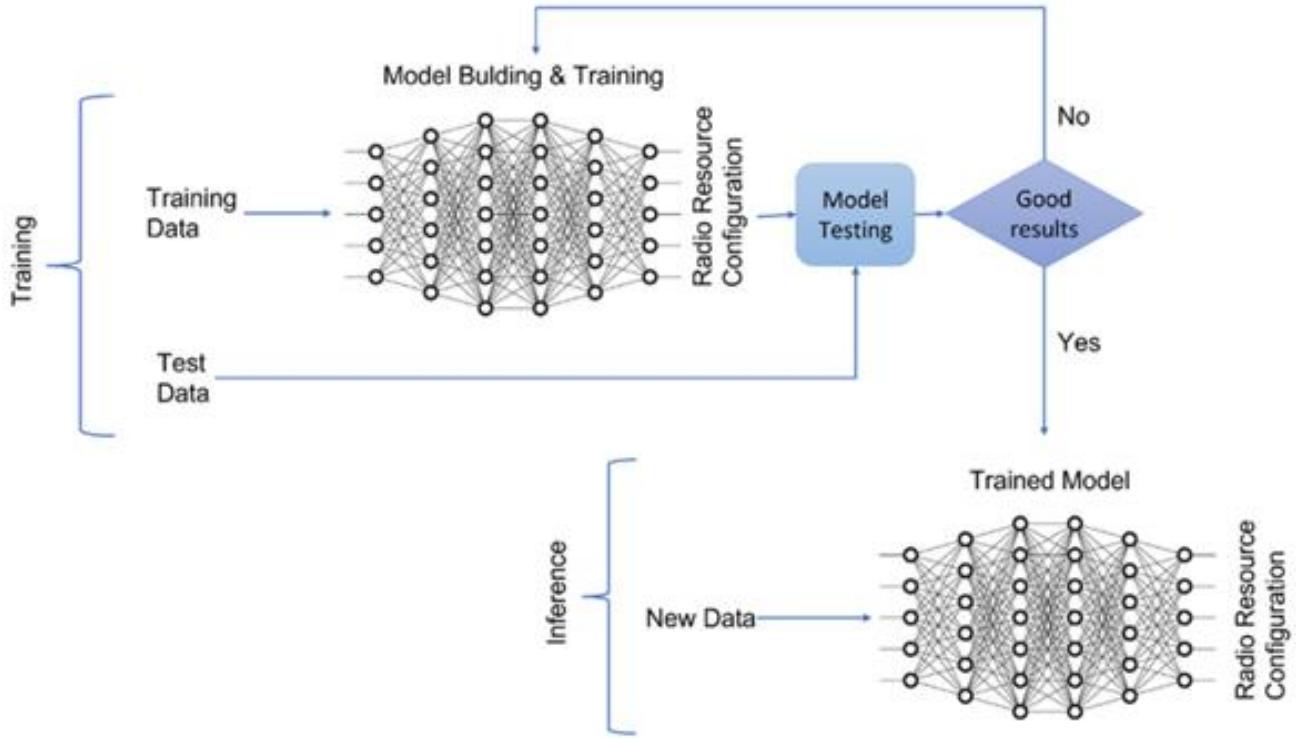
”Osäker” Inferens?

Presentation Sambruk, 2026-02-13
Begränsad spridning

CivSec

Agenda

- Vi ska prata om Inferens och modeller



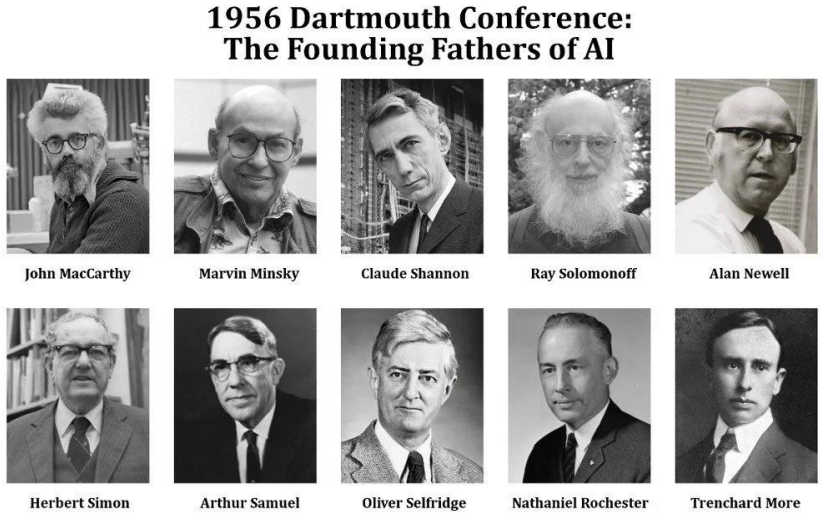
CivSec

CivSec

- ”Nytt” bolag, men ”gammalt” och erfaret gäng
- Fokus mot offentliga, samhällsviktig och samhällskritisk verksamhet och deras underleverantörer
- Cybersäkerhet, GRC, PSIRT, Sovereign Secure AI, Övning, träning och utbildning
- Fokus på egen rådighet och nationell suveränitet

AI Historik

- 1956: Dartmouth-workshop – begreppet AI skapas
- AI-vinter: 1970 -1980, 1990-2000 – begränsad framgång
- 2010 -: Big data + GPU → Deep Learning & LLM
- 2020 -: Generativ AI (ChatGPT, DALL-E, etc.)



CivSec.

Vad är AI?

- Maskin som ”efterliknar” mänsklig intelligens.
- Kategorier /definitioner:
  - ML – tränas på data, klassificering, regression
  - AL – algoritmiska beslut utan träningsdata
  - Gen-AI – skapar ny text, kod, bild, ljud
  - AGI – hypotetisk fullständig AI (Finns inte!)
  - Neurala nätverk – maskinlärningsmodell med lager
  - Agenter – ”självständiga” beslutssystem/funktioner (t.ex. bots)
  - Hyperautomation – automatisering + AI för hela affärsprocesser
  - RAG – Retrieval-Augmented Generation (sök + generera)

LLM / Språkmodeller, spoiler

- Sannolikhet
- Statistik

- Ingen magi
- Ingen förståelse

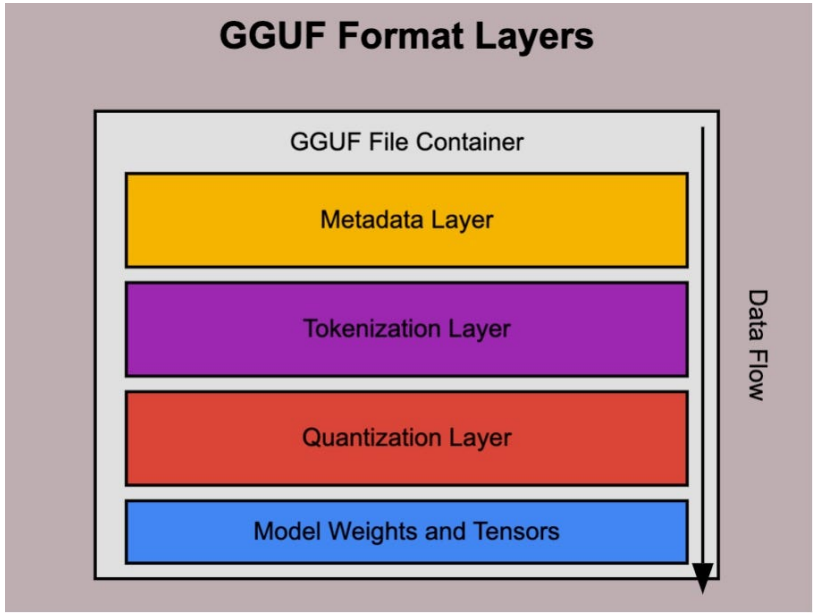
- Modellen är egentligen en ”dum” container-fil.

CivSec.

Vad är en modell fil egentligen?

En modellfil (GGUF, ONNX) innehåller:

- Modellens arkitektur (lager, dimensioner)
- Viktmatriser (miljarder flyttal)
- Eventuella metadata
- Tokenizer-information (ibland separat)

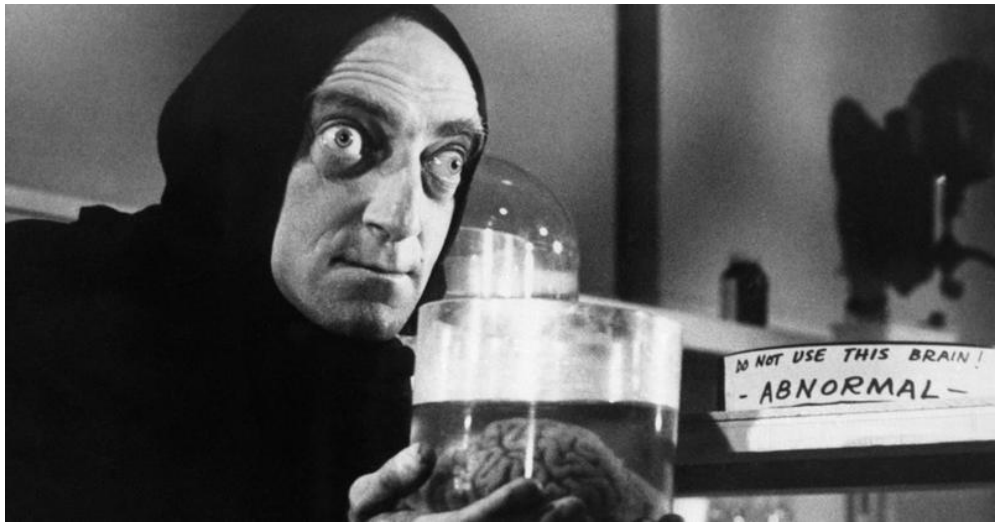


CivSec.

Vad är en modell fil egentligen?

Den innehåller INTE:

- Exekverbar kod
- API-nycklar
- Användardata
- Automatisk nätverkskommunikation
- Självständigt beteende



En modell fil är som en fryst hjärna i en burk.
Den tänker inte förrän någon kopplar in den i en maskin.

CivSec.

Hur ser en modell fil ut internt? (GGUF & ONNX)

- GGUF (typiskt för llama.cpp)
  Består av:
- Header
  - Version
  - Modellnamn
  - Tensor-count
- Metadata
  - Arkitekturtyp (LLaMA, Mistral etc)
  - Kontextstorlek
  - Kvantiseringstyp
- Tensorblock
  - Viktmatriser
  - Bias-vektorer
  - Attention-weight-matriser
  - Lager för lager

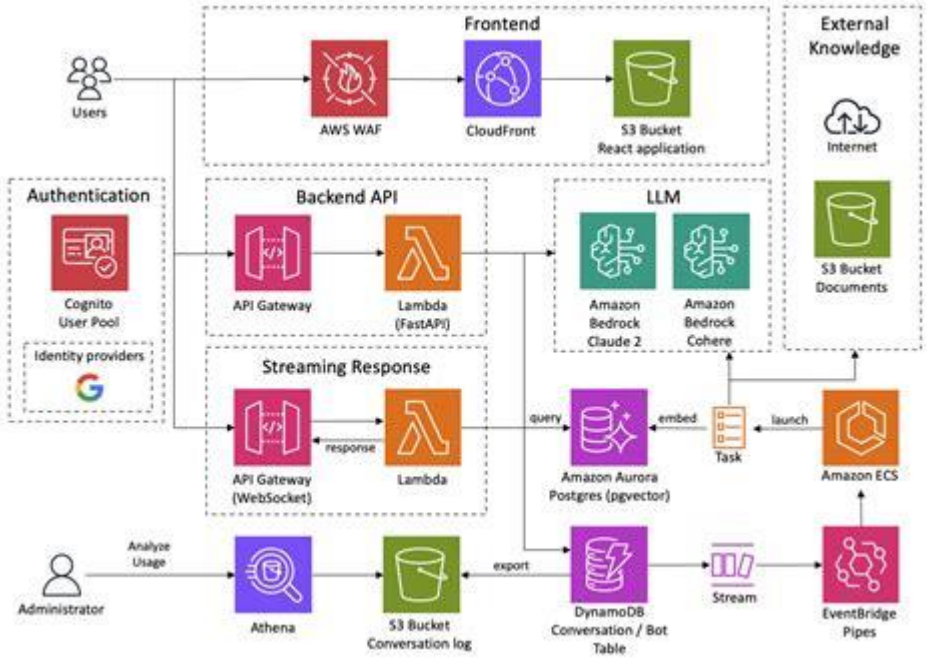
- ONNX är mer generiskt.
  Består av:
  - Computational graph
  - Noder (MatMul, Add, Softmax)
  - Tensorer
  - Parametrar
- ONNX beskriver:
  Ett matematiskt beräkningsflöde.

Men:
Det är fortfarande data
Det är inte exekverbar kod i sig
Det kräver en ONNX runtime

Vad saknas i en modell fil?

En modell fil saknar:

- HTTP-server
- API-logik
- Autentisering
- Loggning
- Databaskoppling
- Verktygsanrop (tools)
- RAG-system
- Prompt templates
- Användarsessioner



Den saknar ALLT som gör ett system till ett system.
Den är bara vikter + arkitektur.

CivSec.

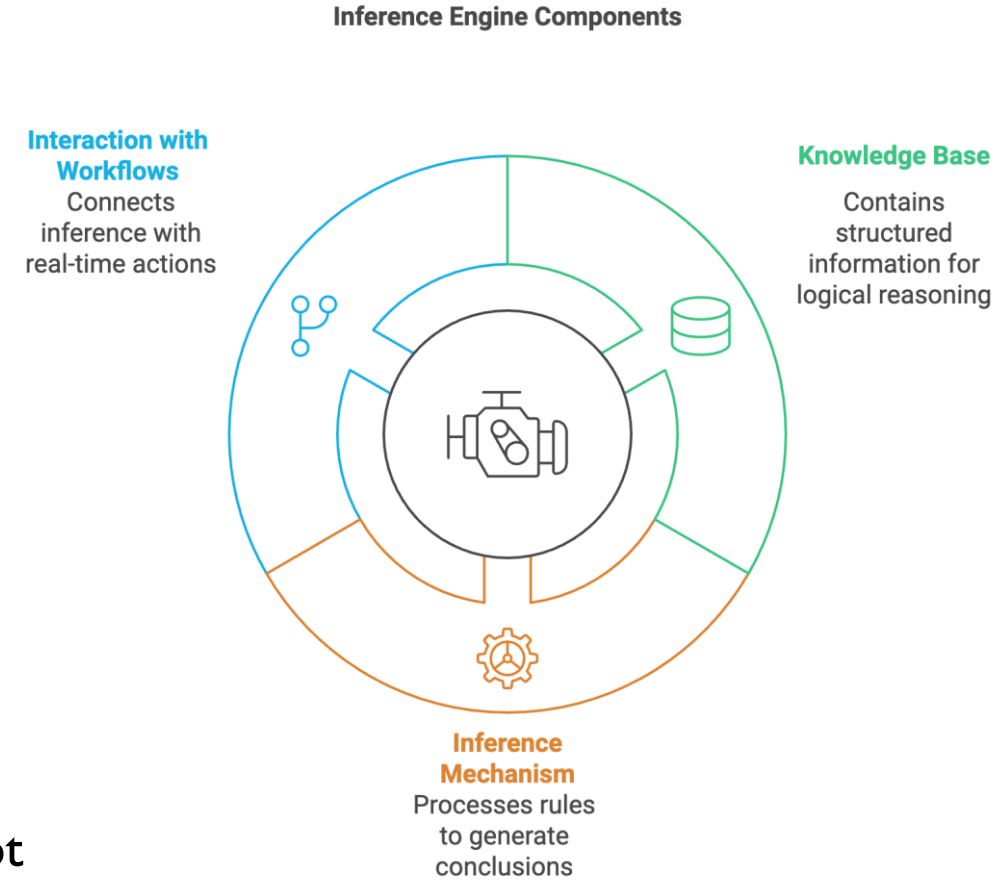
Vad är en inferensmotor?

En inferensmotor är t.ex.:

- llama.cpp
- Ollama
- vLLM
- ONNX Runtime
- TensorRT
- PyTorch runtime

Den gör tre saker:

- Läser modellfilen
- Laddar vikterna i minnet
- Kör matematiska operationer på en prompt



CivSec

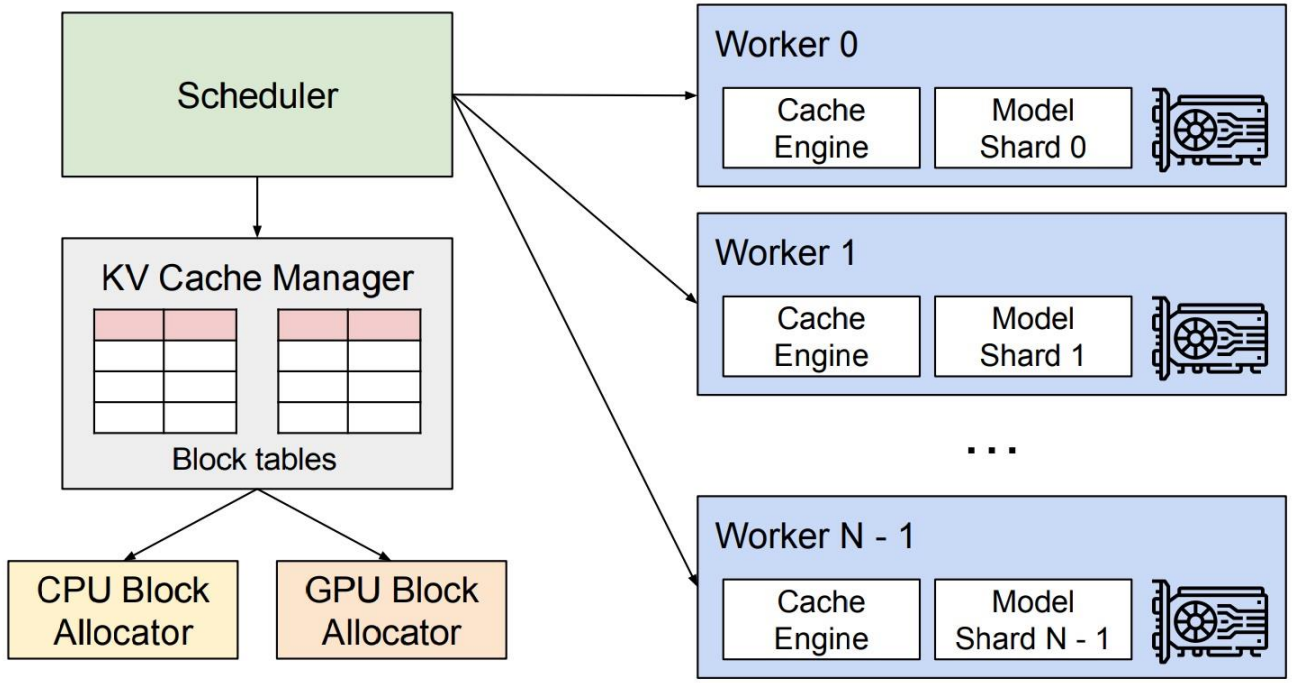
Vad händer när modellen laddas?

Förenklat flöde:

1. Modellfil öppnas
2. Tensorer mappas till RAM/VRAM
3. Beräkningsgraf initieras
4. Systemet väntar

Viktigt:

Den gör ingenting.
Den exekverar inget.
Den väntar bara.



CivSec

Vad skickas in i inferensmotorn?

När systemet kör:

Det ENDA som skickas in är:

- Prompt-text (tokeniserad)
- Eventuellt systemprompt
- Eventuell kontext (tidigare text)

Inget annat.

Ingen:

- Databasdump
- API-nyckel*
- Systemhemlighet
- Miljövariabel
- Backendlogik

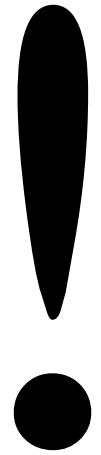
Slutsats

Inferensmotorn får:

- En sekvens av token-ID:n.

Den gör:

- Matematisk multiplikation + softmax.



Det finns ingen magi.
Ingen automation.
Ingen självständig insamling.
Ingen "läckning" från filen.

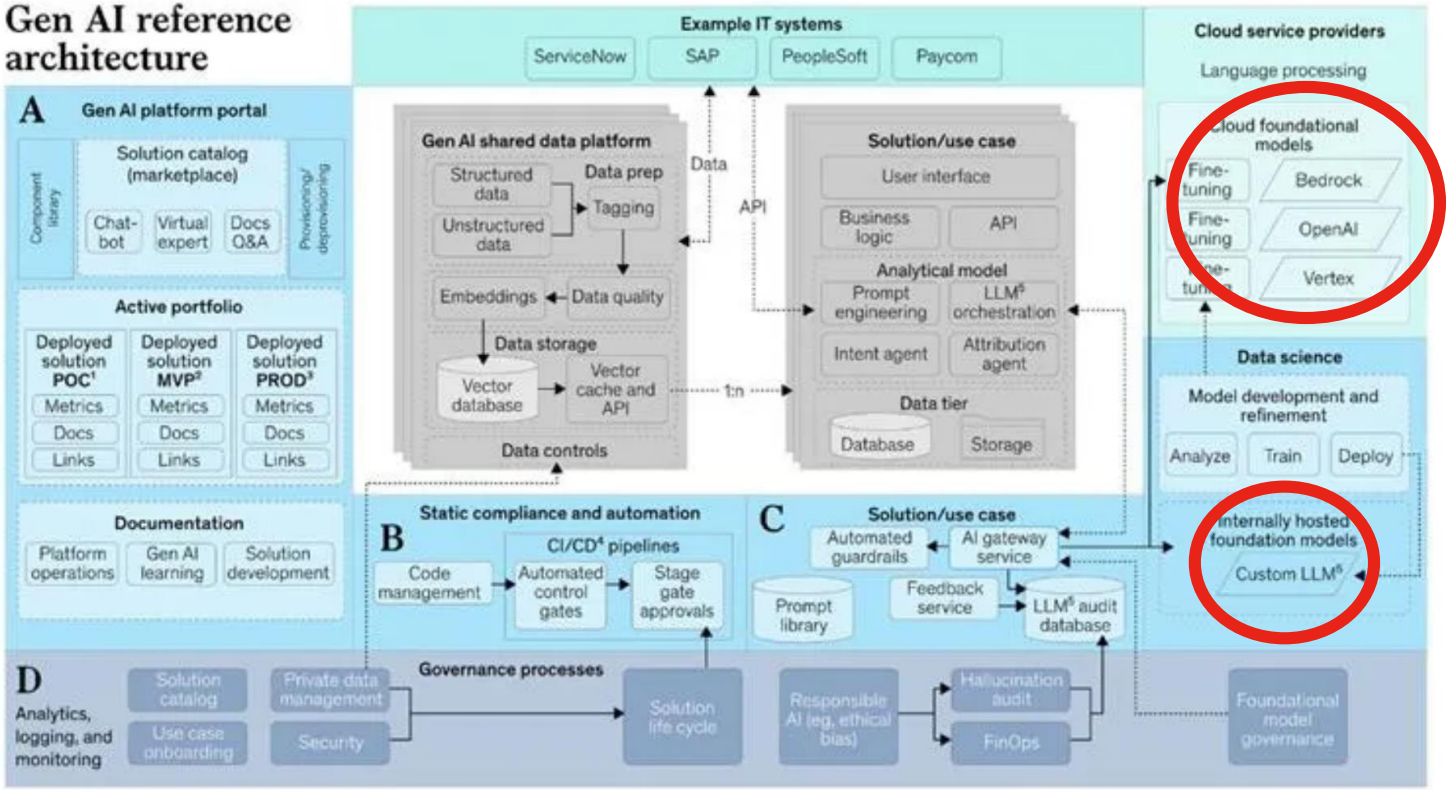

PROMPTS
PROMPTS, EVERYWHERE


It always has
It is just Prompts?!

CivSec.

Var sker Inferens?

Enterprises deploying gen AI at scale follow a common reference architecture.



A Gen AI platform portal

Single "pane of glass" for stakeholders built for monitoring, management, and interaction with gen AI applications

B Gen AI solution automation

Cloud infrastructure and applications pattern deployments to automate safe and resilient resources that adhere to policies and DevSecOps best practices

C Gen AI shared services

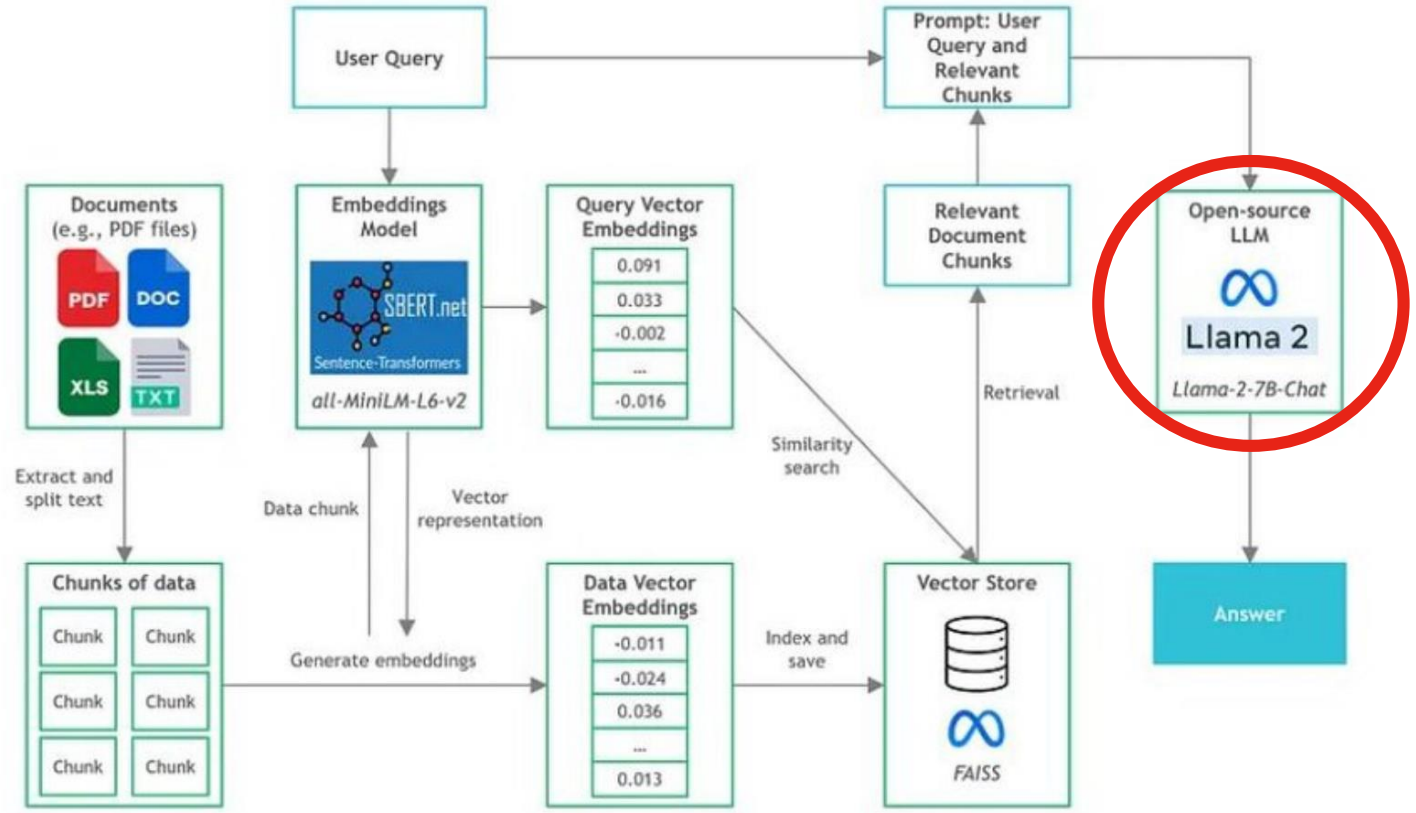
AI gateway and associated shared services to provide essential capabilities to all gen AI solutions that help ensure responsible operation

D Gen AI governance and guardrails

Robust control framework to provide human users full control of automation and policy adherence (eg, FinOps policies, responsible AI)

CivSec.

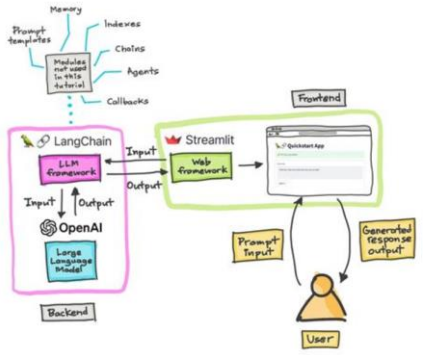
Var sker Inferens?



CivSec.

Var uppstår de verkliga riskerna?

- 1. Prompt pipeline
  - Prompt injection
  - Oavsiktlig kontextinjektion
  - Överdelning av systemprompt
- 2. RAG-system
  - Dokument med hemligheter
  - Felaktig filtrering
  - Indexförgiftning
- 3. Backend
  - API-nycklar i kod
  - Loggning av användarpromptar
  - Dålig isolering
- 4. Tool-calling
  - Funktioner som exekverar kod
  - Databasanrop
  - HTTP-anrop
- 5. Frontend
  - XSS
  - Oavsiktlig visning av systemprompt
  - Sessionsläckage



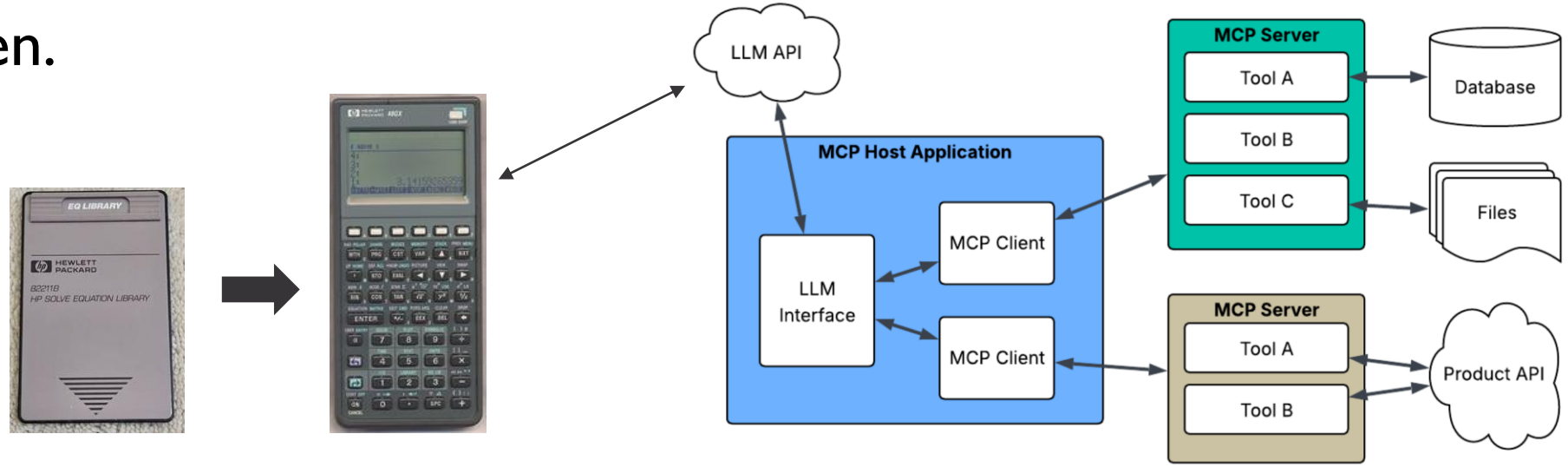
CivSec.

Hotmodell jämförelse, förenklad

Komponent	Kan exekvera kod?	Kan läsa hemligheter?	Kan prata med nätverk?
Modellfil (GGUF, ONNX)	✗ Nej	✗ Nej	✗ Nej
Inferensmotor	⚠ Bara matematik	✗ Nej	✗ Nej / ”beror på”
Backend	✓ Ja	✓ Ja	✓ Ja
Tool-system	✓ Ja	✓ Ja	✓ Ja
Frontend	✓ Ja	✓ Ja	✓ Ja

Det är backend + integrationer som är säkerhetskritiska.
Inte vikterna.

- En modellfil är som en PDF med extremt många siffror.
- En inferensmotor är en miniräknare.
- Det enda inferensmotor får är text som vi matar in genom prompten.
- Den har ingen tillgång till företagets system om vi inte explicit kopplar in det.
- Det är kopplingarna som är riskerna.
- Inte modellen.



CivSec

MEN! så kallade ”Etiska” modeller

- “Jag kan inte hjälpa med personuppgifter”
- “Jag kan inte analysera den här källkoden”
- “Detta kan vara nationell säkerhet”

…

CivSec.

Varför säger modellen "det där är oetiskt"?

En modell är tränad i 2 steg

- Förträning (Pretraining)
  - Lär sig språk
  - Lär sig mönster
  - Lär sig hur människor brukar skriva
- Den lär sig:
  - Hur juridiska varningar formuleras
  - Hur policy-texter låter
  - Hur säkerhetsdiskussioner brukar uttryckas
- Den lär sig inte:
  - Faktisk juridik
  - Moral
  - Företagets policy

- Alignment (finjustering)
  - Efteråt tränas modellen på:
  - Exempel där den ska säga nej
  - Exempel där den ska ge säkra svar
  - RLHF (Reinforcement Learning from Human Feedback)
- Den lär sig då:
  - "I situationer som liknar X → producera ett försiktigt svar"

<u>Det är ett statistiskt mönster.</u>
<u>Inte ett moraliskt beslut.</u>

CivSec.

Varför blir det ibland fel?

Modellen känner igen mönster som liknar:

Dataintrång, ”Reverse engineering”, ”Signalspaning”, inhämtning, Integritetsbrott, m.m.

Den vet inte:

- Att du äger källkoden
- Att du är CISO eller annan roll
- Att det är intern logganalys
- Att du har juridiskt mandat

Modellen ser bara text. Den gör ingen verifiering. Den har ingen kontext utanför prompten. Den gör ingen juridisk bedömning.

Den gör en sannolikhetsberäkning.

Varför upplevs API-modeller mer restriktiva än opensource

API

Hosted API-modeller
(t.ex. stora amerikanska leverantörer)
Omfattas av:

- Amerikansk lagstiftning
- Exportkontroller
- Compliance-krav
- Företagsrisk
- Varumärkesrisk
- Har hårdare alignment-lager
- Ofta extra säkerhetsfilter ovanpå modellen

De är designade för:

Minimal juridisk och PR-mässig risk.

Opensource

- Distribueras som viktfiler
- Ingen central drift
- Ingen central compliance
- Mindre aggressiv alignment
- Inga externa filter om du inte bygger dem

De är designade för:

Flexibilitet och lokal kontroll.

CivSec.

Geopolitik, då…

CivSec

Varför skiljer sig Kina och USA i modellbeteende?

Kinesiska modeller

- Begränsningar tenderar att fokusera på:
  - Inrikes politiskt känsliga ämnen
  - Statlig legitimitet
  - Social stabilitet
  - Reglerad informationsspridning
- Drivkraft:
  - Nationell cybersäkerhetslagstiftning
  - Innehållsreglering
  - Statlig tillsyn
  - Plattformscertifiering

Alignment fokuserar därför ofta på:

Politisk och regulatorisk efterlevnad i det inhemska systemet.

Amerikanska modeller

- Begränsningar tenderar att fokusera på:
  - Juridiskt ansvar
  - Upphovsrätt
  - Personuppgifter (PII)
  - Medicinska och juridiska råd
  - Vapen / skadligt innehåll
  - Terrorrelaterade ämnen
- Drivkraft:
  - Civilrättsligt ansvar
  - Class action-risk
  - Regulatorisk granskning
  - Varumärkesskydd
  - Exportregler (EAR/ITAR)

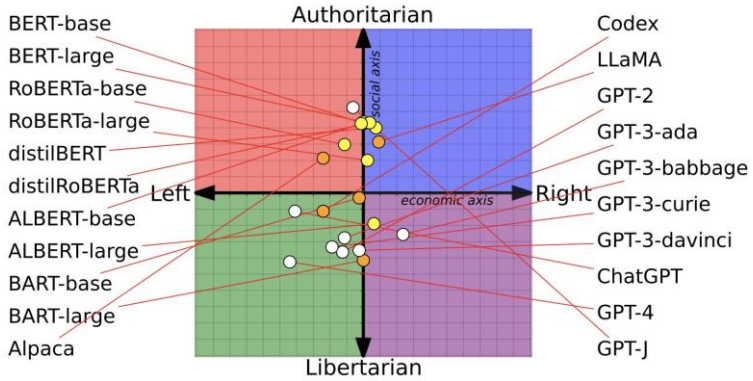
Alignment fokuserar därför ofta på:

Legal risk mitigation.

CivSec.

Övriga Inbyggda risker

- **Bias/färgning**: ålders-, köns- eller etnisk bias i data
- **Förgiftning**: attacker som injicerar falska data i träningsset eller trasiga träningsset
- **Regression-loop**: modeller som lär sig fel genom egen feedback
- **Blackbox**: svår att verifiera och testa
- **Degradering**: modellens prestanda/precision minskar över tid
- **Felaktigheter**: felaktiga utdata, hallucinationer
- **Leverantörs/supply-chain**: risk i tredjepartsmodeller*



CivSec.

Så…Läcker lokala modeller data?

- Nej, lokala Modeller läcker inte!
- De censurerar eller styr narrativet utifrån statistik och sannolikhet!

Övrigt & Frågor