Lokaal LLM (AI) draaien

Inleiding

AI is momenteel booming. Overal op het internet verschijnen nieuwe AI-diensten en aanbieders. Voor mij biedt een Large Language Model (LLM) enorme kansen, omdat het mijn gedachten kan structureren en ondersteunen op plekken waar ik dat zelf lastig vind. Tegelijkertijd merk ik dat ik er steeds afhankelijker van word.

Het ontwikkelen van AI en het draaien daarvan op servers kost ontzettend veel geld. Gelukkig wordt die rekening nu voornamelijk gedragen door venture capitalists. Maar ik verwacht dat, net zoals bij de dotcom-bubble, er een moment zal komen dat deze bubble in elkaar zakt. Op dat moment zullen de kosten niet langer door investeerders gedragen worden, maar door de gebruikers. Omdat deze kostenverschillen nu zo extreem zijn, verwacht ik dat deze overgang genadeloos zal voelen.

Om voorbereid te zijn, wil ik alvast een open-source lokale LLM draaien. Zo kan ik blijven werken met deze technologie wanneer commerciële aanbieders de toegang beperken. Voor iemand zoals ik, die de ziel van een stuk tekst wil overbrengen, is een LLM een enorme uitkomst: ik lever de ziel, de LLM zorgt dat de ander het begrijpt.

Hoe pak ik dit aan?

Om te onderzoeken welke mogelijkheden er zijn, ben ik met ChatGPT in gesprek gegaan. Hieronder staan de vragen die ik heb gesteld, gevolgd door een samenvatting van dat gesprek. Beide onderdelen zijn opgesteld met behulp van ChatGPT.

Mijn vragen aan ChatGPT

- **Zorgen over de toekomst van ChatGPT**
  → Ik gaf aan bang te zijn dat ChatGPT in de toekomst te duur wordt, omdat het bedrijf nu verlies draait maar later winst moet maken.
  → Ik vroeg hoe ik *ChatGPT of iets vergelijkbaars lokaal kan draaien*.
- **Hardwarebeschrijving**
  → Ik beschreef mijn systeem: een **pc met een 4e generatie Intel CPU** en een **AMD RX580 GPU** met voldoende VRAM.
  → Ik vroeg of deze hardware geschikt is voor het lokaal draaien van AI-modellen.
- **Vraag over AI-type**
  → Ik vroeg wat voor soort AI ChatGPT eigenlijk is, dus welk type model of architectuur erachter zit.
- **Aanbevolen open-source model**
  → Ik vroeg welk **open-source model** het beste past bij mijn gebruik van ChatGPT.

- **Overzicht van open-source LLM’s**
  → Ik vroeg om een **overzicht van alle open-source Large Language Models (LLM’s)** die momenteel beschikbaar zijn.
- **Uitleg over “parameters”**
  → Ik vroeg wat **‘parameters’ (of ‘params’)** betekenen bij AI-modellen en hoe dat zich vertaalt naar prestaties en hardwarevereisten.
- **Specifieke modelaanbeveling**
  → Ik vroeg welk model ChatGPT mij persoonlijk zou aanraden en waarom.
  → Het antwoord was **Mistral 7B**, omdat het krachtig, efficiënt en volledig open-source is.
- **Meer informatie over Mistral 7B**
  → Ik vroeg wie het heeft ontwikkeld, onder welke licentie het valt, en wat de technische kenmerken zijn.
  → We bespraken dat het model is gemaakt door **Mistral AI**, valt onder de **Apache 2.0-licentie**, en bekendstaat als een van de beste open modellen in zijn klasse.

Samenvatting van het gesprek over lokale AI-modellen

In het gesprek werd besproken hoe het gebruik van ChatGPT zich in de toekomst kan ontwikkelen en hoe ik daarop kan inspelen door lokaal met open-source AI-modellen te werken.

Ik gaf aan me zorgen te maken dat ChatGPT op termijn te duur kan worden, omdat het bedrijf nu nog verlies draait maar uiteindelijk winst zal willen maken. Dit leidde tot de vraag hoe een vergelijkbaar model lokaal gebruikt kan worden, zonder afhankelijk te zijn van een abonnement of cloudtoegang.

Mijn systeem — een **Intel i5 (4e generatie)** met een **AMD RX580 GPU** — blijkt geschikt om kleinere, efficiënte modellen te draaien. Vervolgens kwamen de volgende onderwerpen aan bod:

- **Wat voor type model ChatGPT is:** een transformer-gebaseerd Large Language Model (LLM).
- **Beschikbare open-source alternatieven:** onder meer **Mistral**, **Llama**, **Phi**, **Gemma** en andere vrij beschikbare modellen.
- **Betekenis van parameters:** de getrainde gewichten van een model die diens kennis en taalvaardigheid bepalen, en hoe het aantal parameters invloed heeft op prestaties en hardwarevereisten.

Op basis van mijn gebruiksprofiel en hardware werd **Mistral 7B** aanbevolen:

- Ontwikkeld door het Franse bedrijf **Mistral AI**
- Licentie: **Apache 2.0** (volledig open en commercieel bruikbaar)
- Bevat **7 miljard parameters**, en draait goed op een GPU met 8 GB VRAM
- Biedt een uitstekende balans tussen snelheid, kwaliteit en efficiëntie

Mijn gebruikersprofiel

Volgens ChatGPT gebruik ik het model op de volgende manier:

- Ik werk voornamelijk met **technische en inhoudelijke teksten** (adviezen, rapporten, PowerPoints, systeemvisies).
- Ik wil **taal verfijnen, structureren en verduidelijken** zonder “wij/onze”-toon.
- Soms werk ik aan **ICT-architectuur, datamanagement of geo-ICT-vraagstukken**.
- Af en toe gebruik ik het voor **persoonlijke zaken** (gezondheid, training, cadeau-ideeën), maar mijn kern ligt bij **inhoudelijke, nuchtere ondersteuning**.

Daarom heb ik vooral een model nodig dat:

- **goed Nederlands begrijpt en schrijft**
- **instructies nauwkeurig opvolgt**
- **lange redeneringen of stukken tekst kan verwerken**
- **niet te groot is**, zodat het op mijn RX580-pc draait

Afsluiting

Deze verkenning is een eerste stap richting meer technologische onafhankelijkheid. Door te leren hoe LLM’s lokaal kunnen draaien, vergroot ik mijn begrip van AI-technologie en mijn controle over het eigen gebruik ervan. Het doel is niet om commerciële diensten volledig te vervangen, maar om te begrijpen hoe ze werken en hoe ik ze op mijn eigen voorwaarden kan inzetten.

Toekomstige stappen zijn het installeren en testen van Mistral 7B, en het vergelijken met zowel kleinere als iets grotere modellen. Mijn grootste beperking is mijn “oude” hardware, maar ik geloof dat beperkingen creatief maken.

Disclaimer

Dit document is opgesteld met behulp van ChatGPT (GPT-5-mini) als hulpmiddel bij het samenvatten en structureren van informatie. Hoewel het AI-model accurate en actuele informatie biedt, kan het fouten bevatten en vervangt het geen wetenschappelijk onderzoek, officiële documentatie of professionele advisering.

Bronvermelding:

- ChatGPT (GPT-5-mini), OpenAI, 2025
- Prompts gebruikt in dit document (samengevoegd/samengevat):
    1. “Samenvatting van gesprek over lokale AI-modellen, inclusief vragen die ik stelde en antwoorden.”

2. “Maak het vloeiend, professioneel en geschikt voor een HBO-niveau studiedocument.”
3. “Voeg een inleiding, afsluiting en disclaimer toe met bronvermelding.”
4. "Herschrijf de tekst zodat het vloeiend wordt, zonder inhoud te verliezen"

/* Notities uitvoering
RX580 wordt niet meer ondersteund door ollama
Mistral 7B quantized, wat is precies quantized?

- Quantized staat voor kleine en lichter AI model. Meestal houdt dit in hoeveel bits er worden gebruikt. FP32, FP16: FLoating point xx-bit, Q8/INT8, Q4 / INT 4, x bit integer.

Vanwege niet ondersteunde GPU werk ik met alleen een CPU, dit werkt niet zoals ik wil.

Mijn conclusies is dat mijn hardware te oud is

Ik zou kunnen onderzoeken of ik een lichter model kan gebruiken, maar dan ga ik te ver weg van hou ik op dit moment chat gpt gebruik.
*/