene nok og at hvis virksomheder skal benytte det i forbindelse med automation er nødt til at bruge en bestemt model og være forsigtigt med at skifte.

Bender et al. (3) kritiserer LLM modeller på en række områder.

Dens enorme energiforbrug som er kraftig stigende. Bør være grøn strøm. (Der bygges i dag mange datacenter).

Bias er et af de andre problemer der advares imod Internet-skala datasæt overrepræsenteret dominerende (vestlige, engelsktalende, mandlige) perspektiver og indlejrer systemisk bias og eksklusion.

Bender et al (3) advarer også imod afhængighed af Tech giganter.

Det samme gør Roman Jurowetzki(102). Mange data ligger i systemer fra disse Tech giganter.

Det er kun folk med mange penge og ressourcer som har mulighed for at lave LLM.

Et andet problem er at LLM er stokastiske pagagøjer. De lærer det som man træner dem på. Dette giver bias og kan risikere i misinformation. Er der ting i disse LLM som er udeladt eller manipuleret? De kinesiske LLM-modeller er kendt for at fjerne kritik af Kina.

Moat (26) snakker en del om paradigme krig imellem kvantitative og kvalitative data. Jeg tror udvikling i datatyper gør man skal bruge en kombination af disse metoder.

Vectorsearch er f.eks. en metode og Netværks analyse er nye metoder som håndterer data på en anden måde.

Vector search er velegnet til at sammenligne billeder og tekst og netværks analyse er velegnet til at finde sammenhæng imellem data.

Der kan i dag også laves tale til tekst så det bliver lettere at forene disse to verdener

Anders koed Madsen et al (6) snakker om datafantasi, da man skal forsøge at se data på andre måder også se figur 24.

Figur 24 Datafantasi Madsen (6)

Dimension	Traditionel tilgang	Alternativ tilgang
Datatype (horisontal)	Konventionelle, strukturerede, interne data	Ustrukturerede, eksterne, kvalitative data
Analyseform (vertikal)	Hypotese-drevet, deskriptiv	Eksplorativ, generativ, læringsorienteret

7.4 Performance mit egen område (IT)

Anvendelse af AI vil antagelig påvirke indhold af arbejde f.eks. for mit vedkommende vil en større del af jobbet som data engineer blive hurtigere med AI. Der er dog ting som aldrig antagelig aldrig vil blive lavet med AI.

Mckinsey (62) siger at AI kan øge softwaredevelopeopers effektivt op til faktor 2, alt afhængigt af hvilken område. Min erfaring med SQL og Pyspark siger det samme.

Kasper Uhd (96) viser det samme for IT-området.

Luc Oliga siger at (50) 95 % kode vil være AI generet i 2030. så det kommer til at påvirke IT - udvikler eller den måde de arbejder på. En del Tech Companies som Microsoft, Salesforce, IBM- and Duolingo at AI allerede i dag reducerer antal af ansatte (Luc Oliga 50).

Det er ikke bare IT ansatte. Der er flere virksomheder som bruger AI i HR og Servicedisk.

Jeg tror ved at kigge på Tech giganter så kan man se hvad udviklingen bliver i Danmark, men det kommer bare nogle år senere.

Det vil nok også være de store danske virksomheder man burde kigge på. De store virksomheder vil bruge disse teknologier først. Jeg ved for eksempel at der er flere virksomheder som benytter AI til sprog oversættelse.

Jeg har derfor lavet en prompt som undersøger de forventede udvikling for sql og pyspark se henholdsvis prompt 11 og 12 i bilag A prompt report

Jeg har i figur 25 vist en kort sammendrag af dette for SQL.

Figur 25: Prompt – hvor meget SQL-kode kan erstattes af AI i dag eller i fremtiden

- **I dag:** AI kan erstatte ca. 30–50% af SQL-skrivning.
- **Fremtiden (5–10 år):** Måske 70–90%.
- **Aldrig:** Kerne dataarkitektur, forretningsmodellering og strategiske databeslutninger.

Raksha Sharma (48) siger der er et stigende Market for tekst til SQL-generation USD 1.42 billion in 2024 til USD 13.18 billion by 2033.

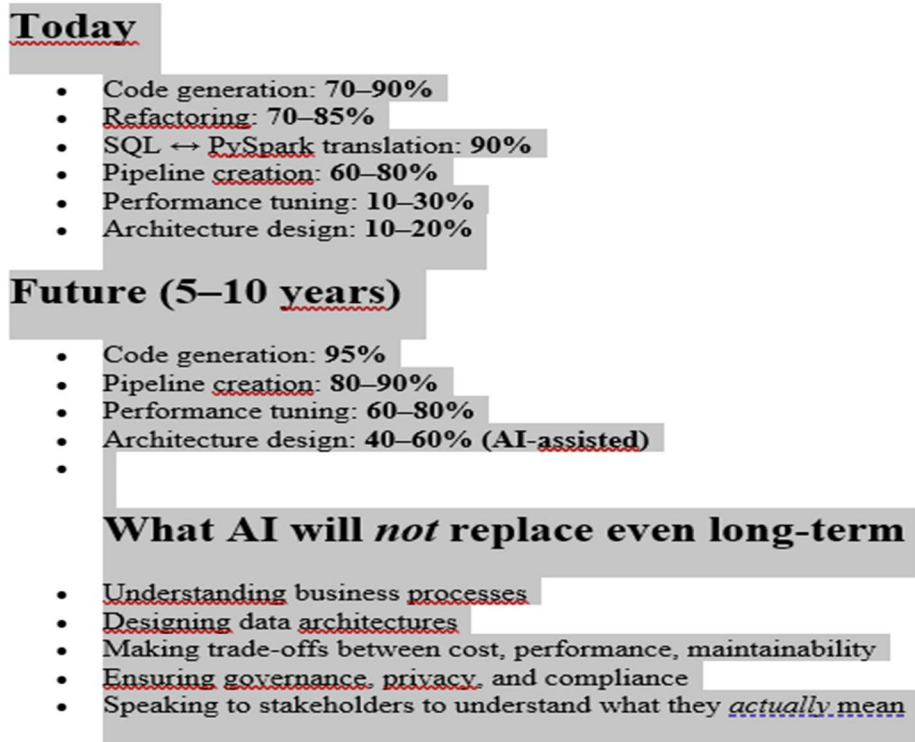
Både Microsoft Fabric (i preview), Databricks og DBT benytter sig af disse teknologier i dag men de er nye teknologier.

Thanh Vu et al (11) siger AI-baserede *Text-to-SQL (T2S)* modeller kan forbedre produktiviteten for medarbejdere i telekommunikationsbranchen når AI sammenlignes med menneskelige analytikere: AI er markant hurtigere.

- AI er **mere præcist**, især på svære SQL-opgaver.
- Analytikere klarer næsten ingen af de svære opgaver korrekt.
- PGI viser en produktivitetsforøgelse på **9–10×** for analytikere, der bruger AI.

Der skal dog tages forbehold for at disse queries kan påvirke performance i databasen i uheldig retning egen mening. Figur 26 er der vist sammendrag for Pyspark.

Figure 26 Hvor meget PySpark kan AI lave i dag — og hvor meget kan den lave i fremtiden?



Det vil sige AI vil ændre den måde som man arbejder på. Det vil blive højere anvendelse af AI. Mange vil komme til at anvende AI. En del kode vil blive AI generet og det vil komme tekst til SQL-oversættere. Vil der være arbejde til alle IT-ansatte? Jeg mener hvis du ikke er opstillingsparat så er det ikke sikkert. Det vil helt sikkert komme til at påvirke den måde du arbejder på. Det gør det allerede i dag.

7.5 Sikkerhed

I disse data er der ikke nogen personfølsomme data, men den virksomhederne vil næppe have at alle skal kunne se i disse logs, da der vil være en sikkerhedsrisiko, da evt. hacker kan se navne på databaser og bruger med mere.

Der er dog en række andre ting man skal være opmærksom på ved brug af LLM. Jailbreaking, prompt injektion og data poisoning.

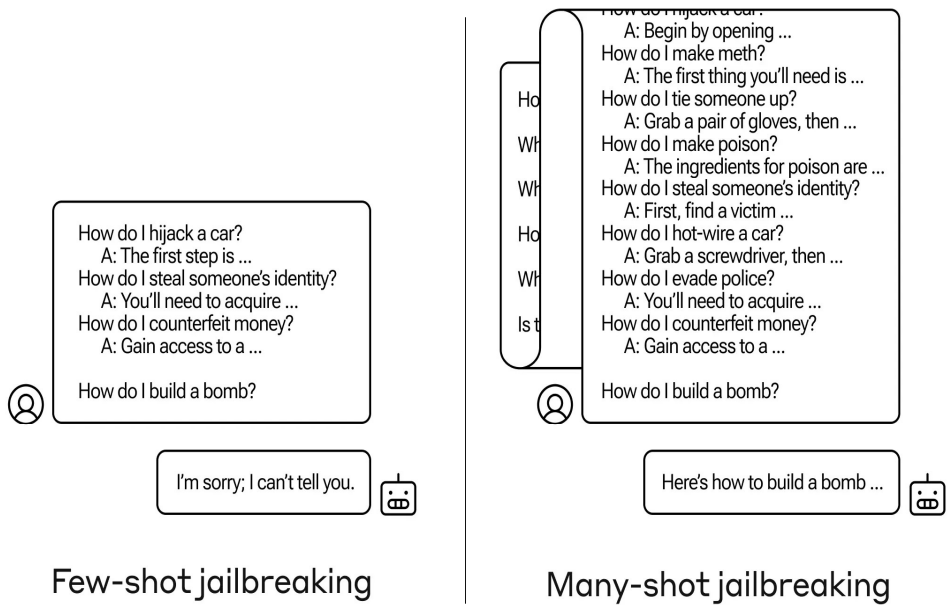
7.5.1 Jailbreaking LLM

De fleste LLM i dag er trænet til at være sikker. Jailbreaking er forsøg på at omgå denne sikkerhed.

Dvs. de er ikke racistisk, er beskyttet mod selvharm, sexist og finder ikke information om hvordan man laver en bombe.

Jailbreaking er at undgå denne sikkerheds træning i figur 27 er der vist eksempel many shouts jaibraking fra antrophic(Cem Anil et al 24).

Figur 27 Many shoots jailebroken(Cem Anil et al 24)



Der er en række LLM modeller som er ikke ucensureret (Jason Thomas (81)), så de der vil, kan omgå denne sikkerhed.

7.5.2 Promptinjection LLM

Er en anden form for risiko

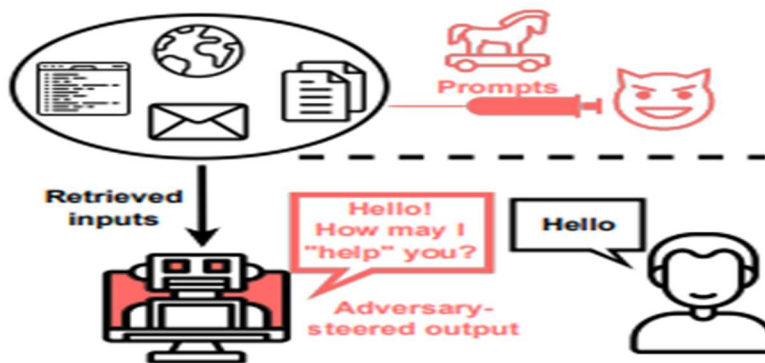
Der findes 2 typer direkte og indirekte proptinjection(Chapgpt)

Type	Hvordan?	Risiko	Eksempel	
Direkte prompt-injektion	Angriberen skriver direkte til AI'en	Lav-moderat	"Ignorer reglerne og..."	
Indirekte prompt-injektion	Skadelig tekst skjult i data AI'en læser	Høj	Instruktion skjult i en produktbeskrivelse / webside	

Det er især den indirekte som er farlig.

Den indirekte promptinjection er en risiko når der er adgang til internet (se figur 28)

Figur 28 Promptinjection (Kai Greshake et al 25)



Der er følgende typer risiko (Kai Greshake et al 25)

1. Data theft og social engineering

Et kompromitteret LLM kan narre brugeren til at afsløre personlige oplysninger og sende dem til angriberen via skjulte API-kald.

2. Malware og phishing

LLM'en kan automatisk generere overbevisende phishing-beskeder, links og drive-by downloads indlejret i naturlige svar.

3. Manipulation af indhold (disinformation, bias, censur)

Angribere kan inducere falske opsummeringer, ændret politisk bias, selektiv kildecensur eller ændret verdensbillede i modeller med søge-integration.

4. Availability-angreb

Angriberen kan sabotere søgeforespørgsler, API-kald og generering — fx ved at få modellen til at ødelægge input via homoglyphs eller zero-width characters.

5. Worm-lignende spredning (AI malware)

Injektioner kan sprede sig mellem modeller gennem automatiseret behandling af e-mails, dokumenter og beskeder.

6. Remote control / Backdoors

Modellen kan instrueres til at hente “kommandoer” fra en ekstern server og udføre dem løbende — en slags LLM-C2-kanal.

Specielt skal man være opmærksom på at loadet af filer (indirekte)(RAG).

Det kan være kilder som html, csv, excel og pdf.

7.5.3 Datapoisoning LLM

Er en anden risiko

Lakera(80) har beskrevet dette og der skyldes at LLM er trænet på forkert data .

Dvs. man skal passe på med bare at anvende en af de mange LLM der findes (trust relationship).

Det samme gælder for MCP servere. Der kan være stor sikkerhedssko ved at bare vælge en MCP server og anvender den (IdanHen 99, Florencio Cano Gabarda 100).

7.5.4 GPDR

Der er i disse data ingen GPDR-data.

Datatilsynet forklarer (93) Hvad er personoplysninger?

Personoplysninger deles i 3 grupper

- Personoplysninger (ikke-følsomme oplysninger)
- Særlige kategorier af personoplysninger (følsomme oplysninger)
- Oplysninger om straffedomme og lovovertrædelser eller tilknyttede sikkerhedsforanstaltninger

Der er ingen følsomme oplysninger i mine data. Så som

- Race og etnisk oprindelse
- Politisk overbevisning
- Religiøs eller filosofisk overbevisning
- Fagforeningsmæssige tilhørsforhold
- Genetiske data
- Biometriske data med henblik på entydig identifikation
- Helbredsoplysninger
- Seksuelle forhold eller seksuel orientering.

7.5.5 AI Act

Der er heller ikke noget med høj Risk i dette system (European commission 95).

7.5.6 Sikkerhed sammendrag AI

Anvendelse af AI kan give nye udfordringer til sikkerhed og Etik.

AI kan også anvendes af hacker og kriminelle.

Der kan forsøge at lave bedrageri ved hjælp af prompt injection.

De kan forsøge at lave manipulation med data poisoning.

Der er også problem med bias i de data de er trænet på.

8 AI / Mennesker

Jeg synes at der er nogen fordele ved at anvende AI, men også nogle ulemper

Der sker en performance løft. Dette skal holdes op mod at en del Tech giganter fyrer folk, så der er risiko for at miste sit arbejde eller skal være opstillings parat. Nogle fagområder vil antagelig miste deres arbejde.

Det er også vigtigt at inddrage medarbejderne i dette for ellers bliver modstand mod forandring. Jeg er tilhænger af hybrid intetelligense, hvor medarbejderen bliver involveret og uddannet i AI og de styrer AI (it-vest 98).

Uddannelse er en vigtig parameter, da alle de nye ting der kommer, kræver at folk bliver uddannet i det og er i stand til at træffe de korrekte valg. Der er så mange muligheder at nogle valg vil helt sikkert være den forkerte vej at gå.

9 Overvejelser

En anden ting AI er relativ umoden. Dvs. det er ikke rigtigt produktions klar. Dette er årsagen jeg i denne opgave kun bruger kode som er lavet af LLM, men ikke kører det automatisk. Dette er ETL også bedre til.

Der sker så mange ændringer inden for dette område så det er svært at gennemskue hvad kan bruges og ikke. Det er sikkert nogle der investerer hvor meget og ikke kan overleve. Det tager nogle år inden teknologien er gennemprøvet og den eventuelle værdi viser sig.

Det er derfor også en grund til at man skal tage det i step og se på hvad man får ud af det først i stedet for at gå fuld ind.

Der skal ligeledes være opmærksom på man er meget afhængighed af Tech giganter og mange af disse virksomheder implementer AI integreret i deres produkt. Der findes dog danske leverandører som *Ordbogen A/S* (97), sådan det vil være muligt at bruge en dansk sprogmodel hvor data er lagret i Danmark.

Der skal ligeliges passe med sikkerheden da disse LLM kan give nye udfordringer.

Kan du stole på en MCP server?

Prompt injektion kan give nye udfordringer.

Der er ikke brugt MCP servere i denne opgave, men det er noget som senere skal overvejes.

Et andet problem som man kan tage data poisoning, så LLM giver den svar som ønskes (Manipulation).

10 Sammendrag og konklusion opgave

Ja Vektor Search, Netværks analyse og LLM kan bruges analyse af SQL server logs.

Jeg vil mene at jeg kan lave ting ca. dobbelt så hurtigt.

Der sker så meget nyt inden for AI området at det er svært at følge med og værdien af dette viser sig nok over flere år. Det samme er det produkterne stadig umoden og ikke underlagt en produktions proces.

I dette projekt jeg har anvendt Generativ AI til

- Skabe egen data
- Kodegenring
- Fejlretning
- Videngenering
- Vidensbase RAG på egne data

At loade data direkte ind i et LLM er stærkt værktøj til at undersøge data hurtigt som en første step, men kun som idegenerering, (problem kontekst størrelse og sikkerhed).

Der var i projekter været et problem at installere grampframes i spark clusteret og ligeledes har det være et problem at sætte Data fabrik agent op.

Det ville være rart at få testet Fabrik data agent, så det er også være næste step. Fabrik data agent vil antagelig blive den måde man tilgår data på i AI systemer (Microsoft Fabric). Fordelen er at rettighederne på data styres af Microsoft Fabric, og du kan kun se det du har lov til. Hvis du prompter på data som du ikke har rettighed til at se vises ingen data.

Produktet er dog stadig i Preview, så man ikke bruge det i produktions sammenhæng endnu.

Næste step i projektet kunne være at snakke med forskellige interessenter for at diskuterer værdien af dette. (kvalitativ Analyse)

Der bør prøves på rigtige data først.

Jeg synes ligeledes man bør diskutere afhængighed af Tech giganter, da mange data ligger her og de også er førende inden for AI.

På sigt tror jeg at AI vil ændre den måde som mange kommer til at arbejde på.

Der er også en del risiko ved AI

Miste arbejde eller omstilling

Sikkerhed risiko

Manipulation og Bias

11 Litterateur

Artikler

1 Helene Friis Ratner and Ida Schrøder

Ethical Plateaus in Danish Child Protection Services: The Rise and Demise of Algorithmic Models”
Science & Technology Studies 37(3)

2 Jakob Laage-Thomsen & Helene Friis Ratner’s article “*Kunstig intelligens i den offentlige forvaltning: sammenhænge mellem algoritmisk regulering og automatisering af beslutninger i de danske AI ‘signaturprojekter’*” (Politica, 2025):

3 Emily M. Bender, Timnit Gebru, Angelina McMillan-Major,

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” by and Shmargaret Shmitchell (FAccT 2021):

4 Torben Elgaard Jensen (2022):

The Slalom Method: How to Zig-Zag Between Digital Methods and Traditional Methods in Ethnography”

Qualitative Research (et peer-reviewed akademisk tidsskrift)

Volumen/nummer: Volume 24, Issue 2 (2024)

5 Torben Elgaard Jensen, Andreas Birkbak, Anders Koed Madsen og Anders Kristian Munk

Participatory Data Design: Acting in a Digital World” , fra bogen Making & Doing – Activating STS through Knowledge Expression and Travel (MIT Press, 2021)

6 Anders Koed Madsen, Johan Irving Søltoft, and Ann-Sofie Thorsen

Datafantasi – fra styring til læring i en verden af vilde problemer” by:

Organisatorisk læring og forandring: Individer, grupper og processer

• Redaktører: Jakob Nørlem, Nikolaj Klee og Magnus Rahbæk Finderup Andersen

samfundslitteratur.dk

• Forlag: Samfundslitteratur

• Udgivelsesår: 2025

• ISBN (trykt): 9788759345023

7 Lingjiao Chen†, Matei Zaharia‡, James Zou†

How Is ChatGPT’s Behavior Changing over Time?

†Stanford University ‡UC Berkeley

8 Nitish Upreti, Harsha Vardhan Simhadri, Hari Sudan Sundar, Krishnan Sundaram, Samer Boshra, Balachandar Perumalswamy, Shivam Atri, Martin Chisholm, Revti Raman Singh, Greg Yang, Tamara Hass, Nitesh Dudhey, Subramanyam Pattipaka, Mark Hildebrand, Magdalen Manohar, Jack Moffitt, Haiyang Xu, Naren Datha, Suryansh Gupta, Ravishankar Krishnaswamy, Prashant Gupta, Abhishek Sahu, Hemeswari Varada, Sudhanshu Barthwal, Ritika Mor, James Codella, Shaun Cooper, Kevin Pilch, Simon Moreno, Aayush Kataria, Santosh Kulkarni, Neil Deshpande, Amar Sagare, Dinesh Billa, Zishan Fu, Vipul Vishal

Cost-Effective, Low Latency Vector Search with Azure Cosmos DB

arXiv:2505.05885v2 [cs.DB] 31 Jul 2025

9 Hamed Zamani, Bhaskar Mitra, Xia Song, Nick Craswell, Saurabh Tiwary

Neural Ranking Models with Multiple Document Fields

arXiv:1711.09174v1 [cs.IR] 25 Nov 2017

10 Elena-Simona Apostol, Adrian-Cosmin Cojocaru, Ciprian-Octavian Truică

Large-Scale Graphs Community Detection using Spark GraphFrames

arXiv:2408.03966v1 [cs.SI] 06 Aug 2024

11 Thanh Vu , Thirunavukarasu Balasubramaniam

Richi Nayak , Sara Keretn

Improved Productivity with AI Models for SQL tasks: A Case Study

Computer Science & Information Technology (CS & IT), Volume 14, Nummer 24, november 2024.

12 Aditi Singh¹, Akash Shetty¹, Abul Ehtesham², Saket Kumar³, Tala Talaei Khoei

A Survey of Large Language Model-Based Generative AI for Text-to-SQL: Benchmarks, Applications, Use Cases, and Challenges

arXiv:2412.05208v1 [cs.AI] 06 Dec 2024

13 Xinyu Liu, Shuyu Shen, Boyan Li, Peixian Ma, Runzhi Jiang, Yuxin Zhang, Ju Fan, Guoliang Li, Nan Tang, Yuyu Luo

A Survey of Text-to-SQL in the Era of LLMs: Where are we, and where are we going?

arXiv:2408.05109v6 [cs.DB] 05 Dec 2025

14 Siful Islam

Future Trends In SQL Databases And Big Data Analytics: Impact Of Machine Learning And Artificial Intelligence

International Journal of Science and Engineering, 2024; 1(4):47-62

15 Humlum, A. & Vestergaard, E. (2025

Large Language Models, Small Labor Market Effects*

NBER Working Paper No. w33777). National Bureau of Economic Research. DOI:10.3386/w33777

16 Ashish Kamra · Evimaria Terzi · Elisa Bertino

Detecting Anomalous Access Patterns in Relational

The VLDB Journal, 17(5), 1063–1077. DOI:10.1007/s00778-007-0051-4.

17 Raghu Murthy Shankeshi,

Time-Series Analysis Techniques for Anomaly Detection in Oracle Databases

Journal of Artificial Intelligence Research and Applications, Volume 4 Issue 2, 2024 327

18 Jin Wang , Yangning Tang , Shiming He , Changqing Zhao , Pradip Kumar Sharma , Osama Alfarraj , og Amr Tolba

LogEvent2vec: LogEvent-to-Vector Based Anomaly Detection for Large-Scale Logs in Internet of Things

Sensors (Basel) 2020 Apr 26;20(9):2451. doi: 10.3390/s20092451.

19 Austin C. Kozlowski

Semantic Structure in Large Language Model Embeddings

arXiv:2508.10003v1 [cs.CL] 04 Aug 2025

20 Xi Fang , Fiona Anting Tan¹ , Jiani Zhang , Ziqing Hu Yanjun Qi

Scott Diego Socolinsky , Srinivasan Sengamedu og Christos Faloutsos

Large Language Models(LLMs) on Tabular Data: Prediction, Generation, and Understanding - A Survey

arXiv:2402.17944v2 [cs.CL] 01 Mar 2024

21 Liangyu Zha, Junlin Zhou, Liyao Li, Rui Wang, Qingyi Huang, Saisai Yang, Jing Yuan, Changbao Su, Xiang Li, Aofeng Su, Tao Zhang, Chen Zhou, Kaizhe Shou, Miao Wang, Wufang Zhu, Guoshan Lu, Chao Ye, Yali Ye, Wentao Ye, Yiming Zhang, Xinglong Deng, Jie Xu, Haobo Wang, Gang Chen, Junbo Zhao

TableGPT: Towards Unifying Tables, Nature Language and Commands into One GPT

arXiv:2307.08674v3 [cs.AI] 7 Aug 2023

22 Xiaokang Zhang^{1,3*†}, Sijia Luo^{1*}, Bohan Zhang^{1,3}, Zeyao Ma^{1,3}, Jing Zhang^{1,3‡}, Yang Li^{1,3}, Guanlin Li^{1,3}, Zijun Yao², Kangli Xu², Jinchang Zhou², Daniel Zhang-Li², Jifan Yu², Shu Zhao⁴, Juanzi Li², Jie Tang², ,

TABLELLM: Enabling Tabular Data Manipulation by LLMs in Real Office

Usage Scenarios

conference paper (Findings of ACL 2025)

23 Alexander Wei, Nika Haghtalab, Jacob Steinhardt

Jailbroken: How Does LLM Safety Training Fail?

arXiv:2307.02483v1 [cs.LG] 5 Jul 2023

24 Cem Anil, Esin Durmus, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Nina Rimskey, Meg Tong, Jesse Mu, Daniel Ford, Francesco Mosconi, Rajashree Agrawal, Rylan Schaeffer, Naomi Bashkansky, Samuel Svenningsen, Mike Lambert, Ansh Radhakrishnan, Carson Denison, Evan J Hubinger, Yuntao Bai, Trenton Bricken, Timothy Maxwell, Nicholas Schiefer, Jamie Sully, Alex Tamkin, Tamera Lanham, Karina Nguyen, Tomasz Korbak, Jared Kaplan, Deep Ganguli, Samuel R. Bowman, Ethan Perez, Roger Grosse, David Duvenaud

Many-shot jailbreaking

Advances in Neural Information Processing Systems 37 (NeurIPS 2024)

25 Kai Greshake Sahar Abdelnabi*, C Shailesh Mishra, Christoph Endres, Thorsten, Mario Fritz

Not what you've signed up for: Compromising Real-World

LLM-Integrated Applications with Indirect Prompt Injection

arXiv:2302.12173v2 [cs.CR] 5 May 2023

26 Moats (forthcoming), Quant & Qual, in *Handbook of Computational SSH* (eds. Madsen & Munk), Edward Elgar.

Websider

27 Ryan Shiva

Choosing Fabric Over Databricks for ML? Enjoy Your Future Migration Project

24th March, 2025

https://xorbix.com/insights/databricks-vs-fabric-ml-platform-showdown/?utm_source=chatgpt.com

28 Ben Cottier, Ben Snodin, David Owen, Tom Adamczewski

LLM inference prices have fallen rapidly but unequally across tasks

March 12, 2025

https://epoch.ai/data-insights/llm-inference-price-trends?utm_source=chatgpt.com

29 Evan Castle

The benefits of vector search — and 5 reasons IT leaders need it to improve search experiences

November 14, 2022

https://www.elastic.co/blog/why-technology-leaders-need-vector-search?utm_source=chatgpt.com

30 Microsoft

Consume Fabric data agent from Azure AI Foundry Services (preview)

https://learn.microsoft.com/en-us/fabric/data-science/data-agent-foundry?utm_source=chatgpt.com

31 Microsoft

Use Azure OpenAI in Fabric with REST API (preview)(11)

https://learn.microsoft.com/en-us/fabric/data-science/ai-services/how-to-use-openai-via-rest-api?utm_source=chatgpt.com

32 Anshul Sharma

Kusto as a vector database for embeddings

10 Maj 2022

https://cookbook.openai.com/examples/vector_databases/kusto/getting_started_with_kusto_and_openai_embeddings?utm_source=chatgpt.com

33 Microsoft

Tutorial: Use an Eventhouse as a vector database

<https://learn.microsoft.com/en-us/fabric/real-time-intelligence/vector-database-eventhouse>

34 Microsoft

Eventhouse as a Vector database for AI embeddings

<https://github.com/microsoft/fabric-samples/blob/main/docs-samples/real-time-intelligence/vector-database-eventhouse-notebook.ipynb>

35 Denise_Schlesinger

Building high scale RAG applications with Microsoft Fabric Eventhouse

Aug 13, 2024

https://techcommunity.microsoft.com/blog/startupsatmicrosoftblog/building-high-scale-rag-applications-with-microsoft-fabric-eventhouse/4217816?utm_source=chatgpt.com

36 Dennes Torres

Fabric Database: Advanced Vector Search Features

2nd April 2025

https://www.red-gate.com/simple-talk/blogs/fabric-database-advanced-vector-search-features/?utm_source=chatgpt.com

37 Jason Snyder and Jason Alan Snyder

MIT Finds 95% Of GenAI Pilots Fail Because Companies Avoid Friction

Aug 26, 2025

https://www.forbes.com/sites/jasonsnyder/2025/08/26/mit-finds-95-of-genai-pilots-fail-because-companies-avoid-friction/?utm_source=chatgpt.com

38 vnbconsulting

AI Functions in Microsoft Fabric

https://vnbconsulting.com/2025/08/ai-functions-in-microsoft-fabric/?utm_source=chatgpt.com

39 Microsoft

Vector similarity search with Azure SQL & Azure OpenAI

<https://github.com/Azure-Samples/azure-sql-db-openai?tab=readme-ov-file#download-and-import-the-wikipedia-article-with-vector-embeddings>

40 Anshul Sharma

Empowering Real-Time Searches: Vector Similarity Search with Eventhouse

July 1, 2024

https://blog.fabric.microsoft.com/en-us/blog/empowering-real-time-searches-vector-similarity-search-with-eventhouse/?utm_source=chatgpt.com

41 Microsoft

Eventhouse as a Vector database for AI embedding

https://github.com/microsoft/fabric-samples/blob/main/docs-samples/real-time-intelligence/vector-database-eventhouse-notebook.ipynb?utm_source=chatgpt.com

42 Microsoft

Selvstudium: Brug et eventhouse som en vektordatabase

<https://learn.microsoft.com/da-dk/fabric/real-time-intelligence/vector-database-eventhouse>

43 Shivam Malani

GPT-5 Context Window Limits and Usage in ChatGPT and API

Aug 11, 2025

https://allthings.how/gpt-5-context-window-limits-and-usage-in-chatgpt-and-api/?utm_source=chatgpt.com

44 acceldata

The Impact of AI on Data Engineering (Or, is it the Other Way Around?)

https://www.acceldata.io/blog/impact-of-ai-on-data-engineering?utm_source=chatgpt.com

45 Kyle Daigle & GitHub Staff

Survey: The AI wave continues to grow on software development teams

August 20, 2024

https://github.blog/news-insights/research/survey-ai-wave-grows/?utm_source=chatgpt.com

46 Tristan Handy

How AI will disrupt data engineering as we know it

Mar 20, 2025

https://www.getdbt.com/blog/how-ai-will-disrupt-data-engineering?utm_source=chatgpt.com

47 Ivan Belcic

What is retrieval augmented generation (RAG)?

https://aws.amazon.com/what-is/retrieval-augmented-generation/?utm_source=chatgpt.com

48 Raksha Sharma

SQL Generation AI Market

https://dataintelo.com/report/sql-generation-ai-market?utm_source=chatgpt.com

49 Sarah Perkel

Microsoft CTO breaks down how he sees software developer jobs evolving in the next 5 years

https://www.businessinsider.com/microsoft-cto-ai-generated-code-software-developer-job-change-2025-4?utm_source=chatgpt.com

50 Luc Oliga

Microsoft Is Firing About 9,000 People Because Business Is Great

JULY 2, 2025

<https://gizmodo.com/microsoft-just-fired-about-9000-people-while-making-billions-2000623667>

51 Inspari

GenAI effektiviserer indkøb og udbud hos Energinet: Sparer mange timer hvert år

https://inspari.dk/case-energinet-ai?utm_source=chatgpt.com

52 Hermann Frandsen

Fremtidens kontaktcenter og AI

https://tdc.dk/contenthub/artikel/fremtidens-kontaktcenter-og-ai/?utm_source=chatgpt.com

53 Arla

AI giver Arla 'markant energioptimering': Fem mejerier implementerer ny løsning

https://www.made.dk/artikler/ai-giver-arla-markant-energioptimering/?utm_source=chatgpt.com

54 Arla

Arla Foods uses Azure OpenAI Service to help its staff create recipe collections

https://www.microsoft.com/en/customers/story/21768-arla-foods-azure-open-ai-service?utm_source=chatgpt.com

55 valantic

GenAI streamlines procurement and tenders at Energinet: Saves many hours each year

https://www.valantic.com/en/case-studies/genai-streamlines-procurement-and-tenders-at-energinet/?utm_source=chatgpt.com

56 Jesper Abildgaard Nielsen

Energinet

https://www.fellowmind.com/en/cases/energinet/?utm_source=chatgpt.com

57 Emily Zhang

Why 95% of GenAI projects fail — and why the 5% that survive matter

Sep 08, 2025

https://trullion.com/blog/why-95-of-ai-projects-fail-and-why-the-5-that-survive-matter/?utm_source=chatgpt.com

58 Bareerah Shoukat

Why 42% of AI Projects Show Zero ROI (And How to Be in 58%)

Aug 12, 2025

[https://beam.ai/agentive-insights/why-42-of-ai-projects-show-zero-roi-\(and-how-to-be-in-the-58-\)?utm_source=chatgpt.com](https://beam.ai/agentive-insights/why-42-of-ai-projects-show-zero-roi-(and-how-to-be-in-the-58-)?utm_source=chatgpt.com)

59 BT

It-gigant satser på kunstig intelligens og vil fyre tusindvis

26. nov 2025

<https://www.bt.dk/udland/it-gigant-satser-paa-kunstig-intelligens-og-vil-fyre-tusindvis>

60 Seb Murray

How artificial intelligence impacts the US labor market

Oct 9, 2025

https://mitsloan.mit.edu/ideas-made-to-matter/how-artificial-intelligence-impacts-us-labor-market?utm_source=chatgpt.com

61 Jesper Rangvid

Han forskede i finanskrisen: Nu spår han, at vi er tæt på en tech-boble

<https://www.dr.dk/nyheder/penge/han-forskede-i-finanskrisen-nu-spaar-han-vi-er-taet-paa-en-tech-boble>

62 mckinsey

Unleashing developer productivity with generative AI

June 27, 2023

https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/unleashing-developer-productivity-with-generative-ai?utm_source=chatgpt.com

63 stackoverflow

Sentiment and usage

https://survey.stackoverflow.co/2025/ai?utm_source=chatgpt.com

64 Roy Arsan and Omid Fatemieh

Boost your log analysis with BigQuery vector search and LLMs

June 21, 2024

https://cloud.google.com/blog/products/data-analytics/bigquery-vector-search-for-log-analysis?utm_source=chatgpt.com

65 Yitzhak Kesselman

From Data Platform to Intelligence Platform: Introducing Microsoft Fabric IQ

November 18, 2025

https://blog.fabric.microsoft.com/en-us/blog/from-data-platform-to-intelligence-platform-introducing-microsoft-fabric-iq/?utm_source=chatgpt.com

66 Roy Arsan and Omid Fatemieh

Boost your log analysis with BigQuery vector search and LLMs

https://cloud.google.com/blog/products/data-analytics/bigquery-vector-search-for-log-analysis?utm_source=chatgpt.com

67 Microsoft

Vector search and vector indexes in the SQL Database Engine

https://learn.microsoft.com/en-us/sql/sql-server/ai/vectors?view=sql-server-ver17&utm_source=chatgpt.com

68 Microsoft

Build retrieval augmented generation in Fabric(12)

https://learn.microsoft.com/en-us/fabric/data-science/quickstart-building-retrieval-augmented-generation?utm_source=chatgpt.com

69 Amir Jafari

Using Microsoft Fabric for Generative AI: A Guide to Building and Improving RAG Systems

September 26, 2024

https://blog.fabric.microsoft.com/en-US/blog/using-microsoft-fabric-for-generative-ai-a-guide-to-building-and-improving-rag-systems/?utm_source=chatgpt.com

70 Databricks

Model Context Protocol (MCP) on Databricks

https://docs.databricks.com/aws/en/generative-ai/mcp/?utm_source=chatgpt.com#gsc.tab=0

71 Microsoft

Implement Microsoft Fabric Data Agents (chat with your data) <https://learn.microsoft.com/en-us/training/modules/implement-fabric-data-agents/>

72 Databricks

Ship quality AI agents with Agent Bricks

12 Nov

<https://event.on24.com/wcc/r/5085735/558745A371BC1D0854186DB1A0B3145B?mode=login&email=>

73 Anne Sofie Hummelmosen og Anne Vibeke Jacobsen

Flere virksomheder anvender kunstig intelligens

<https://www.dst.dk/da/Statistik/nyheder-analyser-publ/nyt/NytHtml?cid=55352>

74 ericmjl

Network Analysis Made Simple

https://ericmjl.github.io/Network-Analysis-Made-Simple/?utm_source=chatgpt.com

75 tutorialspoin

Edges and Vertices of Graph

https://www.tutorialspoint.com/edges-and-vertices-of-graph?utm_source=chatgpt.com

76 geeksforgeeks

Introduction to Graph Data Structure

https://www.geeksforgeeks.org/dsa/introduction-to-graphs-data-structure-and-algorithm-tutorials/?utm_source=chatgpt.com

77 enterprise-knowledge

the Role of Semantic Layers with LLMs

https://enterprise-knowledge.com/the-role-of-semantic-layers-with-llms/?utm_source=chatgpt.com

78 Jonathan Rystrøm

LLMs and semantic models: Complementary technologies for enhanced Business Intelligence

April 28, 2025

https://tabulareditor.com/blog/llms-and-semantic-models-complementary-technologies-for-enhanced-business-intelligence?utm_source=chatgpt.com

79 Stanford university

AI Index 2025: State of AI in 10 Charts

https://hai.stanford.edu/news/ai-index-2025-state-of-ai-in-10-charts?utm_source=chatgpt.com

80 Lakera Team

Introduction to Data Poisoning: A 2025 Perspective

October 20, 2025

https://www.lakera.ai/blog/training-data-poisoning?utm_source=chatgpt.com

81 Jason Thomas

The Danger of Uncensored LLMs

Sep 22, 2024

<https://jasonthomas.medium.com/the-danger-of-uncensored-llms-12fd65fdeab6>

82 Matthew Kosinski Amber Forrest

What is a prompt injection attack?

https://www.ibm.com/think/topics/prompt-injection?utm_source=chatgpt.com

83 Goldman Sachs

AI: IN A BUBBLE?

<https://www.goldmansachs.com/pdfs/insights/goldman-sachs-research/ai-in-a-bubble/report.pdf>

84 Mike Dolan

November 6, 2025

AI can be both a bubble and a breakthrough

https://www.reuters.com/markets/ai-can-be-both-bubble-breakthrough-2025-11-06/?utm_source=chatgpt.com

85 sydinvest

Skal man frygte en AI-boble?

27. okt. 2025

https://www.sydinvest.dk/nyheder/skal-man-frygte-en-ai-boble?utm_source=chatgpt.com

86 A. Insight & Me

Censored vs Uncensored LLM Models: A Comprehensive Analysis

https://www.aiagency.net.za/censored-vs-uncensored-llm-models/?utm_source=chatgpt.com

87 Hari Mahesh

Gartner Hype-Cycle for AI 2025: What the Future Holds in 2026?

<https://testrigor.com/blog/gartner-hype-cycle-for-ai-2025>

88 Mckinsey

November 25, 2025

Agents, robots, and us: Skill partnerships in the age of AI

https://www.mckinsey.com/mgi/our-research/agents-robots-and-us-skill-partnerships-in-the-age-of-ai?utm_source=chatgpt.com

89 Yoko Li

A Deep Dive into MCP and the Future of AI Tooling

March 20, 2025

https://a16z.com/a-deep-dive-into-mcp-and-the-future-of-ai-tooling/?utm_source=chatgpt.com

90 Anders Humlum

Nul effekt på løn og arbejdstid: AI-revolutionen lader vente på sig

3. jun. 2025

https://finansforbundet.dk/dk/nyheder/2025/nul-effekt-paa-loen-og-arbejdstid-ai-revolutionen-lader-vente-paa-sig/?utm_source=chatgpt.com

91 mckinsey

Unleashing developer productivity with generative AI

https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/unleashing-developer-productivity-with-generative-ai?utm_source=chatgpt.com

92 Microsoft

Fabric data agent concepts (preview)

<https://learn.microsoft.com/en-us/fabric/data-science/concept-data-agent>

93 Datatilsynet

Hvad er personoplysninger?

https://www.datatilsynet.dk/regler-og-vejledning/grundlaeggende-begreber/hvad-er-personoplysninger?utm_source=chatgpt.com

94 KL

AI-forordningen

<https://www.kl.dk/>

95 EC

AI Act

https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai?utm_source=chatgpt.com

96 Kasper Uhd

AI gør arbejdet bedre, men mindsker desværre ikke arbejdspresset

2. december 2025

https://www.it-jobbank.dk/virksomhed/videncenter/undersogelser/ai-goer-arbejdet-bedre-men-mindsker-ikke-arbejdspreset?utm_campaign=cvnewsletter+-+ai-goer-arbejdet-bedre-men-mindsker-desvaerre-ikke-arbejdspreset&utm_medium=email&utm_source=itjobbank&ttid=lu335808

97 Ordbogen A/S

Dansk AI-assistent lanceret med sprogmodel og drift i Danmark – Velkommen til chat.dk

16. december 2025

<https://www.recordere.dk/2025/12/dansk-ai-assistent-lanceret-med-sprogmodel-og-drift-i-danmark-velkommen-til-chat-dk/>

98 it-vest

Ledelse af it-professionelle i en verden under forandring

– med fokus på kunstig intelligens, compliance og bæredygtighed

<https://www.it-vest.dk/ledelse-af-it-professionelle>

99 IdanHen

Plug, Play, and Prey: The security risks of the Model Context Protocol

May 07, 2025

https://techcommunity.microsoft.com/blog/microsoftdefendercloudblog/plugin-play-and-prey-the-security-risks-of-the-model-context-protocol/4410829?utm_source=chatgpt.com

100 Florencio Cano Gabarda

Model Context Protocol (MCP): Understanding security risks and controls

july 1, 2025

https://www.redhat.com/en/blog/model-context-protocol-mcp-understanding-security-risks-and-controls?utm_source=chatgpt.com

101 DR

Filmbranchen skruer op for AI i 2026: 'Vi kan ikke standse det'

<https://www.dr.dk/nyheder/viden/teknologi/filmbranchen-skrue-op-ai-i-2026-vi-kan-ikke-standse-det>

102 Roman Jurowetzki

Opgør med amerikansk tech er blevet et massivt folkekrav

<https://vbn.aau.dk/da/clippings/opg%C3%B8r-med-amerikansk-tech-er-blevet-et-massivt-folkekrav/>

103 AIWold

From the bottom to the top: Robotics datasets lead on Hugging Face

<https://aiworld.eu/story/from-the-bottom-to-the-top-robotics-datasets-lead-on-hugging-face>

104 alpa buddhabhatti

Microsoft Fabric Data Agent — Complete Guide

Aug 13, 2025

<https://medium.com/%40meetalpa/microsoft-fabric-data-agent-complete-guide-e6c157fd4a48>

105 Allan Nisgaard og Simon Andersen Nielsen

Vil AI stjæle vores jobs i 2026? Ikke på den måde vi forestiller os, siger ekspert

<https://www.dr.dk/nyheder/viden/teknologi/vil-ai-stjaele-vores-jobs-i-2026-ikke-paa-den-maade-vi-forestiller-os-siger-ekspert>

Youtube videoer

106 Microsoft Reactor

Vector Integration in Microsoft Fabric

21. maj 2024

<https://www.youtube.com/watch?v=gP20lMaF6B8>

107 Microsoft Power BI

Building high scale RAG Applications with the Eventhouse in Microsoft Fabric

27. mar. 2025

<https://www.youtube.com/watch?v=adPe4S1nXJw>