Implementera AI,

låter enkelt, men mycket att tänka på

Sambruk

CivSec.

Agenda

Saker som vi kommer beröra idag:

- Vad är Digital suveränitet, Rådighet, Reseliens och Robusthet och koppling till AI
- Generella hot och exempel Hot mot AI, hot genom AI
- Digital Suveränitet och Rådighet, modell perspektiv (Öppna vs Stängda)
- Digital Suveränitet och Rådighet, teknikstack mjukvara att tänka på
- Digital Suveränitet och Rådighet, teknikstack hårdvara att tänka på
- Avslut summering, frågor

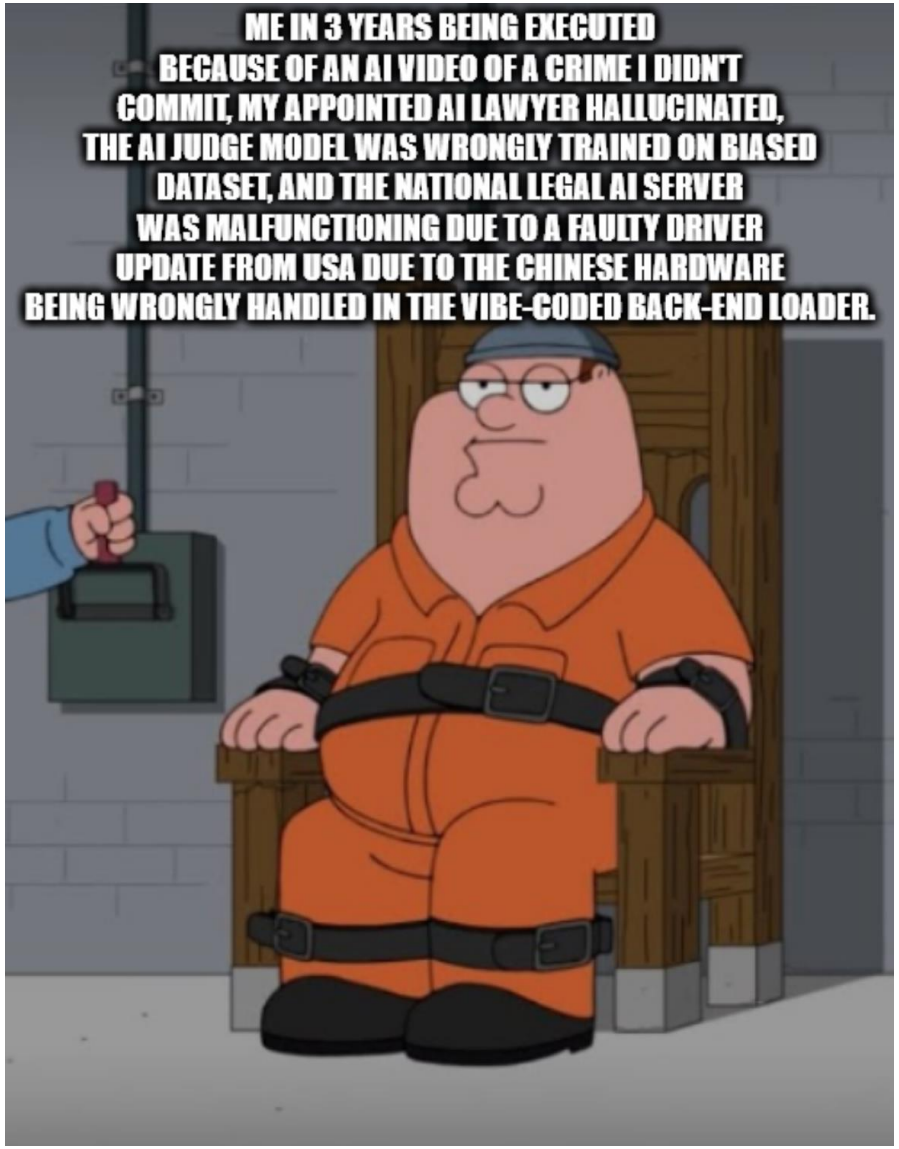
CivSec

- Fokus mot offentliga, samhällsviktig och samhällskritisk verksamhet och deras underleverantörer
- Cybersäkerhet, GRC, PSIRT, Säker AI, Övning, träning och utbildning
- Eget on-prem AI labb
  - Flera olika servrar (16 GPU, 192GB Vram, samt massa annat)
  - Olika hårdvara, NVIDIA, AMD, Huawei, Intel, m.m.
- Säker utveckling
- Fokus på egen rådighet och nationell suveränitet

”Artificiell Intelligens”

eller

”Artificiell Information”



CivSec.

Digital… vad pratar vi om ?

- Resiliens
- Robusthet
- Rådighet
- Suveränitet

CivSec.

Antagonistens nyttjande av LLM/AI

CivSec.

Geopolitik, då…

CivSec

Threat Taxonomy, Example

Threat Type	Description
Prompt Misuse / Injection	Malicious concatenation or structuring of input prompts to manipulate masked prediction or output.
Embedding Poisoning	Training-time attacks that subtly modify token embeddings or introduce malicious associations.
Training Data Leakage	Exposure of sensitive information via overfitting or embedding inversion.
Overconfidence in Nonsensical Outputs	Model produces confident predictions on meaningless or adversarial input.
Data Memorization	Memorized sensitive content retrieved via probing or adversarial queries.
Contextual Entanglement	Spurious associations across unrelated tokens or document spans.
Resource Exploitation	Excessive memory, CPU/GPU usage from long or malformed inputs.

GenAI/ML/LLM

- Alla modeller är inte lika!

Modell klass/typ	Modell exempel	Användningsfall
Decoder-only	GPT-2, GPT-3/4, PaLM, Claude, Mistral, LLaMA	Freeform generation, chatbots, coding, agents
RAG	Facebook RAG, LlamaIndex, Haystack, LangChain, GPT-RAG, Cohere RAG	Knowledge-grounded generation, QA, assistant bots
Encoder-Decoder	T5, BART, mT5, PEGASUS	Sequence-to-sequence generation (e.g., translation, summarization, QA), Text-to-text generation
Encoder-only	BERT, RoBERTa, DeBERTa	Embeddings, classification, QA

Exempel på Hot, OBS - ej uttömmande, samt beroende på infra!

Threat / Vulnerability	Encoder-only (BERT, RoBERTa)	Encoder-Decoder (T5, BART)	Decoder-only (GPT, PaLM)	RAG-based Models
Prompt Injection	⚠ Rare but possible in fine-tuned heads	⚠ Via prefix / prompt confusion	🔥 High risk without guardrails	🔥 Injected via retrieval
System Prompt Leakage	✕ Not Applicable	⚠ Only in sequence generation	🔥 Common (requires protection)	⚠ Via retrieved docs
Hallucination	⚠ Possible via ambiguous contexts	⚠ Common in long outputs	🔥 Common, especially in reasoning	🔥 When retrieval is weak
Embedding Poisoning	⚠ In fine-tuned tasks	⚠ In fine-tune or supervised tasks	⚠ Via user fine-tunes	🔥 Via knowledge base poisoning
Context Poisoning	✕ Not Applicable	⚠ During long prefix chains	🔥 Long-range prompt attacks	🔥 Through vector retrieval
Token Overflow / DoS	⚠ Capped input size	⚠ Token overflow during generation	🔥 Runaway completions, loops	⚠ High retrieval cost
Memorization / PII Leakage	⚠ In embedding heads	⚠ In pretrain / output	🔥 High in long contexts	🔥 If private docs embedded
Output Steering / Bias	⚠ In classification output	⚠ Decoder label bias	🔥 Prompt phrasing control	⚠ Retrieval ranking bias
Best-Suited Use Case	Classification, embeddings	Structured NLP pipelines	Interactive generation & agents	Knowledge-grounded QA, bots
Security Leverage Points	Input schema, DP finetune	Template enforcement	Role-token enforcement, red teaming	Retrieval guardrails, context filters

Implementationsutmaningar

- Hårdvara
  - Leverantörer
  - Legacy
  - Supply chain
- Mjukvara generellt
  - OS
  - Firmware
  - Drivrutiner
  - GPU/NPU ramverk
  - Beroenden
  - Modell motorer
  - Runtom system inkl. databaser
- Mjukvara specifikt
  - Funktionsstöd
  - Begränsningar
  - Support
- Modeller
  - Typ
  - Licenser
- Geopolitik
  - Sanktioner
  - Begränsningar
  - Hot mot suveränitet och egen rådighet

Precision då?

Precision	Size	Speed	Accuracy	When to Use
INT8	8-bit	Fastest	Lower	Inference on mobile, edge devices
FP16	16-bit	Fast	Good	Balanced inference/training, cloud/edge
FP32	32-bit	Slower	Highest	Full training, precision-sensitive tasks

Use Case	**Description**	**Recommended Precision**	**Why?**
Chatbot (Customer Support, Virtual Assistant)	Conversational AI with fast response times	INT8 or FP16	INT8 for fast, cheap inference; FP16 if moderate accuracy needed
RAG (Retrieval-Augmented Generation)	Combines search with generation for contextual responses	FP16	Needs balance between speed and understanding from external docs
Document Handling (Summarization, Classification)	Processing large volumes of text documents	FP16	FP16 supports fast yet sufficiently accurate batch processing
Mathematical Reasoning	Complex math problem solving, symbolic reasoning	FP32	Requires high precision to avoid numerical errors
Code Generation / Understanding	Generating or interpreting code (e.g., Python, JavaScript)	FP16 / FP32	FP16 for general tasks; FP32 for high-accuracy or edge-case handling

CivSec.

Power of prompting

- Prompt engineering
- Context engineering
- Chain of thought
- ”Few Shot” prompting

Exempel, tydlighet, struktur…

AI kan inte läsa dina tankar!

```
{
    "task": "Research about Broadcoms ASIC compute-core architecture,
limitations and performancs",
"context_summary": "In-depth professional technical research to
evaluate if the broadcom technology can compete in FP32+ precision
area or if it is only designed for INT8/FP16 ",
"next_steps":
    [
        "Perform a very detailed web search, no special time limit
        and with highest detail to find related data about
        Broadcoms AI architecture and their ASIC compute-core
        implementation, use research papers, test reports, press
        releases, investment data any related information",
        "Deep analyze the identified research and papers that is
        connected, focus on architecture that can reveal what kind
        of precision that the compute cores are optimized for and
        limitations ",
        "By references, compile a text that summarizes the findings
        and then compare the result to equivalent hardware are from
        Nvidia, AMD, Huawei, Moore threads. "
    ],
"preferred_output_format": "Highly technical text, scientific and
formal and professional, evidence based "
}
```


CivSec.

Tänk efter före, våga ställa frågorna:

- Varför
- till vilket syfte
- vad ska vi få ut av detta

Övrigt och frågor