

Hvordan bruke OpenAI mot interne datakilder og gjøre informasjon enklere tilgjengelig?

Marius Sandbu

Sky evangelist @ Sopra Steria

Demokode: <https://github.com/msandbu/gpt-ai>

Presentasjon:

MVP
Dagen

Litt om Marius

(Snart...) 37 år, gift, 2 barn og 4x4

Sky evangelist @ Sopra Steria

Litt forfatter og litt mer blogger

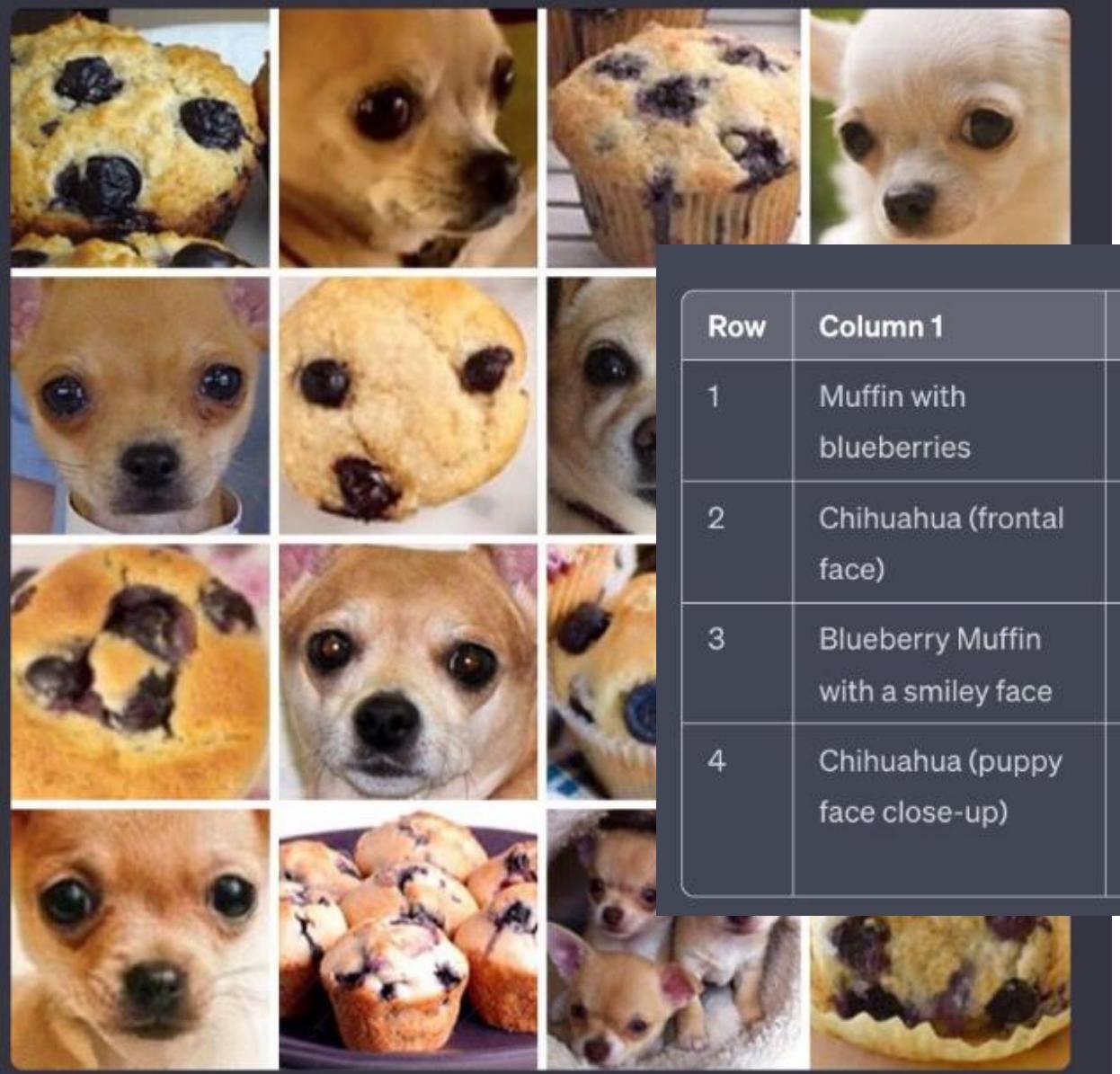
Microsoft MVP - 10 år

Klient → Azure → Sikkerhet





**Prompt: in norwegian woods, in the
mountain in the background --ar 16:9 --v
5.0**



Row	Column 1	Column 2	Column 3	Column 4
1	Muffin with blueberries	Chihuahua (side profile)	Blueberry Muffin	Chihuahua (profile looking up)
2	Chihuahua (frontal face)	Cookie with chocolate chips	Chihuahua (face close-up)	Blueberry Muffin with sugar on top
3	Blueberry Muffin with a smiley face	Chihuahua (frontal face close-up)	Blueberry Muffin with blueberries on top	Chihuahua (slightly showing teeth)
4	Chihuahua (puppy face close-up)	Group of Cookies with chocolate chips	Group of Chihuahua puppies	Muffin with blueberries

Litt bakteppe rundt bruk av språkmodeller

ChatGPT har nå estimert over 180 millioner aktive brukere

Mennesker kan lese mellom 170 til 200 ord i minuttet

GPT-2 besto av 1.5 Milliarder parametere
GPT-3 består av 175

Prosesseringskraften fra å bruke GPT-4 sammenlignet med 3 er over 10x høyere

Generering av innhold med lokal språkmodell via CPU vil gi ca 450 ord i minuttet

GPU Minnekrevet for en LLM er vanligvis 2x av antall milliarder parametere

Trening av en språkmodeller på 70 milliarder parametere koster ~250 millioner

Generering av innhold med lokal språkmodell via GPU vil gi ca 4000 ord i minuttet

Bildegjenkjenning og GPT-V

Network and Internet > Network Connections >

Diagnose this connection

Rename this connection



Ethernet

Network cable unplugged

Realtek PCIe FE Family Controller



Bildet viser et vindu fra "Network Connections" på en Windows-datamaskin. Basert på det som vises:

1. **Ethernet:** Det ser ut som Ethernet-tilkoblingen er frakoblet fordi "Network cable unplugged" vises. Dette betyr at det enten ikke er en Ethernet-kabel koblet til datamaskinen, eller at det er en feil med kabelen eller porten den er koblet til.
2. **Wi-Fi:** Wi-Fi-statusen viser "Disabled", som betyr at Wi-Fi-tilkoblingen er deaktivert. Det er en "Enable"-knapp ved siden av, noe som indikerer at brukeren kan aktivere Wi-Fi ved å klikke på den knappen.

Årsakene til at Wi-Fi er deaktivert kan være mange:

- Brukeren kan ha deaktivert Wi-Fi manuelt.
- Det kan ha vært en feil eller en programvareoppdatering som førte til at Wi-Fi ble deaktivert.

Hvorfor har denne brukeren problemer

Bildegjenkjenning og GPT-V



Bildet viser en skadet Ethernet-kabel (RJ-45). Hvis brukeren prøver å koble til Internett ved hjelp av denne kabelen, vil den mest sannsynlige årsaken til at de ikke får tilkobling være skaden på kabelen. De indre lederne i kabelen ser ut til å være brutt eller skadet, noe som forhindrer en stabil datatilkobling. For å løse problemet bør brukeren erstatte den skadde kabelen med en ny, hel Ethernet-kabel.

Hvorfor får ikke

Tidslinje rundt LLM og andre innovasjoer

2021-06-11: GPT-3 Beta

2022-03-15: GPT-3 og Codex Public API

2022-06-21: Github Copilot

2022-11-30: ChatGPT lansert

Alle de store leverandørene
integrerer språkmodeller inn
i sitt økosystem

2023-01-16: Azure OpenAI-tjenester lansert

2023-02-06: Google Bard info

2023-02-07: Bing Search med ChatGPT lansert

2023-02-24: Meta introduserte LLaMa

2023-03-01: OpenAI med Whisper API

2023-03-14: ChatGPT med GPTv4 lansert

2023-03-15: Midjourney v5 lansert

2023-03-16: Microsoft 365 Copilot introdusert

2023-03-21: GPTv4 i Azure OpenAI-tjenester, Google Bard (Preview)

2023-03-23: ChatGPT-plugins annonsert, Bing Search Image

2023-03-30: BloombergGPT annonsert

2023-04-12: Dolly 2.0 lansert

2023-04-14: AWS CodeWhisperer

2023-04-19: NVIDIA – tekst-til-video

2023-05-10: Microsoft 365 Copilot – Utvidet Preview

2023-05-11: Google I/O (Bard, Søk, Bildesøk)

2023-05-15: OpenAI tilgjengelig gjør plugins in OpenAI

2023-06-13: ChatGPT Functions

2023-07-06: ChatGPT v4 API tilgjengelig

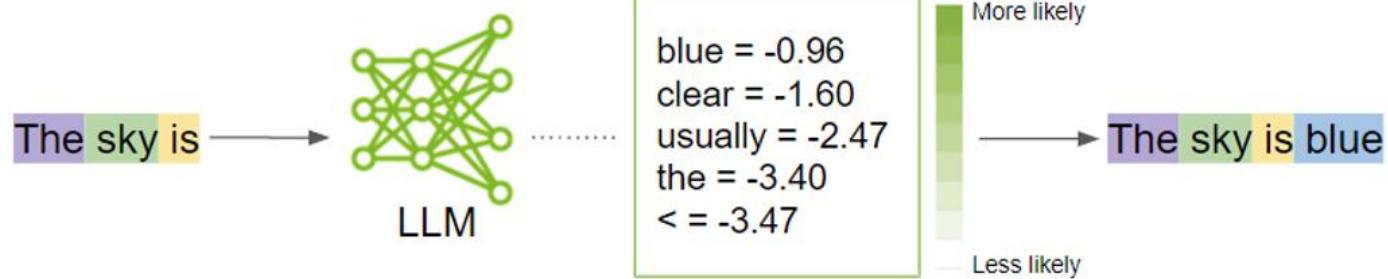
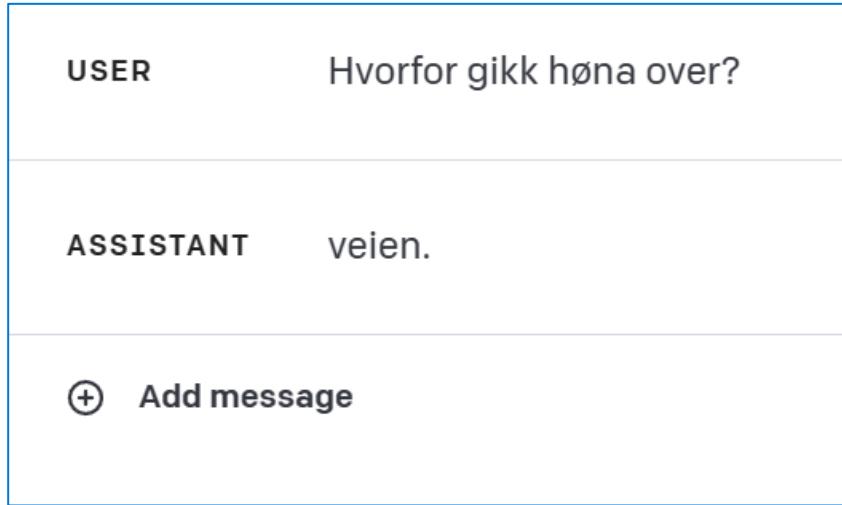
2023-07-16: Google Bard tilgjengelig i Norge

2023-07-18: LLaMa2 Lansert

2023-08-15: RuterGPT Lansert

2023-10-17: Fine-Tuning supporter på Azure OpenAI

Hvorfor er himmelen (blå)

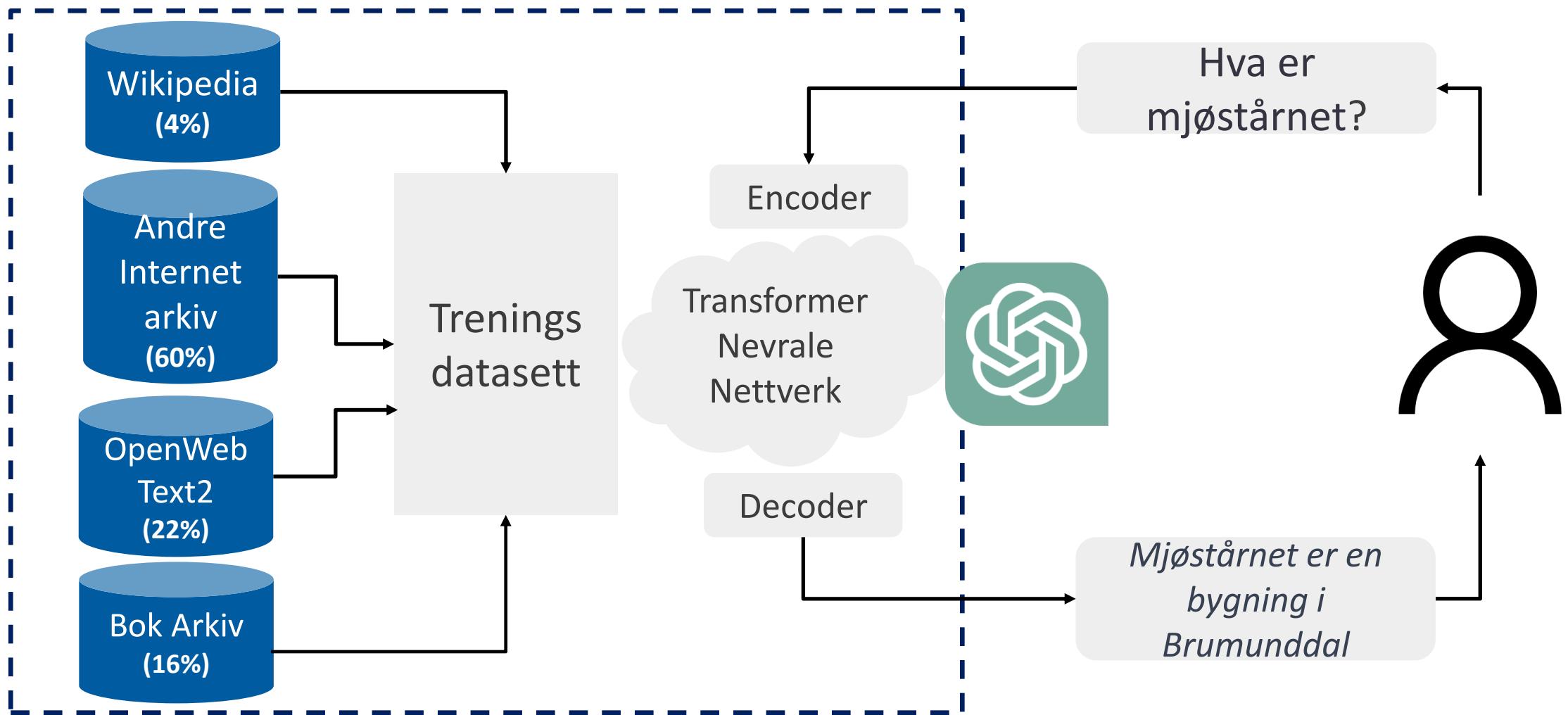


Leter etter Relasjoner mellom
ord, kontekst og betydninger

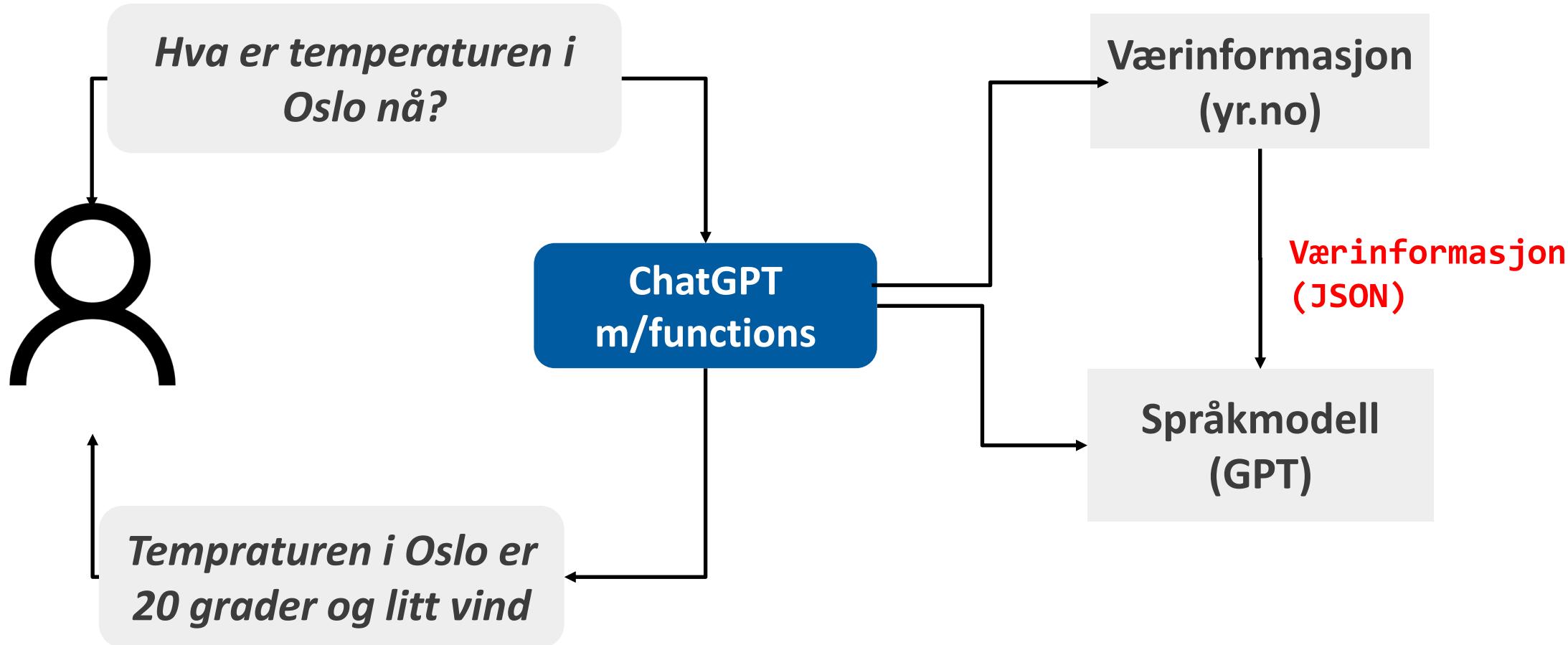
Henter ut semantikken

Datagrunnlaget i ChatGPT frem til Sep 2021 (OpenAI)

Januar 2022 for GPT 4



GPT og Functions



API Bruk – m/Python

```
import os  
import openai  
  
openai.api_key = "API Nøkler"  
  
response = openai.ChatCompletion.create(  
    model="gpt-4",  
    messages=[  
        {  
            "role": "system",  
            "content": "You are an AI assistant that helps the users with getting answers to their questions."  
        },  
        {  
            "role": "user",  
            "content": "What is the best way to get started with OpenAI?"  
        }  
    ],  
    max_tokens=256,  
    temperature=1,  
    frequency_penalty=0,  
    presence_penalty=0  
)  
print(response)
```

API Nøkkel fra OpenAI

Definere hvilken språkmodell

System: Beskrive kontekst til hvordan språkmodellen skal tolke informasjonen og respondere

Temprature: Tall fra 0 – 2 definerer hvor kreativ eller faktabasert modellen skal være

Bruksområder for språkmodeller

- Formuleringsstøtte & Dokumentasjon
- Intern Virtuell Assistent
- Brukerstøtte Chatbot
- Ekstern Datainnsamling & Tolking
- Innholdssammendrag & Generering
- Opplæring & Kompetanseutvikling
- Kode forklaring (Script deteksjon)
- Automatisering av Oppgaver & Prosesser
- **Rett å slett å komme i gang...**

Eks: med bruk av **OpenChat** og innhenting av nettinnhold for en chatbot

Hvordan kontakter jeg Sopra Steria?

Du kan kontakte Sopra Steria ved å sende en e-post til firmapost.no@soprasteria.com eller ved å kontakte din kundeansvarlig.

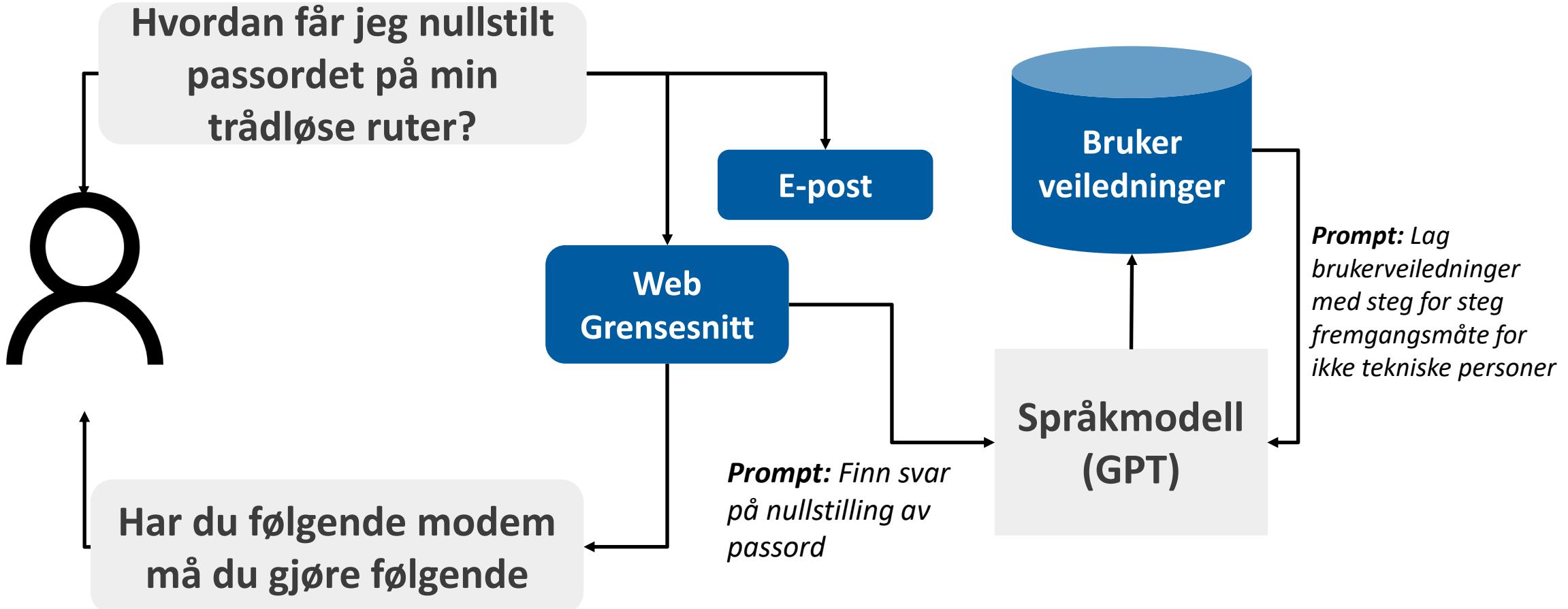
Hvilket type konsulentfirma er Sopra Steria?

Sopra Steria er et konsulentfirma innen digitalisering, innovasjon og bærekraft.

Bot · now

Gi sluttbrukere et enklere grensesnitt mot informasjon

Operere som både ChatBot og søkefunksjon mot data



Automatisk generering av dokumentasjon

The screenshot shows a dark-themed Table of Contents page. At the top left is a yellow icon of a book labeled "Table of Contents". Below it is a list of ten items, each with an icon and a link:

- [Table of Contents](#)
- [Overview](#)
- [Features](#)
- [Project Structure](#)
- [Modules](#)
- [Getting Started](#)
- [Roadmap](#)
- [Contributing](#)
- [License](#)
- [Acknowledgments](#)

A blue callout box with white text is positioned on the right side of the page, containing the following text:

Leser kildekode og lager automatisk tilhørnede dokumentasjon

[eli64s/readme-ai:](#)

Begrensningene med språkmodeller



koble deg til denne nettsiden og skriv en oppsummering av artikkelen
<https://www.ivanti.com/blog/cve-2023-35078-new-ivanti-epmm-vulnerability>



Beklager, men jeg kan ikke koble meg direkte til eksterne nettsider. Imidlertid, hvis du gir meg en beskrivelse eller kopierer innholdet fra artikkelen, kan jeg prøve å hjelpe deg med å oppsummere det.

Ikke Internett tilgang

Ikke tilgang til nyere informasjon

Kan ikke lære nye ting

Ingen kildevalidering

Ingen evne til å forstå større kontekst

Designet for å være chat basert

Annen funksjonaliteten fra OpenAI

Codex

Kan tolke og lage applikasjonskode.
Brukes i dag blant annet til **Github Co-Pilot**

Whisper

Tale-til-tekst.
Brukes allerede blant annet hos UiO for transkribering av forelesninger

DALL-E

Kan forstå og tolke naturlig språk og deretter **skape visuelt innhold** som svarer til beskrivelsen.

GPT

En samling av modeller som kan **tolke og generere naturlig språk.** (Det som brukes til ChatGPT)

CLIP (GPT-V)

Bildegjenkjenning som kan brukes til å beskrive innhold i bilder

Hva med sikkerheten rundt bruk av GPT?

Prompts vil ikke bli gjort tilgjengelig for andre brukere

Prompts kan bli brukt av OpenAI for videreutvikling

Spørninger via API **blir ikke** brukt men lagres i 30 dager

All data prosesseres på et datasenter i USA

Ikke tilgjengelig som egen tjeneste innenfor Europa

The screenshot shows a 'Settings' page with a navigation bar on the left containing 'General', 'Beta features', and 'Data controls'. The 'Data controls' tab is selected. A large blue callout box highlights the 'Chat history & training' section. This section contains the text: 'Save new chats on this browser to your history and allow them to be used to improve our models. Unsaved chats will be deleted from our systems within 30 days. This setting does not sync across browsers or devices.' Below this, there are three sections: 'Shared links' with a 'Manage' button, 'Export data' with an 'Export' button, and 'Delete account' with a red 'Delete' button.

Alternativer og økosystemet

Azure OpenAI

OpenAI
Språkmodeller

Google Vertex

Google sin
språkmodeller

AWS Bedrock

AWS sin
språkmodeller
+ Claude

LLaMa2

Utviklet av Meta
kommer i ulike
størrelser samt
optimalisert for
Chat

GPT4ALL

Lokal GPT med
støtte for en
rekke ulike open
source
språkmodeller

Mosaic

LLM Platform
som tilbyr både
ChatGPT
alternativ og
fine-tunings
mekanismer

Kan kjøres lokalt

Integrasjonsrammeverk

Langchain

LLaMaindex

Microsoft Semantic Kernel

Streamlit

Vertex AI Extensions

Chainlit

Norske alternativer?

ARENDA SUKA

NorGPT er en modell som ennå ikke har lært seg folkeskikk



Ruter As
8,843 followers
4d ·

...

+ Follow

I dag lanserer vi Norges første store språkmodell – RuterGPT. Hør hva vår leder for data science [Umair M.Imam](#) og adm dir. [Bernt Reitan Jenssen](#) har å si om denne spennende milepælen 🚀

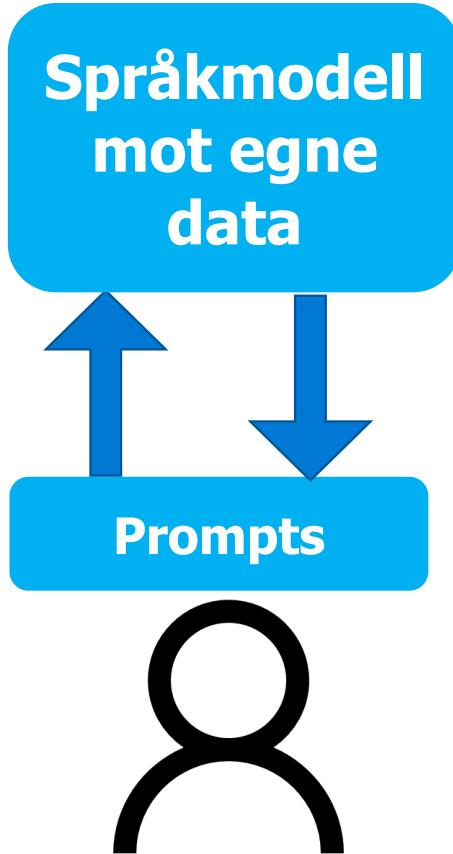
Bedre støtte
for Norsk
innhold

RuterGPT

LLaMa2 LLM

Kan vi trene språkmodeller med våre egne data?

- 1: Prompts
- 2: Fine-tuning
- 3: Grounding (RAG)



Prompt

Informasjon direkte inn i en prompt (begrenset data)

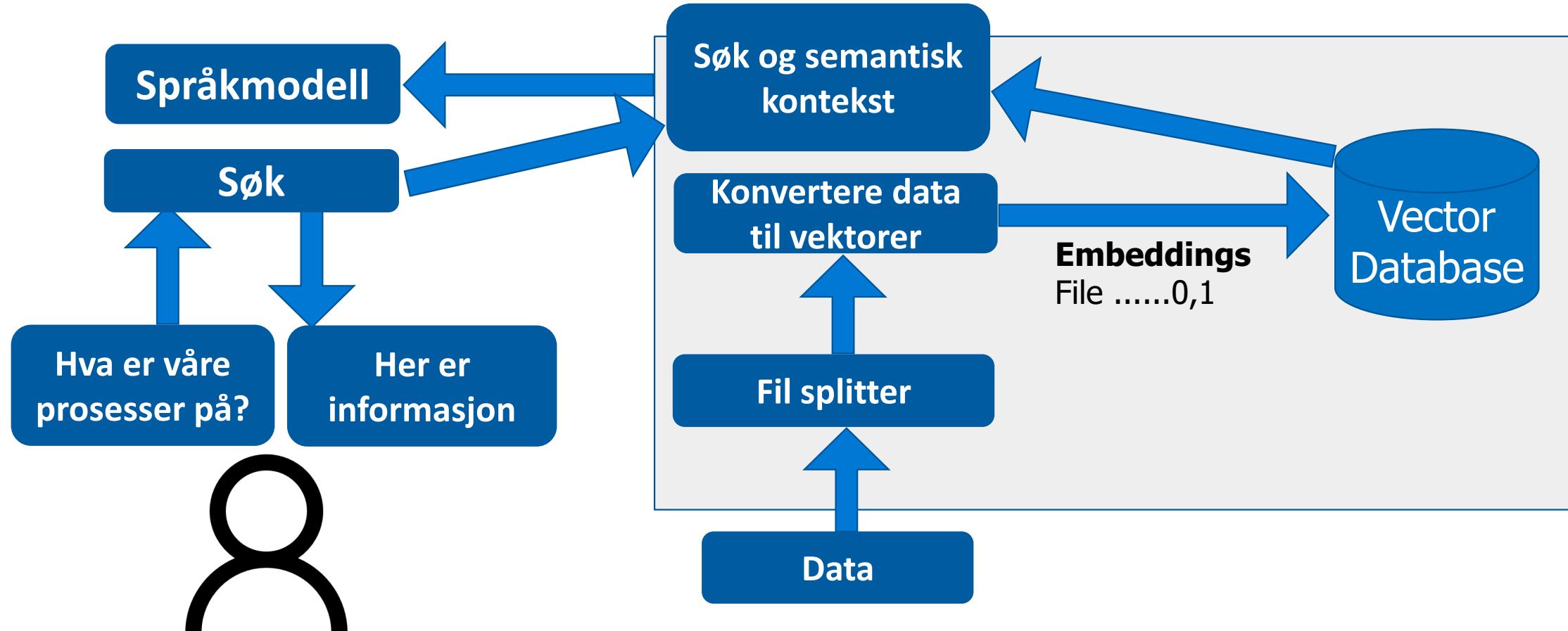
Fine-tuning

Trene opp språkmodellen på egne datakilder.

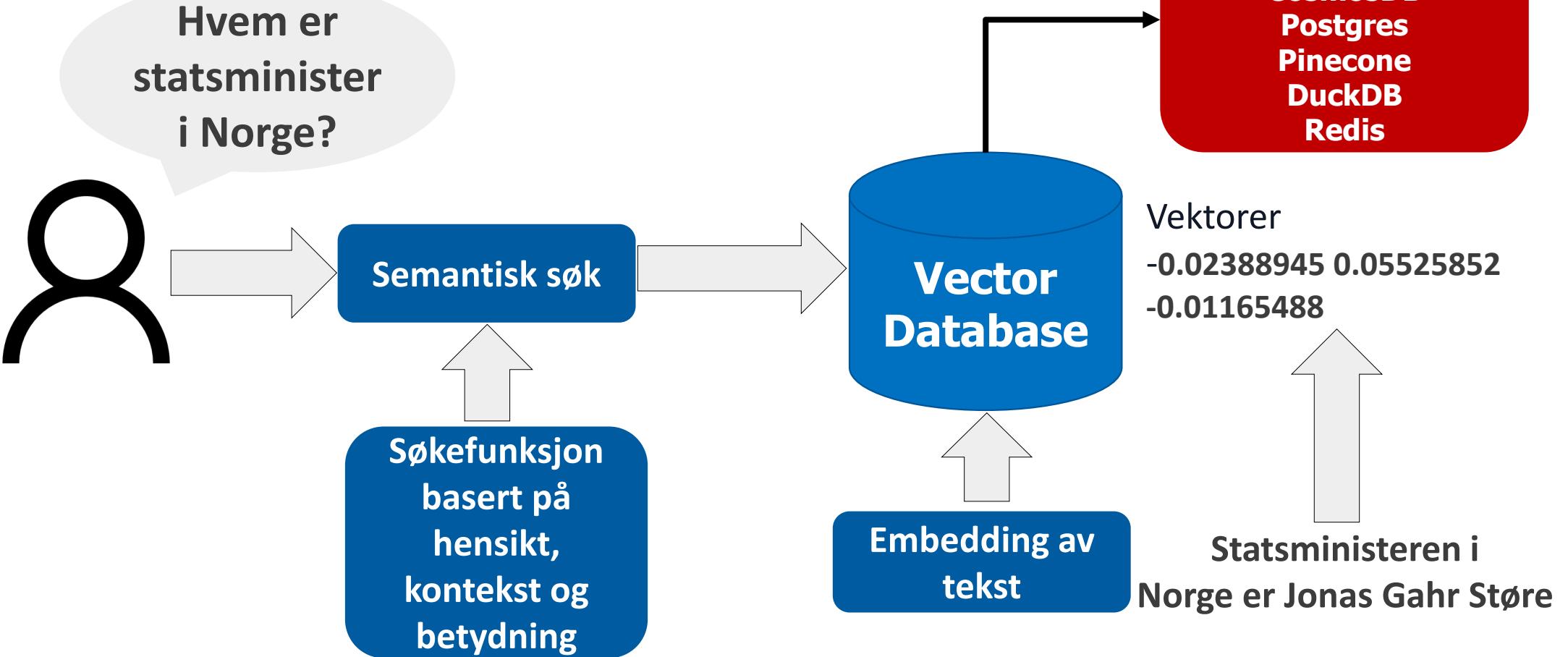
Grounding (RAG)

Tilgjengeliggjør data utenom språkmodellen. Bruker søk for å hente inn data

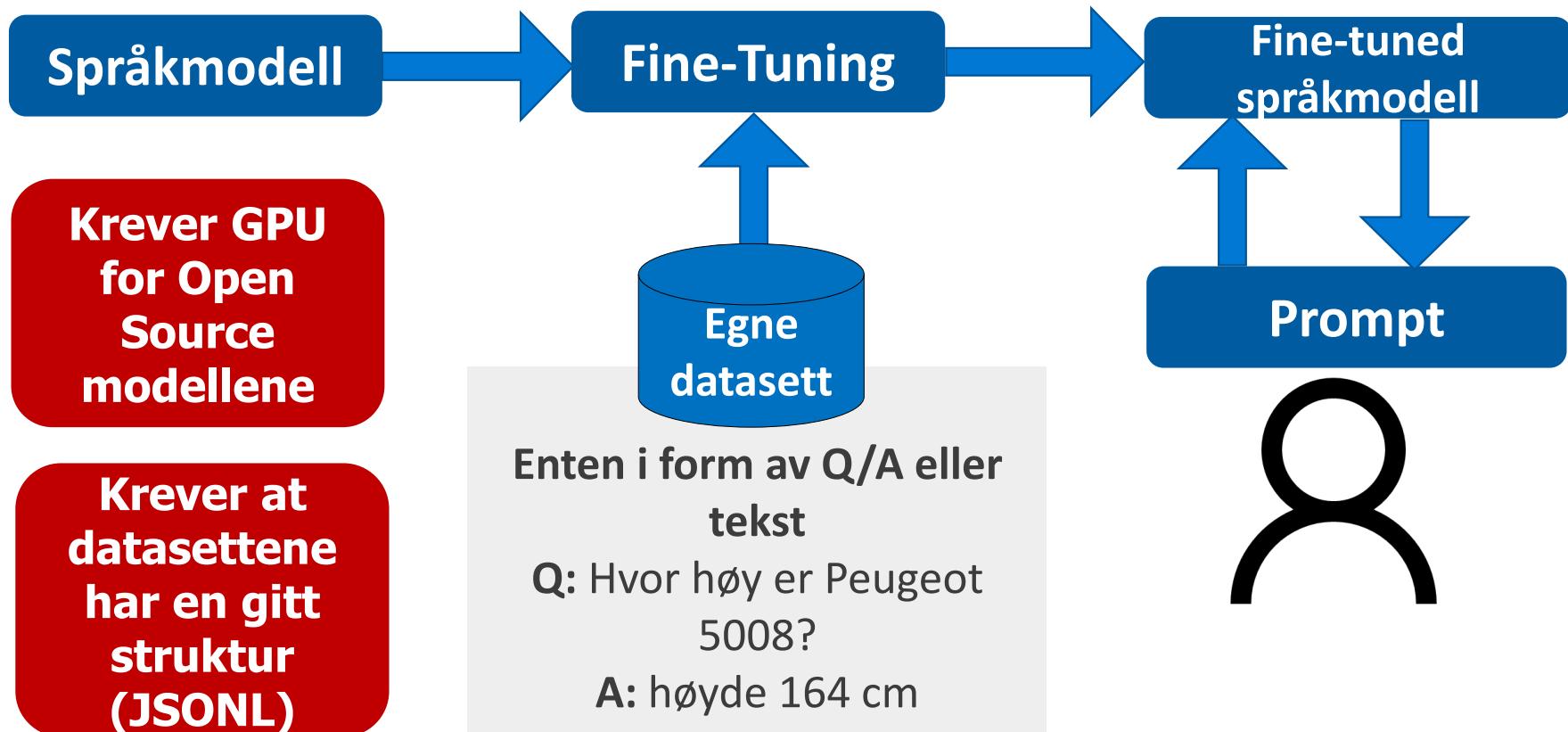
RAG (Retrieval Augmented Generation)



Vector databaser og semantisk søk

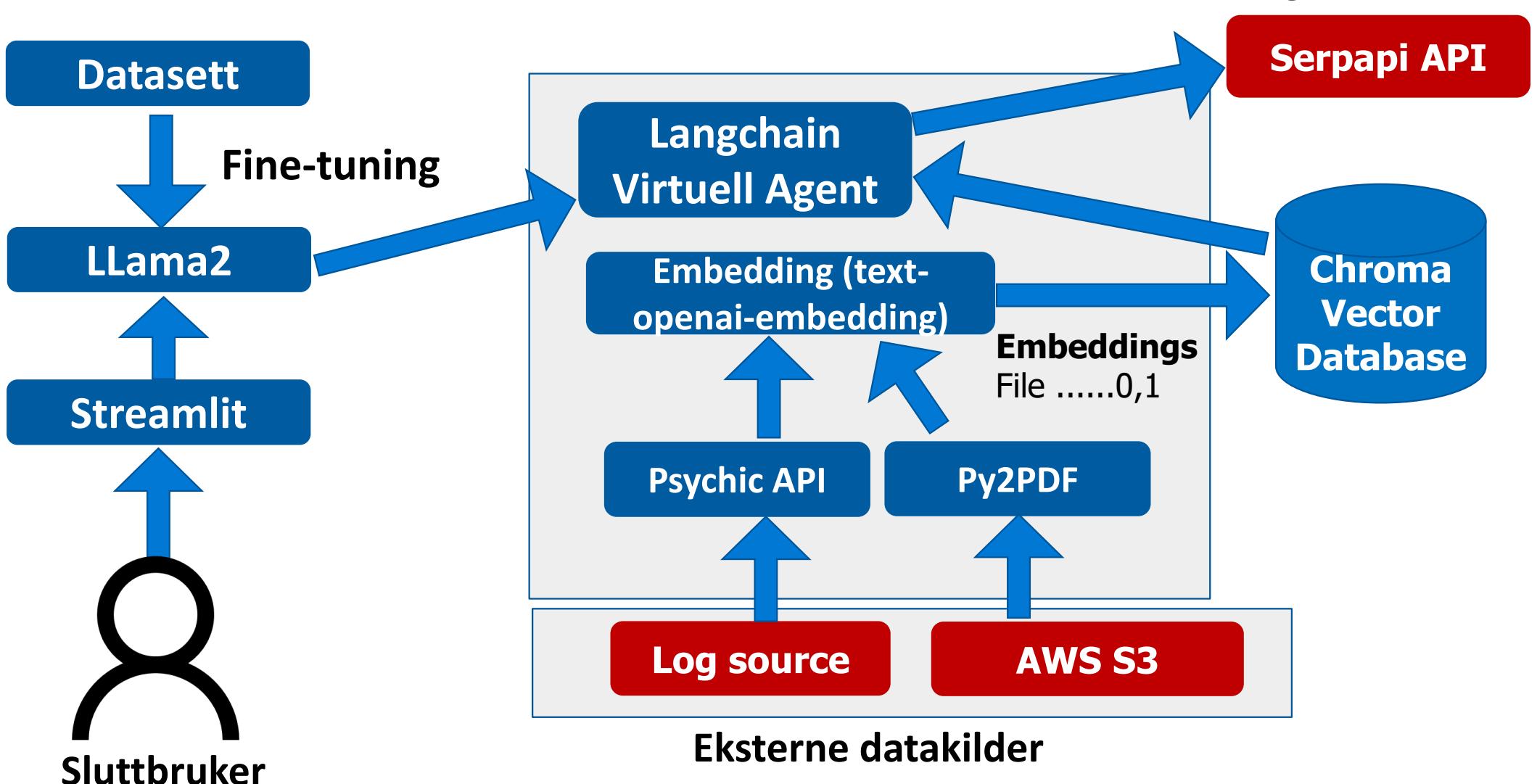


Fine-tuning



Eks på ferdige datakilder
<https://huggingface.co/datasets>

Eksempel løsning



Hva er Azure OpenAI?



**Microsoft administrert tjeneste
av OpenAI Språkmodeller**

**Mer fleksibilitet i forhold til
styring av nettverk samt
tilgangskontroll**

**Leveres fra Microsoft sine
datasentere**

**(Nesten..) Samme
funksjonalitet som OpenAI**

**Kan også søke unntak om 30
dagers logging**

Azure OpenAI

Model ID	Base model Regions	Fine-Tuning Regions	Max Request (tokens)	Training Data (up to)
gpt-35-turbo ¹ (0301)	East US, France Central, South Central US, UK South, West Europe	N/A	4,096	Sep 2021
gpt-35-turbo (0613)	Australia East, Canada East, East US, East US 2, France Central, Japan East, North Central US, Sweden Central, Switzerland North, UK South	North Central US, Sweden Central	4,096	Sep 2021
gpt-35-turbo-16k (0613)	Australia East, Canada East, East US, East US 2, France Central, Japan East, North Central US, Sweden Central, Switzerland North, UK South	N/A	16,384	Sep 2021
gpt-35-turbo-instruct (0914)	East US, Sweden Central	N/A	4,097	Sep 2021

Instruct = Model ikke designet for chat

Vil automatisk oppdatere til nyere versjon

Kan søke unntak om 30 dagers logging

1000 Prompts = 37,- i mnd m/GPT 3.5 (4K)

1000 Prompts = 977,- i nmd m/GPT 4 (8K)

Azure OpenAI

 Azure OpenAI

 Playground

 Chat

 Completions

 DALL-E (Preview)

 Management

 Deployments

 Models

 Data files

 Quotas

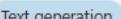
 Content filters (Preview)

Azure AI Studio

Welcome to Azure OpenAI service

Explore the generative AI models, craft unique prompts for your use cases, and fine-tune select models.

Get started

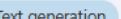
 Text generation



Chat playground

Design a customized AI assistant using ChatGPT. Experiment with GPT-3.5-Turbo and GPT-4 models.

[Try it now](#)

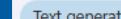
 Text generation



Completions playground

Experiment with completions models for use cases such as summarization, content generation, and classification.

[Try it now](#)

 Text generation



Bring your own data

PREVIEW

Connect and ground your data. Deploy to a web app or Power Virtual Agent bot (coming soon).

[Try it now](#)

 Text generation



DALL-E playground

PREVIEW

Create unique images by writing descriptions in natural language.

[Try it now](#)

[Deploy Azure OpenAI using Terraform with Private Endpoint - msandbu.org](#)

Azure OpenAI



Start chatting

This chatbot is configured to answer your questions

Type a new question...



Web applikasjon med ferdig integrasjon mot egne data

Bruker Azure tjenester for publisering og lagring av data

Kan settes opp med autentisering via Microsoft 365 brukerkontoer

Kan publiseres direkte som bot til Microsoft Teams

Azure OpenAI og Azure ChatGPT



FORNYINGS- OG
ADMINISTRASJONSDEPARTEMETET

Veileder

Veileder til reglene om
offentlige anskaffelser

Azure OpenAI og Azure ChatGPT



Azure AI

Når kan de ulike anskaffelsesprosedyrene benyttes?

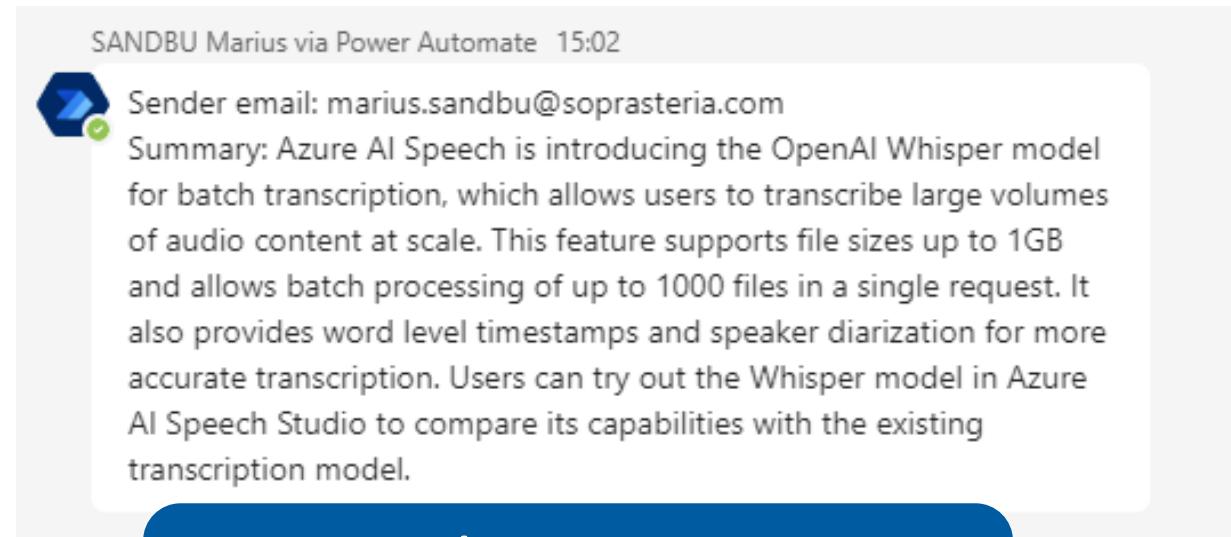
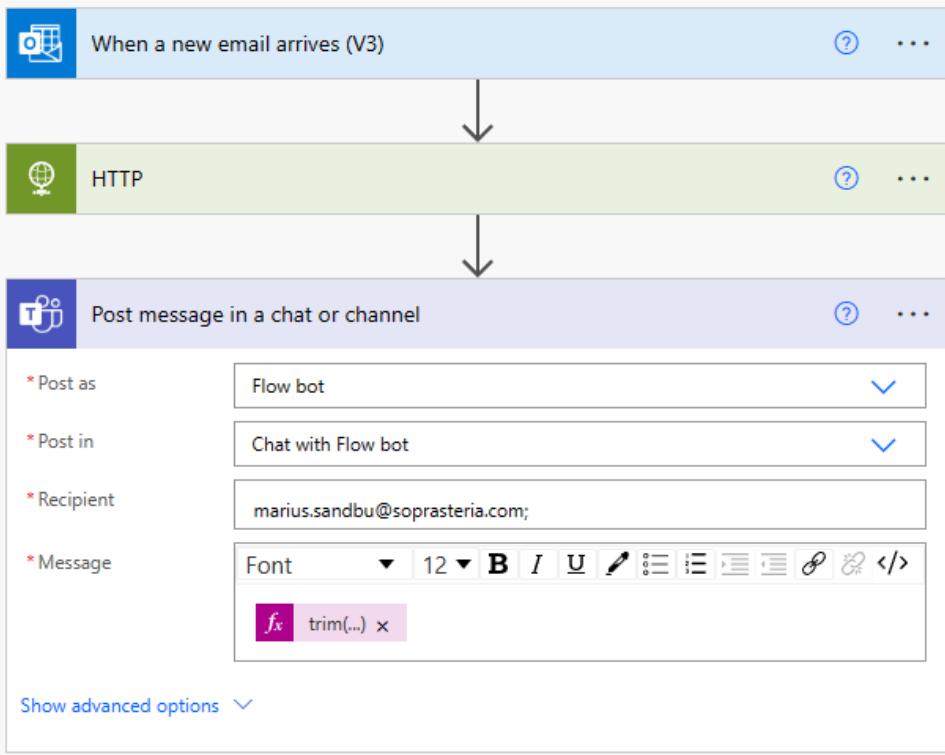
Her er en oversikt over når de ulike anskaffelsesprosedyrene kan benyttes:

- Anskaffelser under 500 000 kr kan velge mellom følgende prosedyrer dersom vilkårene i § 5-1 annet ledd er oppfylt:
 - Fremgangsmåten etter forskriftens del I
 - Åpen anbudskonkurranse
 - Begrenset anbudskonkurranse
 - Konkurranse med forhandling¹
- Anskaffelser under EØS-terskelverdiene og uprioriterte tjenester kan velge mellom følgende prosedyrer dersom vilkårene i § 5-1 annet ledd er oppfylt:
 - Åpen anbudskonkurranse
 - Begrenset anbudskonkurranse
 - Konkurranse med forhandling¹

Type a new question...



Andre integrasjoner og LLM funksjoner



**Logic App / Power Automate Kan
brukes til å lage predefinerte
arbeidsflyter**

Demo – Tabletalker

- Bruker Azure Speech Recognition, ElevenLabs, OpenAI with Functions, Langchain og SQLAlchemy
- 1# Analyserer tale som tekst ved bruk av Azure Speech Recognition
- 2# Tekst håndteres som inndata til Langchain agent-kontekst
- 3# Avhengig av nøkkelord vil en OpenAI-funksjon brukes
 - For eksempel – db.chain.run for SQL-relaterte forespørsler
 - For eksempel – search.run for Google-søk
- 4# SQL-spørsmål vil bli sendt til SQL-database
- 5# Utdata tolkes ved bruk av Elevenlabs
- 6# Kontekst tilbake til brukeren



Microsoft Copilot

GPT Virtuell Assistent i Microsoft 365

Integrert i Office

Lisens på 30\$ Per bruker i måneden (minimum 300)

Tilgjengelig fra 1 November (Ingen Trial funksjon)

Gir GPT integrasjon mot **egne data** i Microsoft 365

Bruker Azure OpenAI for LLM operasjoner

Delvis støtte for norsk ved lansering

Full norsk støtte kommer lengre

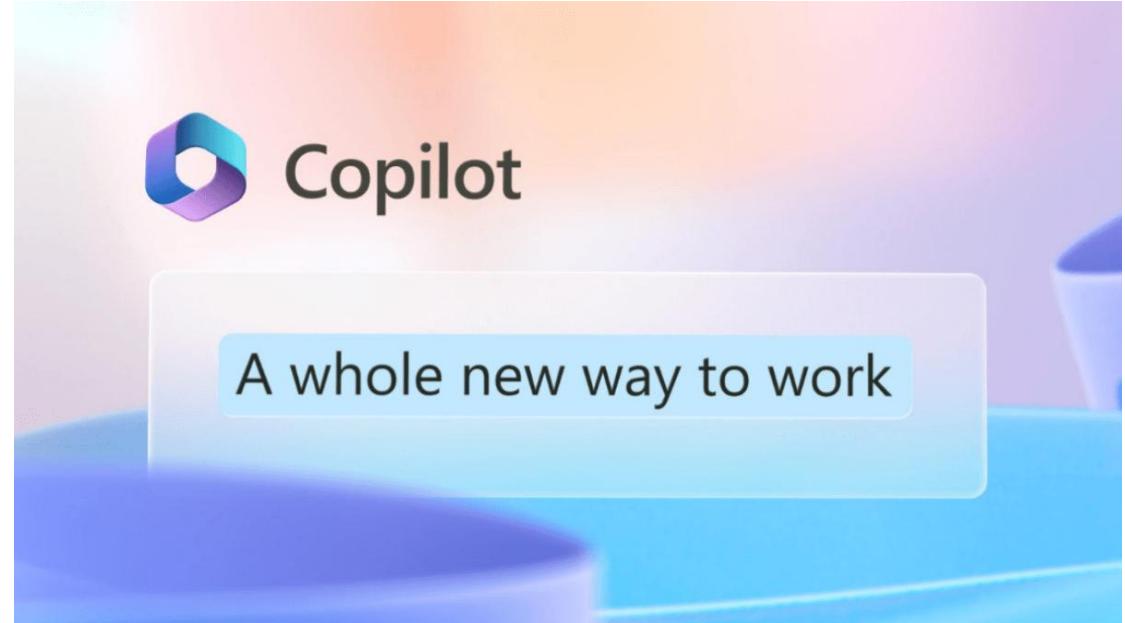
Kan også utvides med **3.parts datakilder**

Service Now (Ferdig integrasjon)

Confluence (Ferdig integrasjon)

Filservere (Ferdig integrasjon)

....andre data kilder løsninger



Hvordan funker det?

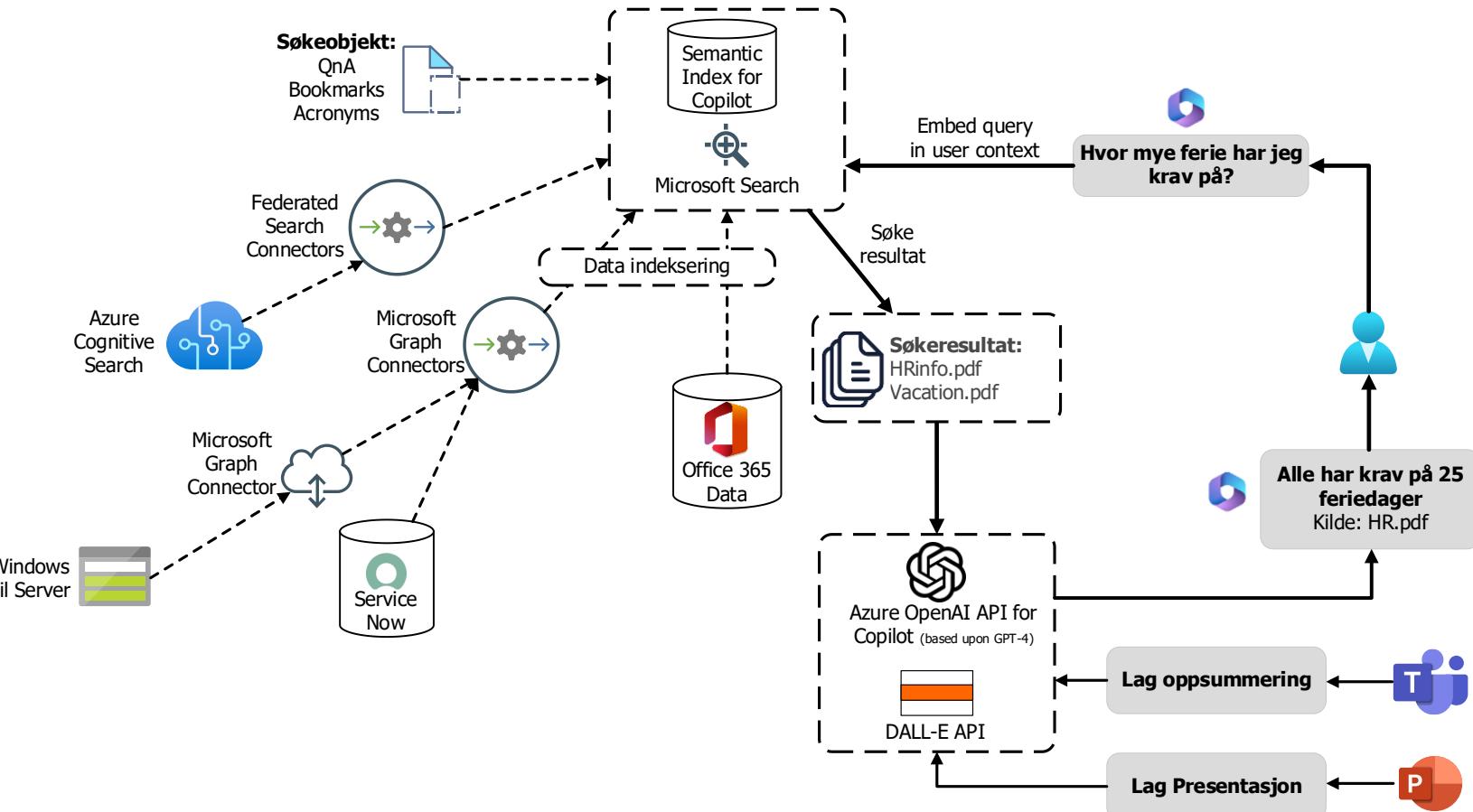
All data og innhold blir indeksert i en felles katalog

Data og innhold blir indeksert basert på semantisk søk

Data blir indeksert per bruker (Semantic Index)

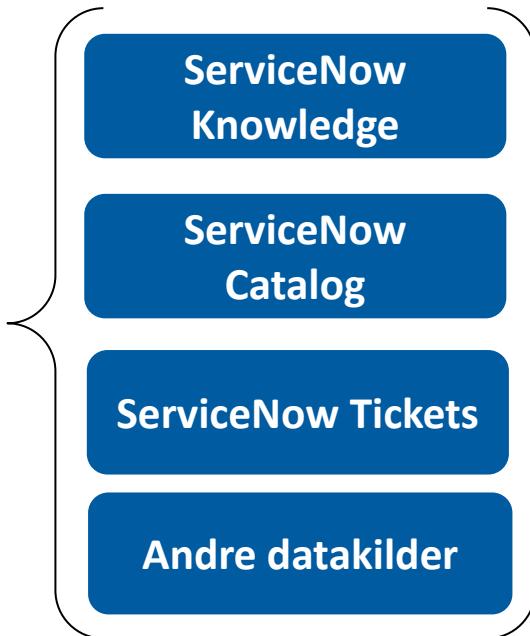
Data fra 3.parts kilder kan også integreres

CoPilot bruk fra enten Teams eller Office applikasjoner

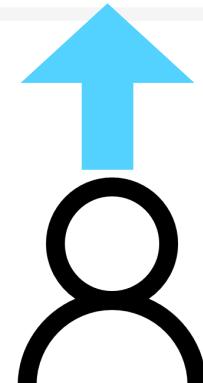
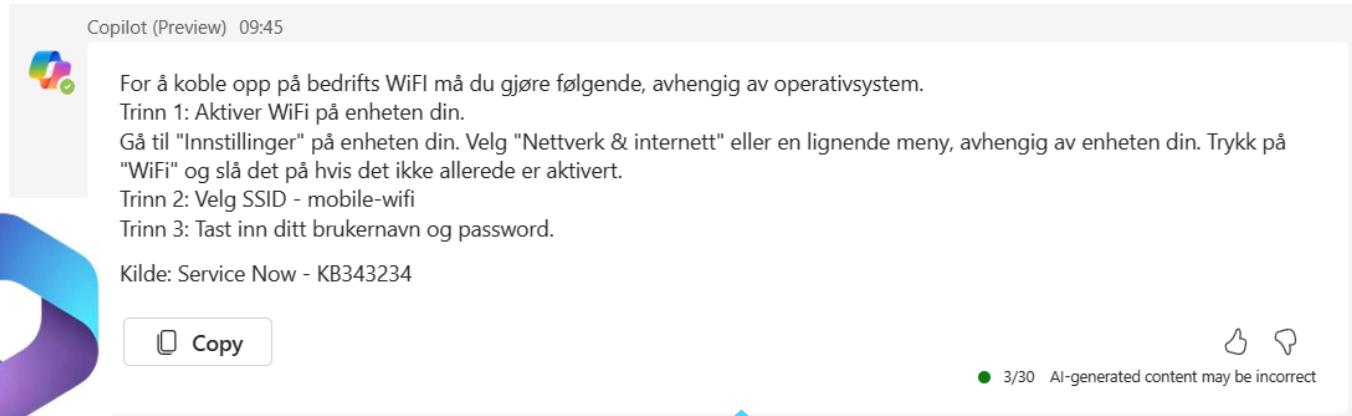


Integrasjonsmuligheter mot 3.Parts datakilder

Microsoft Graph Connectors



Microsoft 365 Copilot



«Hvordan
kobler jeg til
Wifi?»

Data indeksering og 3.parts datakilder

Kan lage egne
integrasjoner som kan
brukes for indeksering
av data

Hver organisasjon kan
ha opptil 30
integrasjoner (eksterne
kilder)

Bookmarks
QnA
Akronymer

The screenshot shows the Microsoft 365 Admin center interface for a user named 'Copilot User1'. The 'Licenses and apps' tab is selected. A message box at the top states 'Semantic indexing is complete.' Below it, a dropdown menu shows 'United States'. The 'Licenses (2)' section lists the following licenses:

- Azure Active Directory Premium P2**
100 of 100 licenses available
- Communications Credits**
Unlimited licenses available
- Exchange Online (Plan 1)**
1 of 1 licenses available
- Microsoft 365 Business Basic**

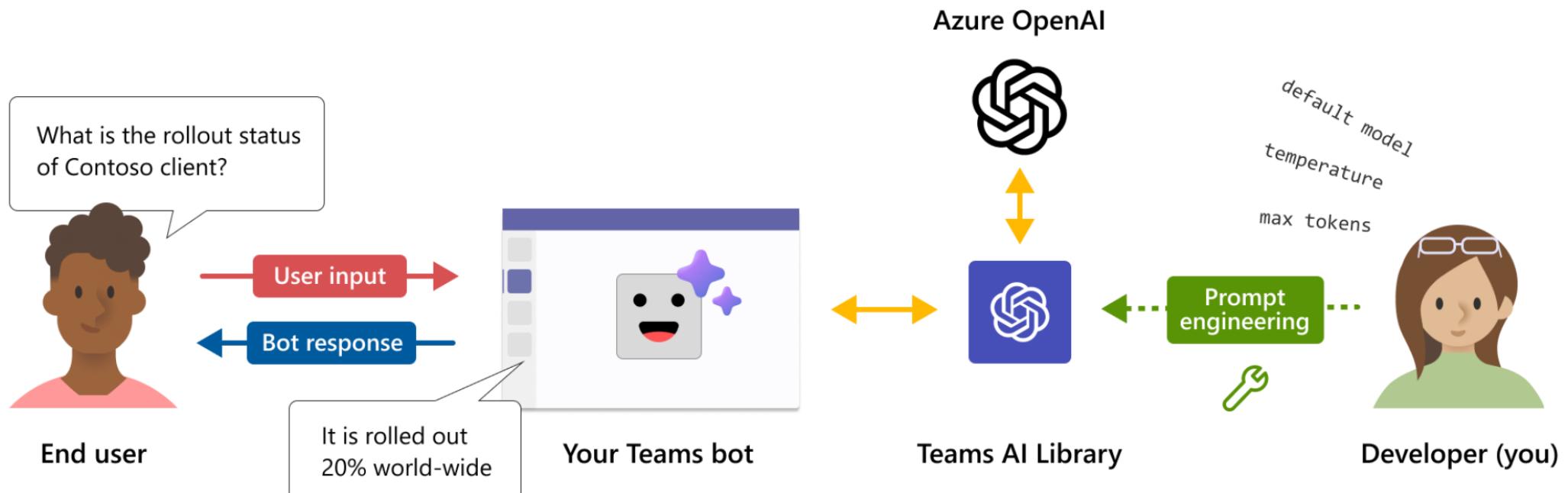
Mulighetene fremover!

Teams AI
bibliotek

Lag egne språkmodell
aktiverte boter

Med ulike integrasjoner mot
3.parts kilder

Krever ikke
Copilot 😊





DEMO

Tabletalker og andre snacks

Alternativer

[microsoft/azurechat: 🚧 💼 Azure Chat Solution Accelerator
powered by Azure Open AI Service \(github.com\)](#)

[microsoft/gpt-review \(github.com\)](#)

[microsoft/chat-copilot \(github.com\)](#)

[microsoft/semantic-kernel: Integrate cutting-edge LLM
technology quickly and easily into your apps \(github.com\)](#)

[microsoft/autogen: Enable Next-Gen Large Language Model
Applications. Join our Discord: <https://discord.gg/pAbnFJrkZ>
\(github.com\)](#)

[microsoft/promptflow: Build high-quality LLM apps - from
prototyping, testing to production deployment and monitoring.
\(github.com\)](#)

SIEM integrasjon

Berikelse av informasjon

Innhenting av ekstern informasjon

Automasjon via Azure OpenAI API



Comment created from playbook - gptcomplete 19.09.23, 12:02



The alert indicates that there was a failed logon attempt from multiple external IP addresses to various user accounts, including \ADMINISTRATOR, \admin, \test, \user, \scan, \pc, \AZURE, \AZUREADMIN, \AZUREUSER, \SUPERADMIN, and \ADMINUSER. The IP addresses involved in the logon attempts are 188.132.130.62, 45.143.201.62, 186.96.215.69, 206.119.117.84, and 198.211.97.72. The logon attempts occurred on the host named "vm01."

Based on the eventID 4625 and the nature of a network logon, it suggests that the failed logon attempts were likely a result of a brute force attack over RDP (Remote Desktop Protocol).

Hvordan ligger Microsoft an i forhold til de andre?

Features	Google	Microsoft	Amazon Web Services
LLM Service/Runtime	Vertex AI	Azure OpenAI	Bedrock
LLM Models available	PaLM, LLaMa2, Falcon, Claude2*	GPT, LLaMa2, Falcon, Databricks Dolly	Titan, Claude2, Cohere
LLM Models Code	Code-Bison	Codex	
LLM Models Security	Sec-PaLM	Tbd*	
LLM Models Catalog	Model Garden	Model Catalog	Model Providers
LLM Token Size FM	32k (PaLM2)	32k (GPT4)	8k (Titan)
LLM Availability			
LLM Integration framework	Vertex AI Extensions	Microsoft Semantic Kernel	
LLM Safety filter		Azure AI Content Safety	
LLM Fine-tuning support	Code-bison(PaLM), text-bison(PaLM)	GPT-3.5 supports Fine-tuning but not in Azure yet	
LLM Agent	Vertex AI Conversation	Power Virtual agents	Bedrock Agents, Amazon Lex

[gpt-ai/cloudgpt.md at main · msandbu/gpt-ai \(github.com\)](https://gpt-ai/cloudgpt.md at main · msandbu/gpt-ai (github.com))

Hvordan ligger Microsoft an i forhold til de andre?

Features	Google	Microsoft	Amazon Web Services
Vector database	Cloud SQL (Pgvector), AlloyDB, Vertex AI Vector Search	Azure Cosmos DB	Amazon RDS (Pgvector)
Embedding	Text embedding API Gecko	Ada OpenAI Embedding	Titan Embeddings
Integration services - Langchain	Vertex AI, Google Search	Azure Cognitive Search	
Code assistant-based AI	Duet AI	Github Copilot	Amazon Code Whisperer
Collaboration GPT	Duet AI	Microsoft Copilot	
Digital Watermarking	Synthid (Image)		
Security Powered LLM	Security Command Center AI	Microsoft Security Copilot	

[gpt-ai/cloudgpt.md at main · msandbu/gpt-ai \(github.com\)](https://gpt-ai/cloudgpt.md at main · msandbu/gpt-ai (github.com))

Andre interessante tillegg

ChatGPTWriter

Tillegg til nettleser for å kunne gi GPT funksjonalitet direkte

Elevenlabs.io

AI Tekst-til-tale programvare, har også mekanismer for kloning av stemmer

HyperWrite

Addon til nettleser som kan automatisk oppsummere innhold og har rekke integrasjoner

Dify

Platform for å lage bots/virtuelle assistenter

OpenAGI

Bilde Analyse. Kan gjøre om bilder til tekst med beskrivelse

Jarvis

Integrasjon mellom språkmodeller og ML operasjoner

Auto-GPT

Agent basert tilnærming for utføring av oppgaver

Gpt-llm-trainer

Enklere måte å lage datasett for trening samt kjøre fine-tuning

Prompttools

Verktøykasse for å teste språkmodeller, prompts og vektordatabaser

PrivateGPT

Lokal installasjon av GPT basert på GPT4All, Langchain og LLaMA

GPTEngineer

Virtuell kode assistent som bygger apper basert på instruksjer

ChatGPT Retrieval

For de som ser på å integrere ChatGPT mot egne data (fra OpenAI)

Og noen til....

Bellzebub

Honeypot basert på GPT som emulerer en Linux Host

Selefra

Verktøy som bruker OpenAI til å analyse og se etter sårbarheter i Sky miljø (AWS, Google, Azure)

HyperWrite

Addon til nettleser som kan automatisk oppsummere innhold og har rekke integrasjoner

PentestGPT

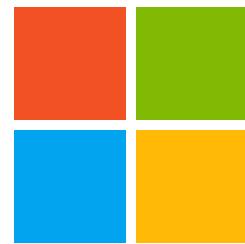
Pentest rammeverk som bruker LLM for å oppsummere og lage oppgaver

AgentGPT

Agent basert tilnærming for utføring av oppgaver

PrivateGPT

Lokal installasjon av GPT basert på GPT4All, Langchain og LLaMA



Microsoft

Evidi

?



MVP-Dagen 2023





Tusen takk!

MVP-Dagen