Shahjahan "Shadde" Khan

AI-STRATEG · GRUNDARE · KEYMAN

sk@mybaseai.com
shadde@ai-agentsx.com
+46 709 900 597

SKANNA FÖR LINKEDIN

⏱️ 30 minuter
🇸🇪 På svenska
🔒 100% lokal AI

— Gift, pappa till 2 barn

— Dedikerat mitt yrkesliv till AI sedan GPT-2 och BERT lanserades

— Grundare av **My Base AI** – lokal Datasäker RAG-plattform

— Precis avslutat **eSam** projekt för test av agentisk utveckling på myndighet

— *Boyband-sångare i **40 Plus*** 🎵

↓ Ladda ner PDF **↓ Ladda ner PPTX**

presentation/presentation.pdf | presentation/presentation.pptx

2019

GPT-2 & BERT
AI-resan börjar

100%

Lokal inferencing
utan molnet

My Base AI

RAG-system för
myndigheter

Keyman

Workshop &
Implementation

AI-terminologi – enkelt förklarat

TOKEN

Vad är en token?

En token är en liten del av text – ungefär ett ord eller en stavelse. Modellen läser och skapar text token för token. "Hej världen" = 2 tokens. 100 tokens är ungefär 75 ord, vilket innebär att en token är ungefär 0.7 till 0.8 ord.

INFERENCING

Vad är inferencing?

Inferencing är när en tränad AI-modell används i praktiken. Du ger modellen en fråga och den genererar ett svar baserat på vad den lärt sig. Det är det vi kör lokalt.

GPU VS CPU

GPU kontra CPU

GPU (grafikkort) kör tusentals parallella beräkningar – idealisk för AI. CPU (processor) är tillräcklig för kvantiserade modeller på modern hårdvara – utan dyrt grafikkort.

GENERATIV AI

AI som skapar nytt innehåll – text, bild, kod, ljud – baserat på mönster från enorm mängd data. Inte en sökmotor, utan en kreativ modell som resonerar.

LLM – LARGE LANGUAGE MODEL

En stor språkmodell tränad på miljarder textdokument. Den förstår kontext, för resonemang och genererar sammanhängande svar. Exempel: Llama, Mistral, Qwen, DeepSeek.

Generativ AI – en rad som sammanfattar allt

"AI tar in din fråga och genererar nytt, relevant innehåll – text, kod, bild, video eller musik – utan att du behöver vara expert."

TEXT & CHAT

Lokal ChatGPT

Kör din egen säkra chattbott. OpenWebUI ger ChatGPT-liknande gränssnitt, My Base AI är ett avancerat RAG-system.

Ollama · LM Studio · OpenWebUI

BILDGENERERING

Lokal AI-bild

Generera professionella bilder lokalt utan molnet.

Flux 2 · Flux 2 Klein · Z Image
Qwen Image Edit · ComfyUI

VIDEOGENERERING

AI-video

Generera video från text eller bild, lokalt.

LTX 2.3 · Wan 2.2

Se exempel:
instagram.com/reel/DT-GEcqjCGc

MUSIKGENERERING

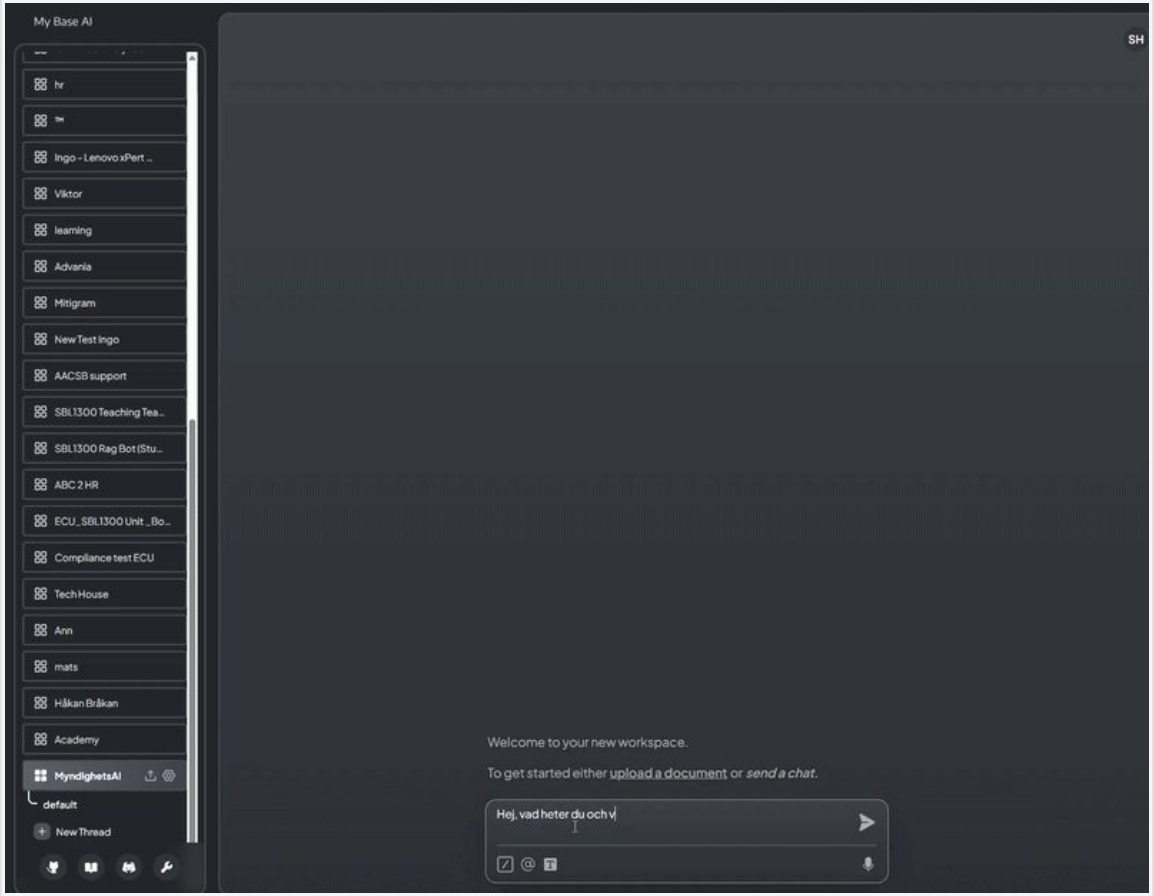
AI-musik

Skapa musik med text-prompts. Ace Step 1.5 genererar komplett musik lokalt.

Spotify-spellista:
https://open.spotify.com/playlist/38kWz0zwjrWSLe3uVwg2IH

Se lokal AI i aktion – MyndighetsAI

▶ DEMO VIDEO



VERKTYGSÖVERSIKT – TEXT & CHAT

→ **OpenWebUI – Open source, ChatGPT-liknande UI**

→ LM Studio – Enkel lokal modellhantering med GUI

→ Ollama – Snabb CLI-inferencing, API-kompatibel

→ My Base AI – RAG-system för myndighetsdokument

BILDGENERERING

→ Flux 2 & Flux 2 Klein

→ Z Image · Qwen Image Edit

→ ComfyUI – avancerat workflow-verktyg

VIDEO & MUSIK

→ LTX 2.3 · Wan 2.2 – videogenerering

→ Ace Step 1.5 – musikgenerering

AI behöver inte vara dyrt

KOM IGÅNG FÖR

10 000 kr

EN MODERN LAPTOP RÄCKER

En laptop upp till 5 år gammal kan köra moderna, kapabla modeller som Qwen 3.5 i 0.8B–4B storlek. Med CPU-inferencing via Ollama eller LM Studio behövs inget grafikkort.

Även äldre hårdvara duger – dagens optimeringar gör lokal AI tillgänglig för alla organisationer.

[4-bit Q4_K_M]

[8-bit Q8]

[CPU Inferencing]

[GGUF-format]

KVANTISERING – MODELLKOMPRIMERING FÖRKLARAT

Vad är miljarder parametrar (B)?

En modells "B" (miljarder) anger antalet interna vikter – ungefär som hjärnceller. En 7B-modell har 7 miljarder parametrar. Fler = smartare, men kräver mer minne och beräkningskraft.

16-bit är onödigt – kör 8-bit eller lägre

Att köra modeller i 8-bit eller 4-bit kvantisering ger bara ~1% kvalitetsförlust men halverar minneskravet. Tänk på det som en video: originalet är 70–100 GB, men en 300–700 MB-fil ser nästan identisk ut.

Kvantisering tar bort lager som inte bidrar märkbart till kvaliteten – smart komprimering utan att offra precision.

Kör lokal AI – plattformar & ekosystem

MY BASE AI – MIN LÖSNING

My Base AI

Ett komplett RAG-system för myndigheter och företag. Ladda upp era egna dokument och få precisa svar – 100% lokalt, inga data lämnar er server.

[RAG] [On-Prem] [GDPR]

OPEN SOURCE

OpenWebUI

En öppen, självhostad ChatGPT-klon. Kopplas mot Ollama. Stöd för filuppladdning, RAG, röst och plugins.

Ollama

Kör hundratals modeller från terminalen. Perfekt för servermiljöer och API-integration. Extremt enkel setup.

ANVÄNDARVÄNLIGT GUI

LM Studio

Ladda ner och kör modeller med ett klick. Inbyggt chat-UI, GPU-acceleration och lokal API-server. Idealisk för icke-tekniska användare.

HUGGINGFACE

Världens största öppna AI-bibliotek – hundratusentals modeller, nya releases varje dag. Härifrån hämtas alla lokala modeller.

vLLM & Unsloth Studio — För produktionsmiljöer erbjuder vLLM hög-throughput inferencing, medan Unsloth Studio möjliggör finjustering (fine-tuning) av modeller på er egen data – direkt på er hårdvara.

Agentisk kodning – AI skriver er kod

VAD ÄR AGENTISK AI-KODNING?

En AI-agent analyserar krav, skriver kod, testar den, hittar fel och itererar – utan att du behöver vara utvecklare. Agenten tar hela flödet från idé till fungerande kod.

KILO CODE – FULLT ÖPPEN KÄLLKOD

Vad är Kilo Code?

En VS Code-plugin för agentisk AI-kodning. Fullt open source – du kan forka och anpassa fritt. Kopplas mot lokala modeller via Ollama eller OpenAI-kompatibla APIer.

→ Agentisk filhantering & kodskrivning

→ Förståelse för hela kodbasen

→ Kör lokalt – ingen data till molnet

→ Fork:a och anpassa efter era behov

ESAM – STATLIG AI-TESTNING

Vad vi testade

Via eSam testade vi state-of-the-art öppen källkods-modeller för statlig kontext:

✓ Kodgenerering för myndighetssystem

✓ Dokumentanalys med RAG

✓ Säkerhet & dataintegritet

✓ GDPR-kompatibel on-prem deployment

REKOMMENDERADE MODELLER FÖR KODNING

[Qwen Coder] [DeepSeek Coder] [Llama 3 Code] [Mistral Code] [StarCoder 2]

KEYMAN & SHADDE KHAN

Halvdagsworkshop i Lokal AI

Vi kommer till er, kartlägger era behov och lär er organisation att komma igång med lokal AI – samma dag.

15 000 kr

4 TIMMAR · INKL. HANDS-ON DEMO

- ✓ Behovsanalys & AI-kartläggning
- ✓ Installation av lokal AI-miljö
- ✓ Hands-on demo med era dokument
- ✓ Handlingsplan för vidare implementation

BOKA & KONTAKTA

Redo att ta nästa steg? Skanna QR-koden eller kontakta mig direkt på LinkedIn. Kontakta Mats på Keyman mats.martensson@keyman.se

LinkedIn – Shadde Khan

+46 709 900 587
sk@mybaseai.com
shadde@ai-agentsx.com

VARFÖR LOKAL AI?

- → 100% datasäkerhet – inget lämnar er server
- → GDPR & sekretesslagstiftning uppfylls
- → Inga löpande API-kostnader
- → Full kontroll över modeller & data

TACK FÖR ER UPPMÄRKSAMHET

Framtiden är lokal & öppen.

AI behöver varken vara dyrt, molnbaserat eller svårt.
Låt oss bygga er AI-infrastruktur – säkert, lokalt och på era villkor.

LinkedIn

Shadde Khan
Keyman / My Base AI
sk@mybaseai.com
shadde@ai-agentsx.com
+46 709 900 587

↓ PDF: presentation/presentation.pdf
↓ PPTX: presentation/presentation.pptx