Masterclass: AI zonder dat je data je computer verlaat

Bron: AI Report (aireport.nl) - automatische transcriptie via Paco.

Welkom. Dit is misschien wel een van de meest gevraagde onderwerpen die we kunnen bespreken hier. Want waar wij ook komen, altijd is de vraag ongeveer hetzelfde. Namelijk, hoe zit het met mijn privacy en hoe kan ik AI gebruiken als ik allerlei gevoelige gegevens heb? En iedereen heeft gevoelige gegevens, van je financiële gegevens tot je e-mail tot... zielen roersteen die mensen tegenwoordig delen met AI. En als het daar al niet over gaat, gaat het wel over de gegevens van je bedrijf. Hoe ga je daarmee om? En hoe breed is het probleem eigenlijk? En wat kun je zelf doen? Dat zijn vragen die denker des vaderlands Wietse Haag vandaag voor jullie gaat beantwoorden. In de eerste 50 minuten van deze webinar, waarbij er in de laatste 10 minuten voor jullie ruimte is om vragen te stellen, gebruik daarvoor de Q&A tab of de V&A tab van Zoom. Dus gebruik niet de comment functie daarvoor. Daar kunnen jullie met elkaar discussiëren. Doe dat ook vooral. Jullie kunnen elkaar ook helpen bij heel veel dingen. Maar als je vragen hebt aan Wietse, stel die dan via die Q&A functionaliteit. En ik zal die uitkiezen op basis van welke vragen de meeste duimpjes omhoog hebben. Dus als je een vraag goed vindt, geef dan een duimpje omhoog. En op die volgorde behandelen we de vragen. Without further ado, Wietse. Go your gang. Dankjewel, ik heb er zin in. We gaan praten over lokale AI. Het voelt een beetje alsof ik jullie naar mijn treinbaancollectie meeneem. Want die doe je heel veel mee. Het is best een technisch puzzeltje wat we gaan leggen met elkaar vandaag. Maar het wordt ook heel gaaf. Ik heb er echt heel veel zin in. Om te beginnen hebben wij eigenlijk even een voorproefje van wat je dan straks kan. Ik ga er al wel bij zeggen. In eerdere webinars doen wij heel vaak een soort... Ja, je toch wel onder de indruk laten zijn van de kracht van AI. Namelijk moet je kijken wat het nu allemaal ineens kan. Lokaal kunnen we niet zoveel als dat we dat kunnen met cloud modellen. Maar we kunnen wel interessante dingen doen. Om te beginnen iets wat saaier is praten met een AI op jouw telefoon. Android of iPhone. Dit is een iPhone voorbeeld waarbij je de vraag stelt hoe maak ik mijn koffie extra lekker. Hij staat in airplane mode. Ik ga ook tijdens de hele presentatie een icoon in de hoek zetten wanneer we... offline zijn en wanneer we online zijn dat je een beetje een idee krijgt wat kan offline wat kan online dus let een beetje op de hoek of wij op daadwerkelijk op internet zitten terwijl ik dit terwijl ik met AI interacteer of niet dan is er de mogelijkheid om met omnie modellen te praten google heeft een model gemma 4 en in dit gemma 4 model kun je bijvoorbeeld een foto tonen dan zeggen wat zie je op de foto en dan krijg je wat erop staat daarmee probeert te zeggen het kan steeds meer wat er lukt Ja,

precies. Dus zet heel even het dingetje uit in je hoofd dat je denkt, maar dit kon AI toch al? Nee, hij staat op airplane mode. Dit gebeurt op jouw toestel. Dat gaat er niet vanaf, het gebruikt jouw processor. Eigenlijk is het technisch best wel een effe een stap. Ook al is hij qua eindgebruikersgemak misschien niet zo indrukwekkend. Dus dat wordt een beetje het thema vandaag. Je moet vooral een beetje onder de indruk zijn van de motor en wat minder van hoe hard die kan. Maar hij kan wel wat harder. Dit is Gemma 4, 31 biljoen. Ik ga straks over parameters praten en wat die dan betekenen. Hier vraag ik om een leuk verhaaltje voor het slapengaan voor mijn neefje. En daar maakt hij best een goed verhaal van. De lokale modellen kunnen inmiddels ook redelijk Nederlands. Dat was ook heel lang een probleem. Onze taal staat gewoon niet zo hoog. Dus de modellen spraken vaak best wel goed Spaans, Engels, Chinees. En Frans, maar gewoon niet zo heel erg goed Nederlands. En de Gemma-modellen doen dat best oké. Daardoor kan je dus ook samenvattingen laten maken. Mailtjes laten schrijven. Met je mail die alleen op jouw computer blijft. Maar daar kom ik straks op. Best wel gaaf. Hier doe ik, misschien verwarrend, een hybride. Wat doe ik hier? Hier mag het model wel het internet op. Maar draait het model toch nog op jouw laptop? Dus alleen als die informatie nodig heeft die die niet heeft... gaat hij het even vragen, in dit geval via DuckDuckGo... en voegt hij de informatie toe. Ik ga straks laten zien wat er gebeurt. Als je dat niet doet bijvoorbeeld... dan krijg je hele andere jaartallen voor de kinder. Maar in dit geval gaat het dus niet langs de service van OpenAI. Exact. En alleen langs de service van DuckDuckGo. Welke searchdienst dan ook. Of jouw persoonlijke database. Noem maar op. Best wel gaaf. Dan gaan we aan het einde... en maak je geen zorgen, ik ga uitleggen wat hier gebeurt. Ik liet dit thuis zien en daar werd gezegd, je moet echt heel veel plaatjes gaan gebruiken om dit uit te leggen. Hier heb ik een applicatie laten maken door Cloud AI, die vervolgens lokale AI gebruikt om zijn werk te doen. Ja, dus het houdt eigenlijk in dat je de kracht van de cloud modellen gebruikt om een stukje software te maken, die je vervolgens offline kunt blijven gebruiken. Ja, je zegt, een groot deel van het enthousiasme over AI is tegenwoordig dat je zelf software kan maken. Daarvoor zijn die lokale modellen nog niet goed genoeg, als ik je goed begrijp. Ja, klopt. En een tussenoplossing is dat je dan wel die software maakt met die cloud AI, die nieuwste modellen. Maar dat je hem zo maakt dat hij daarna geen verbinding meer hoeft te hebben met het internet. Dus alle gevoelige verwerkingen doe je dan met offline. Ja, dus je gebruikt een grote intelligentie om eigenlijk de tool te bouwen voor je, jouw gereedschap. En vanaf de rest van de week ga je dat nieuwe gereedschap wat je uit de cloud hebt gemaakt... lokaal gebruiken. En dat is wat mij betreft, en ik lees daar weinig over... een hele interessante stap om te doen. Dus daar gaan we ook wat meer over praten. Hier zie je bijvoorbeeld hoe ik in Word zit. Je ziet het co-pilot-icoontje daar in de hoek. We hebben een eerdere co-pilot-webinar gedaan. Ik gebruik het co-pilot-

knopje hier niet. Ik gebruik mijn keyboard-shortcut uit de app... die ik heb laten maken door AI... die de lokale AI aanroept om dit stukje tekst... eventjes te redigeren voor mij. En dit had ik dus ook offline... in mijn tent in het bos kunnen doen. En dat is best wel gaaf. Ik neem als het ware een stukje van de cloud mee. Dan de laatste die ik heb gebouwd. Die had ik ook even zelf nodig. Ik had een hele map vol met gedichten. Gescande gedichten. En ik wilde die OCR'en. Dus van een rauwe scan met tekst door verschillende fonds. Gedraaid. Soms met potlood. Omzetten naar tekst. En dan volledig lokaal. Dus deze DocuScan OCR app heb ik gemaakt. Met codecs in de cloud. En die kan ik vervolgens jarenlang lokaal gebruiken. Want ik heb hem heel even in de cloud nodig gehad. En vanaf dat moment kan ik door. En dat is toch wel een heel krachtig iets van deze tijd, vind ik. Goed, dan even een paar stapjes terug. Lokale AI, want mensen hebben het over soevereiniteit, ze hebben het over privacy, dingen als Edge AI. Het gaat er uiteindelijk om dat je je laptop op vliegtuigmodus kunt zetten of je smartphone en nog steeds toegang hebt tot de verschillende modellen die op jouw eigen apparaat staan. En die modellen kunnen van alles doen, plaatjes genereren, plaatjes bekijken, tekst genereren, tekst bekijken, noem maar op wat je gewend bent van de cloud. Het gaat wat trager. Het is wat minder intelligent. Maar daarom heb ik gewacht om deze webinar te doen. Ik vind dat je inmiddels, en daar moet je wel je portemonnee voor trekken. Daar kom ik straks ook op. Thuis en op kantoor. Want dit gaat ook over bedrijven die een stukje soevereiniteit willen. En een stukje van hun data binnen hun bedrijfsmuren willen houden. denk ik dat we economisch gezien op een punt aan het komen zijn... dat edge AI, dus AI op de rand, dus niet in het datacenter... maar bij jou aan de rand, serieus interessant aan het worden is. Waarom heet het edge AI? Nou, eigenlijk van het woord edge compute. En het is een soort net als met de last mile. Als je een pakketje levert ergens dat je zegt... Je kunt helemaal in de cloud gaan zitten met alle spullen. Dan moet het door allemaal kabels heen. En dan uiteindelijk helemaal naar jouw apparaat toe, zeg maar. Maar hier net naar jouw apparaat, voordat hij naar de cloud gaat. De rand, zeg maar. Dan zou je zeggen, we doen het... En voor bedrijven is het vooral belangrijk dat je... Stel dat jij een AI bedrijf bent, gewoon laten we voor het gemak zeggen Google, die vindt het op zich ook best prettig dat er wat meer naar de edge geduwd wordt, naar de rand van het netwerk, zeg maar, waar de gebruikers zitten, zodat je ook al die rekenkracht niet in je datacenter hoeft te hebben. Ja, maar het idee is dus dat niet iedereen bij zijn computer een eigen AI server heeft staan, maar dat dat dan een gedeeld ding is voor het hele bedrijfsnetwerk. Je zou kunnen zeggen, dan moet een soort semantiek over het woord edge gaan doen. Van waar zit dat randje dan en wat zit er nog binnen die rand? Binnen bedrijven zou het on-premise heten namelijk. Dat je zegt, we zetten een stukje AI in ons eigen bedrijf. Maar alles wat ik vandaag vertel gaat net zo goed op voor het bouwen van een kleine... Ja, het kan ook thuis. Ja, het gebruiken van

een smartphone, het gebruiken van een laptop, gebruiken van laptops. Of een stapeltje service op kantoor. In essentie zou dat ook een lokaal model zijn. Wat kan ik aan lokale AI niet? Je gaat er eigenlijk vandaag ook achterkomen... waarom die datacenters zo groot zijn. En waarom er zoveel datacenters bijgebouwd worden. Want de rekenkracht en ook vooral het geheugen... ga ik zo meer over praten... wat je nodig hebt om de vraagjes die wij dagelijks... gewoon maar in klat en chat GPT en Copilot gooien... Om die te kunnen beantwoorden gaat er ergens iets anders in de buurt van Zeewolde ineens heel hard blazen om voor jou die rekenkracht te doen. Maar die horen wij niet, want ja, we zitten met onze telefoon in onze hand. Nu ga je dus ook merken dat je iPhone heel warm gaat worden. Letterlijk. Letterlijk. Dus je voelt als het ware voor het eerst wat meer ook het energieverbruik in je hand. Ik sta hier ook, ze hallucineren vaker. Dat is waarom lokale AI in ieder geval de laatste twee jaar niet zo serieus genomen is. Omdat er zo weinig kennis in de modellen zit. Ik ga straks uitleggen wat ik daarmee bedoel. Maar het komt erop neer. Als jij feitelijke vragen gaat stellen. Dingen die je normaal op Wikipedia zou lezen. Dan kan je lachen vooral. Want het is hilarisch. De koning heet ineens, hoe heette die nou gisteren? Klaas in plaats van Willem-Alexander. Ik ga het straks allemaal laten zien. Ehm... Waarom eigenlijk lokale AI? Alexander zat al een beetje foreshadowing te doen. Want het heeft natuurlijk met privacy te maken. En ik zit al foreshadowing te doen met energieverbruik. Want ik voelde de warmte in mijn hand. Veel redenen voor lokale AI. Meer dan ik eigenlijk dacht toen ik erover na ging denken. Werd er steeds meer. Eén, privacy. Dat is ook de insteek van vandaag. Ik denk voor de meeste mensen. Als je iets geeft om dit onderwerp. Dan geef je eigenlijk om data. Dan wil je als bedrijf of als particulier. Vind je het best wel spannend om een gesprek te hebben over je gezin en hoe je de opvoeding beter kan doen? Vind je het best wel spannend om te praten met een soort virtuele huisarts? Vind je het best wel juridisch ingewikkeld en misschien zelfs onmogelijk om met bepaalde data te werken? Dan ben je eigenlijk iemand die op zoek is naar lokale AI. Als je toch graag een beetje met die modellen zou willen interacteren. Het is ook zo dat data... Ik ga hier later nog wat meer op in hoor. Kijk, we hebben al veel data in de cloud. Dus dat is even een stapje terug. Veel mensen hebben hun foto's bij Google staan of bij Apple. Je hebt je documenten bij Dropbox Drive of OneDrive staan. Dus het is ook een beetje gek. Ik hoor dat ook van veel mensen. Dat ze zeggen, is dat nu ineens een probleem dan met AI... Want al die gevoelige data staat al in de cloud. Dus wat is nou precies erger doordat we AI gaan gebruiken daarmee? Ja, en ik denk dat een van de redenen is dat we door hoe deze technologie zich naar ons uit. Als een life coach, als een gezinscoach, als een dokterachtig systeem. Dat we meer zijn gaan vertellen. dat we meer een dagboek zijn gaan bijhouden online... dan dat we hiervoor ooit hebben gedaan. Ik denk als ik aan mijn vrienden had gevraagd tien jaar geleden... hoeveel van jullie

heeft een dagboek in Dropbox staan? Ik denk niet zo heel veel. Terwijl je eigenlijk wel dagboek aan het schrijven bent... samen met ChatGPT. Dus er is een soort verandering gekomen... van de data die we zijn gaan delen... En je mag best weten, dat ga je ook zien vandaag. Je kunt natuurlijk ook heel veel analyse nu doen met AI op die data. Ook vanuit de bedrijven om voor jou andere reclames te tonen enzovoorts. Dus ik denk dat we ook meer een bewustzijn hebben gekregen van oké. Ik voelde eigenlijk niet zo last van mijn Dropbox of mijn andere diensten. Maar nu we de bedrijven zelf met hun eigen AI natuurlijk ook jouw data kunnen bekijken. Voor zover in de voorwaarden opgenomen. Want ik vind het heel belangrijk. Dus jij zegt eigenlijk met al die data die al in de cloud staat, of het nou je e-mail is of allerlei andere gegevens, eigenlijk omdat de analyseerbaarheid door de partijen die die data bewaren zoveel makkelijker wordt, komt je privacy eigenlijk automatisch meer onder druk te staan. Of jij nou meer AI gebruikt, ja of nee. Sowieso, dus op twee kanten. En daarbovenop komt nog eens dat de AI ons eigenlijk stimuleert om steeds meer te delen, ook al heb je het misschien niet door. We zouden nog een vector kunnen toevoegen aan dit verhaal... als het dan om data gaat. Veel van jullie hebben misschien gehoord over Mythos... een groot model van Entropic of GPT-55 Cyber van OpenAI. Ook Google heeft een model gelanceerd... nu specifiek bedoeld eigenlijk om de boel veiliger te maken op het internet. Maar vooral omdat er steeds meer sterke I komt om in te breken. Dus het aantal datalekken zal zeer waarschijnlijk ook toe gaan nemen. Dat zijn datalekken van cloud data. Dus ik... We ain't seen nothing yet met je Odido gegevens. Ja, dus het is ook nog... Mensen aan mij vragen... waarom is privacy zo'n ding ineens geworden met AI? Want ik had het toch al Dropbox. Ik hoop dat het nu iets duidelijk begint te worden... dat het speelveld aan het veranderen is. Ja, jij maakt je hier zorgen over. Ja, toch wel een beetje. De kosten. Ik bedoel, sommigen van jullie zijn misschien heel erg enthousiast... met AI bezig. Programmeren, grote datasets analyseren, noem maar op. Het kan best in de kosten lopen... Voor de duidelijkheid, lokale AI is niet gratis. Want je hebt een computer nodig. Die computer hangt aan zo'n stroomnet. Maar dat stroomnet heb jij misschien wat meer onder controle. Die computer gebruik je misschien s'avonds om video-edits te doen. Of als gaming machine. Ja. Op zich zou je op een gegeven moment een beetje kunnen gaan rekenen. Die kosten komen nu nog niet uit hoor, wil ik er wel bij zeggen. Er worden veel berekeningen online gemaakt van mensen die zeggen... nou, als ik voor 5000 euro een laptop moet kopen, heb ik een hoop tokens voor. Of dat is een hoop maandabonnementjes. Maar tegelijkertijd, als het een multipurpose device is... en het begint nu wel langzaam de modellen worden... sterker, terwijl ze niet per se aan het groeien zijn... de lokale modellen. En de apparaten worden ook langzaam wat goedkoper. Dus op een gegeven moment gaan ze het elkaar natuurlijk vinden. Voor mij persoonlijk hebben ze elkaar... in heel veel use cases al wat meer gevonden. Ik had altijd al

best een krachtige laptop. Daar zal ik ook wat meer over vertellen. Want ik doe dit allemaal vandaag hier op best wel een... zware laptop, moet wel gezegd worden. Ja. Energie, de geoptimaliseerde modellen, de kleine modelletjes gebruiken sowieso minder stroom. Dit is waarom als jij gratis ChatGPT gebruikt of gratis Gemini of Cloud, dat je Haiku krijgt in het geval van Entropic of Flash in het geval van Gemini of ChatGPT Mini. Dat zijn hun, die zouden bijna lokaal kunnen draaien zou je kunnen zeggen. Die gebruiken zij om de kosten te drukken. En jij gaat ook kleinere modellen draaien, want je moet wel. Dus je gebruikt sowieso minder energie. En je hebt keuze over waar jij je energie vandaan haalt. En ik denk zelf, kan ik ook voor vandaag zeggen, dit is niet een anti-cloud AI verhaal. Dat is een beetje te kortzichtig. Ik wil het hebben over een soort hybride. Waar wel edge AI. Bijvoorbeeld het samenvatten van een mail of een mail reply. Ja, ik ga er vanuit. En let maar op. Zowel Google als Apple met hun grote platforms gaan het allemaal naar de edge duwen. Dat zijn ze met Apple Intelligence wel aan het doen. Ook Google heeft zijn eigen edge model. Je zegt we gaan eigenlijk steeds meer, ook als gewone consumenten, steeds meer lokale AI gebruiken. Zonder dat je het weet. Die kleine modellen worden steeds beter en kunnen in principe van je zonnepanelen komen. En dan valt dat argument over energiegebruik in ieder geval voor je eigen gebruik weg. Ja, het wordt in ieder geval minder prangend. En ik hoop dus zelf eigenlijk, en ik zie op zich wel wat de eerste voortekeningen ervan, dat je de keuze gaat krijgen om te zeggen, ik wil geen cloud AI op mijn telefoon, dan maar matige replies in mijn mail. Maar dan heb ik tenminste controle over mijn eigen energiegebruik. En dan leggen we het wat meer bij de mensen. Dus weerbaarheid, ook weer door AI, krijgen we natuurlijk meer aanvallen op het netwerk. We zitten sowieso in een spannende geopolitieke situatie op dit moment. Dus ja, is het niet prettig dat als jouw bedrijf of jouw eigen particuliere leven inmiddels best wel afhankelijk geworden is van AI, dat het ook doet als de wifi het niet doet? Ik noem maar iets. Compliance natuurlijk. Ik zet hem wat later, want het voelt wat saaier voor veel mensen. Maar heel veel bedrijven en organisaties en ook individuen mogen dit gewoon helemaal niet gebruiken anders. En op het moment dat je het dieper naar de cloud stuurt, zijn je mogelijkheden in ieder geval breder. Ik zeg nu niet vandaag, oh, alles wat je lokaal doet is meteen goedgekeurd. Want je installeert het niet zomaar op een bedrijfslaptop, et cetera, et cetera. Maar neem voor mij aan, en er zijn ook veel organisaties die dit aan het doen zijn. Als het heel duidelijk is waar de computerkracht zit... waar de berekening gedaan wordt... waar de inference, zoals dat heet... dus eigenlijk wat AI-modellen doen... namelijk zware berekeningen om jou het juiste antwoord te geven... of een samenvatting te schrijven of een mail terug te typen. Als dat proces helemaal netjes lokaal plaatsvindt... dan heb je eerder een vinkje van compliance... dan als het naar een server moet. En dan autonomie. Kijk, uiteindelijk prijzen gaan omhoog van modellen. Providers

kunnen zeggen, je mag niet meer met ons werken. Of wat jij bespreekt vinden we niet prettig. Dit soort dingen zijn natuurlijk ook wat vrijer. Ik kan daar al meteen de kanttekening bij maken. Er is natuurlijk ook een reden dat deze AI bedrijven bepaalde dingen niet willen dat jij die bespreekt. Het kan ook over het gaan dat je misschien chemische wapens wil maken of iets. Dus ja, even iets in leuks. Ik wil maar zeggen dat... Ik wil even tegen de mensen zeggen die nu luisteren. Ik zit hier ook niet met de naïviteit van... Oh, leuk man dat iedereen intelligentie thuis heeft... waar ze van alles mee kunnen. Ik denk dat we de komende jaren heel erg gaan zien... dat de lokale modellen waarschijnlijk een beetje gecapt gaan worden. Los van hoe goed je laptop is. Op wat ze kunnen doen qua complexiteit. Omdat het ook best spannend is... om zoveel intelligentie... in de handen te geven van iedereen tegelijk. Hier kunnen we uren over discussiëren. Wolde ik toch even gezegd hebben. Waarom geen lokale AI eigenlijk? Hoe werkt Cloud AI? Oh wacht, ik moet het even uitleggen. Dit was mijn verhaal van net. Waarom geen lokale AI? Omdat... Je uiteindelijk je kunt afvragen of bijvoorbeeld een tiener die best wel gestrest is en spannend vindt om op school te zitten en een moeilijk verhaal heeft met een lokale AI. Ja, daar kan nu niks op ingrijpen. Daar kan nu niks mee lezen. Ja, je wilt zeggen, het moet niet een doel op zich zijn om alle AI lokaal te krijgen, want daarmee verliezen we ook als maatschappij grip. Hoe sluit je dan criminelen buiten als die op hun eigen computer opeens inderdaad allerlei dingen kunnen doen die cloud providers in ieder geval zouden voorkomen? Ja, ik denk dus dat mijn nuance vooral is dat het een mix wordt. Zo'n cliché ding net met de energiemix. Maar het moet een absolute mix worden van dingen waar lokaal op zijn plek is en dingen waar cloud prima kan. Want een hele wereld waarin iedereen superclusters thuis moet hebben en daar intelligentie op draait, dat klinkt misschien... Ook niet efficiënt. Nee, en op allerlei manieren misschien ook wel spannend. Dan, even een stapje terug. Want hoe werkt eigenlijk normale AI? Want ik dacht eigenlijk gewoon dat dat altijd al gewoon op mijn telefoon gebeurde. Ja, GPT staat toch op mijn telefoon? Klinkt misschien even suf wat ik nu zeg. Maar ik denk dat best wel veel mensen niet helemaal snappen wat er op jouw telefoon gebeurt. En wat er niet op jouw telefoon gebeurt. Kijk. eigenlijk is AI een samenkomst van op zijn minst deze twee componenten. Een thin client, dus een dunne computer, letterlijk een laptop, die niet zo heel veel kan, die een klein batterijtje heeft en een klein scherm, die zo snel mogelijk via een glasvezelkabel of via 5G verbinding maakt met de rest van de computer. van de computer. Het datacenter. En in dat datacenter zijn allemaal op elkaar... gestapelde computers die samenwerken... en bruitsware modellen kunnen draaien. En ze ook de hele dag staan te loeien. Want die rekenkracht stoot ook allemaal warmte uit. En kost ook heel veel elektriciteit. Dus als jij zegt, ik gebruik AI... besef je dan dat jouw computer... grotendeels niet bij jou thuis staat. En dit is ook waarom... Naarmate het aantal gebruikers van chat apps, om het maar even zo te

zeggen, toeneemt, ook automatisch het aantal vierkante meters van datacenters toeneemt. En die moet je dicht bij de gebruiker zetten. Waarom? Omdat je anders een hele lange tijd moet wachten op het antwoord. Dus je ziet dat eigenlijk in Nederland bijvoorbeeld een bepaalde plekken zijn, knooppunten, net als bij de snelweg, waar deze datacenters geplaatst worden. Dat is eigenlijk een kwestie van een open AI app op je smartphone of op je laptop. De ChatGPT app of Codex of Cloud Code. En aan de andere kant het datacenter en daar draait het model. Dus ik heb hier even de app een soort squirkel gemaakt. Een vierkantje en het model de cirkel. En die twee vormen een samenwerkende hybride. Dan, hoe zit het nou met die data? Want ik zei het net en ik krijg heel veel vragen over bij lezingen van... Maar hoezo dan... Zijn ze alles aan het opslaan? Is iedereen met alles aan het meelezen? Hoe paranoïde moet ik hier eigenlijk over zijn? Kijk, de bedrijven hebben hun voorwaarden. In de voorwaarden, dat is heel erg afhankelijk van of je betaalt of niet. Voor de niet betaalde versies van deze diensten betaal je op een andere manier met je data. Wat bedoel ik daarmee? Dan gaan ze trainen. Hier is het zo dat er een stukje opslag is... om heel veel gesprekken te blijven opslaan... om die later te kunnen gebruiken om te trainen. Die worden geanonimiseerd grotendeels. Maar weet wel, je gaat mee in de trainingsset. Je gaat niet altijd mee in de trainingsset. Bedrijven kiezen ervoor om duurdere contracten af te sluiten. Sommige partijen zeggen dit doen wij sowieso niet. Het is heel erg afhankelijk van met wie jij aan het praten bent. Maar weet, het bestaat absoluut en het gebeurt ook. Die modellen worden niet vanzelf beter... Is het zo dat alleen bij gratis gebruik van Chachapati wordt opgeslagen wat jij allemaal zegt tegen dat ding en dat dat gebruikt kan worden voor trainingsdaad? Of is dat ook bij betaalde? Wat ik heb kunnen vinden, veel langer bij gratis. En bij betaalde zeggen ze, er staat volgens mij als je erop klikt ergens een dingetje. Je moet hem sowieso uitzetten. Ik heb hem uitgezet. Het is een opt-out. oud uit mijn hoofd in chatgpt. Je mag hem optouten als je betaalt, zeg maar. Maar het is wel zo, want ik sta hier ook het woord logging. Het is niet een fully encrypted iets. Dus er kan ook voor gekozen worden van ja, dit is wel interessant. We gaan wel meelezen. En dat kan natuurlijk vanuit veiligheidsperspectief zijn. Maar ja, wie bepaalt wat veiligheidsperspectief is met een hoeveel gebruikers hebben ze inmiddels? Nou, een hoop. Dat kan natuurlijk ook geen mens zijn die dan meeleest. Dus Ga er in ieder geval van uit dat... In de gratis versies de kans groot is dat jouw prijs die je betaalt in plaats van euro's je data is. In de betaalde versies dat je ontzettend goed moet opletten waar je uiteindelijk echt voor tekenen. Hoe groot is de kans dat als jij, weet ik veel, iets deelt over een ruzie die je hebt met een vriendin. Dat een andere gebruiker van ChatGPT die data terug kan krijgen. En eigenlijk dus iets over jou weet wat jij niet wilt. Ik zeg het even voorzichtig, maar bijna nul. Dus dat is een theoretisch risico. Dat dingen die je ook voor je bedrijf, als jij Excel sheets upload met begrotingen die nog niet

publiek zijn of wat dan ook. Voor wie wil je het veilig houden is uiteindelijk de vraag. Ik denk dat veel mensen denken, als ik het invoer in chat, dan gaan andere gebruikers op een of andere manier daarachter komen. Omdat chat dat traint op wat ik invoer en dat die AI dat weer gaat zeggen tegen anderen. En jij zegt, dat is een theoretisch risico, als ik dat goed begrijp. Ja, omdat eigenlijk... Kijk, uiteindelijk vanuit al die trainingsdata... wordt een brood gebakken... wat een soort gemiddelde is van al die data. Kijk, als heel veel Nederlanders... het tegelijk een beetje over dezelfde relatieproblemen gaan hebben... dan zal er een soort relatieprobleempatroon in terechtkomen. Jouw punt is, is het dan terug te herleiden tot Alexander Klöpping? Niet voor andere gebruikers van die chatapp zomaar. Maar tegelijkertijd als jouw account lekt... Dan natuurlijk wel. En die bedrijven hebben ook nog eens andere belangen erbij om jouw gegevens te bekijken. En je moet dan als bedrijf maar hopen dat als je concurrentiegevoelige gegevens deelt weer bij de een of andere manier concurrent. Het zijn allemaal Amerikaanse bedrijven dat daar niet gebruik van gemaakt wordt. Maar... Ja, en ik denk dat, ik vind het wel goed dat jij net even nog aangeeft. Het gaat ook om data lekken. OpenAI heeft vorige week Advanced Security gelanceerd voor chat GPT accounts. Dus dan moet je best wel zware stappen door om je eigen account te beschermen. Dat doen ze dan toch ook niet voor niets. Dat zijn gewoon persoonlijke accounts van mensen. Omdat het ook zo is, ja, net als bij... je crypto wallet of je bank... ja, als mensen binnen zijn... kunnen ze dingen. En jouw twee jaar aan chat-GPT-logs... plus je memory die ze kunnen extracten... hebben we in de overstappers webinar over gehad. Goed, dat dan. Dus misschien is het... Ik zeg heel vaak als het om... angst rondom privacy gaat, dat heel veel mensen denk ik de juiste intuïtie hebben. Als in, ik vind dit spannend. En dan als ze vervolgens gaan uitleggen waarom, dan komt daar een soort buurman lees mee. Dat ik denk, nee, je moet wel voorzichtig zijn, maar dat is dan niet daarom. Toch? Dat jij voelt dat ook, denk ik, bij mensen. De intuïtie klopt vaak wel, van een soort spannend. Maar veel van ons kunnen dat dan, zelfs ik, kan nu niet helemaal pinpointen. Wat ik bedoel, het komt er wel op neer dat er gewoon heel veel data verzameld wordt over ons nu door onszelf. We geven het zelf terug in die gesprekken. Ik zei het al, chatGPT zijn standaard training, mogen gebruikt worden. Het is wel zo dat je dit extra niet uit kan schakelen. En het is ook zo dat veel partijen inmiddels er hun uniek selling point van maken om te zeggen, wij doen niets met jouw data. We willen het niet eens, betaal ons gewoon in euro's. Nou ja, ik zei het al, datalects hebben mooi vooruitgelopen op deze slides. Dit is eigenlijk wat we net zeiden. Dan... Is er een stukje rekenkracht? En dit is wel een heel belangrijk punt om even te vertellen. Jouw laptop, ook mijn hele dure laptop hier, is niet zo krachtig als het datacenter. En dat ga je merken als je hiermee gaat werken. Het datacenter is trouwens nog veel krachtiger dan op beeldpas. Dat ding gaat echt buiten hier naar een andere stad toe. We zitten in Amsterdam, eindigt in

Utrecht denk ik. Hoe werkt die lokale AI nou? Lokale AI is dit. Dit is de droom. Namelijk een zonnepaneel, een battery bank en iets duurdere computer met erop een model. Praten met je LGBT terwijl je geen bereik... Dit is wat jij voor je ziet. Dat je zo in de bergen zit en dan een beetje zit vibe coden. Dit is jouw droom hè? Je ziet aan de USB-port aan de voorkant dat het een AI-generated plaatje is. Maar dit is wel een beetje wat ik dacht, ik laat het even zien. Zie de zesde vinger en het display-stuk. Je moet het maar even op je vision board maar even naartoe werken. Ja, toch? Ja. Nou, wat gaan we doen? We gaan het model naar ons toe halen. We gaan hem ophalen. Kan je zomaar ChatGPT ophalen dan? Nee, zeker niet. Want dit zijn de diamanten van de AI-bedrijven. We hebben zogenoemde open-weight modellen en close-weight modellen. Ook OpenAI, daar gaan we vandaag ook mee werken, heeft een open-weight model. Een kleinere versie van hun ChatGPT. GPT-OSS. En die mogen we draaien. Google heeft Gemini. En Entropic heeft niets. Interessant detail. Wat gaan we doen? We gaan op die laptop LM Studio zetten. Ga ik zo laten zien. En dan praten we met het model. En dan is het datacenter weg. En dan zijn we nog alleen met onze laptop in het bos. Die rekenkracht is niet super die daarboven zit. Weet dat. Dus dat betekent dat we even moeten praten over... wat bedoelt Wietse dan met rekenkracht? Twee stukjes. Er is een GPU. Dat is de rekenmachine in jouw laptop. Vroeger noemen we het altijd de CPU. GPU is specifiek voor grafische opdrachten. Maar neem van mij aan, voor AI zijn GPU's belangrijk. Ik ga heel even overslaan waarom. Want het wordt anders wel een te lange sessie. Maar dit is een dure chip... Die moeten vaak in de laptop bij zitten. Dus je moet een duurdere laptop hebben. Of je moet het geluk hebben dat je een processor hebt waar die meegebakken is. Apple bijvoorbeeld bakte hem in één ding mee. Waardoor je niet een hele dure grafische kaart hoeft te kopen. iPhones, ook de Pixel lijn van Google, hebben MPU's erin zitten. Dat zijn speciale processoren voor AI. Maar dit is een beetje buiten de scope van de webinar. Een ander ding wat wel heel belangrijk is, is hoeveel geheugen heb je. En dan heb ik het niet over RAM, maar over VRAM. Om het lekker complex te maken. Als het een losse grafische kaart is, zit de RAM daarop. Maar als het een MacBook is of een recente ASUS Strix Halo, iPhone of Pixel, is het allemaal wel dezelfde RAM. Oké, ik ga naar de mediamarkt en ik moet een laptop hebben waar ik een lokaal AI-model wil draaien. Ik werk ernaartoe, ik werk ernaartoe. Beloof het, beloof het. Maar ik moest het even vertellen, omdat je anders dit tabelletje niet gaat begrijpen. Want ik wil even praten hierover en dan, oké, ik wil het draaien, wat moet ik hebben dan? We hebben vier opties waar we uit kunnen kiezen. We gaan vandaag werken met Gemma via een GPT-OSS. Er is ook QEN. Van Alibaba en Mistral uit Europa. Hier staan een aantal dingen bij. Parameters. De parameters zijn hoe slim is dit ding eigenlijk? Daar kan je wel van uitgaan. Hoe meer parameters, hoe groter, hoe duurder de laptop. En uiteindelijk moet je je voorstellen dat een mythos uit mijn hoofd van mij op een triljoen parameters zit

inmiddels. Om even aan te geven. Dit zijn baby garnalen. Absolute baby garnalen. Anders passen ze niet op onze laptop. Ja. Die babygenaal weten ze ook niet wie Willem-Alexander is. Want dat is er allemaal afgesnoeid, die informatie. Die kennis kan er niet meer in zitten. We moesten eigenlijk alles samenvatten naar patronen. Wat een koningshuis is, dat weet hij nog. Maar hij weet niet meer de geschiedenis van de Oranjes. Want dat is hoe je een model klein maakt. Kennis eruit, zoveel mogelijk generaliseren. Dan hebben we de kwant. Ik houd het hier even kort over, maar deze hebben we helaas echt nodig. Want mensen gaan dit thuis proberen en snappen er anders geen reet van. De kwant houdt simpelweg in een soort vorm van compressie. En hoe meer getallen je achter de comma mag hebben... ik sla hem even helemaal plat... hoe groter het model wordt, hoe slimmer het model wordt... hoe minder getallen achter de comma je als het ware geeft... hoe minder slim die wordt... Maar hoe kleiner die wordt en wat is gebleken. Dat is best wel een leuke doorbraak geweest. Je kunt hem best wel klein maken zonder dat die dom wordt. Dus daarom hebben we ook veel mooiere kleine modellen door de quants. Als jij in de local AI community zit, zeggen ze. Draai je een 5 of 6 bit quant? Dan moet je antwoord op hebben dan. Weet je wel? Goed. Lokale AI verwijst serieus hardware. We gaan naar de prijs van de laptop en waar je hem kan halen. De MacBook Pro is een zware machine. Hier staat 1929. Daar ga je het niet mee redden, kan ik je vertellen. Dan een Nvidia GTX Spark. Kan je een los apparaatje kopen bij Nvidia. Die alleen maar gebouwd is voor AI. Ik ga het zo op mij eens vertellen. Of je zegt, ik werk liever met Linux of Windows. Dan zou ik zeggen, ga schrijven naar een framework. Desktop of laptop. Ook die hebben veel VRAM. Want daar gaat het om. Anders heb je er niets aan. Wat kost dat dan? Echt serieuze knaken. Want je wilt 128 gigabyte VRAM hebben. Daar ben ik van mening. Anders kan je er gewoon niet zo heel erg veel mee. Hier zullen de meningen over verschillen. Maar ik kan daardoor veel met AI. En dat betekent dat je van 6.000 tot 3.500 euro in de BTW moet gaan uitgeven voor jouw thuis AI avontuur. Tegelijkertijd, mijn laptop, mag je best weten, heeft 24 gigabyte RAM. En ik heb dit vandaag allemaal hierop gedaan en gedemoot met een hoop blazen en schuiven. Oké. Niet goedkoop. Smartphone. Als je denkt joh. Ik moet toch binnenkort een nieuwe smartphone. Kan ik daar dan AI op? Dat kan. Je kunt gaan voor de iPhone 17 Pro. Of de S525 van Samsung. Of de Pixel 9. Prijzen. En VRAM. Hier kun je hele kleine modelletjes op draaien. Maar die kunnen best wat. Zie even een tiende van die laptop. Maar de prijs is ook wat lager. Voor dat geld krijg je wel een Pixel 10 inmiddels. Maar ja, inderdaad. Kijk, het gaat er vooral om. Check hoeveel RAM ze hebben. Dat is het allerbelangrijkste. Ja, eigenlijk toestellen die in de laatste twee jaar zijn uitgekomen. Daar kan je dit ongeveer mee doen. Ja, ik dacht ik doe nog één goedkopere optie erbij. Goedkoop knippoog, want het is een hoop geld. Dan, leuk. Alexander is naar de mediamarkt gegaan. Die heeft online een framework besteld. Dat doosje komt bij hem thuis.

Zegt hij, wiet, succes man. Wat gaan we doen? We gaan dingetjes installeren. Dit is echt leuk gegroeid. Dit was een jaar geleden heel matig en wordt echt steeds beter. Wat betekent dat dingetje? Windows is het raampje en Mac is Mac. LM Studios, waar we het vandaag over gaan hebben. Die kun je downloaden. Ik ga straks laten zien hoe. We hebben ook Ollama, zeer populair. Allemaal Windows Mac. We hebben ook Jan. Het ziet er gezellig uit. Het wordt nog veel gezelliger, want we hebben ook Osaurus voor de Mac. Er zijn wat ontwikkelaars bezig om het zo gezellig mogelijk te maken. Maar dit zijn dus allemaal software waar je die modellen, waar je het eerder over had. Dus OpenAI heeft een model en QN en hoe heet die meuk allemaal. Dat moet je dus allemaal dan, je kan kiezen in welke app je die wil draaien. En dat ben je nu aan het laten zien. Ja, heel mooi. Goed dat je dit zegt. In al die apps kan je nagenoeg dezelfde modellen draaien. Helder. En ze zien er allemaal net iets anders uit of zo. Ja, ik was gewoon een beetje waar je mee wil werken. Wat mooi, als je zegt Bolt, kan je ook nog weer, kan je wel ook OpenAI doen. Dus dan ga je zeggen, oh, ik ga door elkaar hybrid werken en zo. Maar goed, ik heb ook, ik weet dat heel veel Windows mensen kijken. Dus ik wilde ook jullie zo min mogelijk Apple dingen geven. Dan op de smartphone is er locally AI voor de Apple gebruikers. Ga ik zo laten zien. Kan je dus ook diezelfde modellen opdraaien, maar wel kleiner. Want het was veel minder VRAM. Tuurlijk, want die iPhone er zit veel minder geheugen in. Ja, en Google heeft eigenlijk zelf de Google AI Edge Gallery. Meest rare naam voor een local AI app ever. Maar goed, het komt erom neer dat ze een galerij hebben gemaakt van de mogelijkheden wat betreft Google Edge AI. Ja. Bel een best wel interessante app moet ik zeggen. Ik vind het wel gaaf gedaan. En jij gaat ook vertellen wat we er dan mee kunnen. Ja, absoluut. Met welke modellen, met hoe duur mijn laptop moet gaan worden. Ja, een beetje het verschil. Ik ga hierna naar de Mediamarkt. Alexander is er klaar voor. Als jij nu al overtuigd bent, dan gaat het goed. LM Studio, die gaat Alexander installeren als hij bij de Mediamarkt geweest is. Hij gaat naar de website, hij downloadt hem. Hier staat Anders Windows. Als je op een Windows machine zit, het werkt prima. Oké, dus gewoon de app, lege app. Ja, en je bent nog niet in vliegtuigmodus nu. Niet meteen enthousiast denken, wie zei dat het lokaal werkt? We moeten nog even wat dingen downloaden. Ja, nou, ik ga naar mijn modellen. Daar staat nog niks. Ik heb niks. Ik kan nergens meer. Download je model. Wauw, toch? Ja. Andere wereld, hè? Dat doen we. Het is een soort modellen app store. Ja, en ik denk dat in hoe ik jou ken, als jij hierop gedrukt had en je had niet met mij hier gezeten, dat je had gezegd, laat maar. Ja. Mogelijk had je het ChatGPT-logo daarboven gezien. Ja, ik herken het Google-logo. Dus dat helpt. Ik moet zeggen, dit hebben ze wel goed aangepast. Inmiddels is het een standaard rijtje geworden. Want de meeste mensen willen natuurlijk met de S4 ChatGPT. En download ik één model of download ik er meerdere? We beginnen even met één. Ja, maar is er een reden om meerdere modellen te downloaden? Absoluut,

want ze zijn allemaal beter in het een en ander. Ah, jeetje. Oké. Oh, en dit is een heel liefschattig lijstje in vergelijking met wat daadwerkelijk gebeurd is. Je weet de open source wereld, het is een bazaar hè? En wat gebruik jij als jouw go-to? Gemma 431B. Gemma 431B, mensen. En trouwens, 4-bit kwant, hè? 4-bit kwant. Schrijf mee. Hij gaat hem downloaden. Daar is hij. Ik ga ook uitleggen waarom. Oh, jij hebt met de ham nu iets gezocht wat niet het meest gelijkt is. Tuurlijk. Dit is wat jij als sommelier natuurlijk weet. Ja, en ik ga ook nog uitleggen waarom die wijn zo heet en uit die regio komt. Oké, nou, wij volgen jou gewoon. Ja, oké. Oh, Jezus. Haha. Ja, wat is dit nou weer? Oké, MLX Community, kennen jullie natuurlijk allemaal. Nee, MLX Community is een community op Hugging Face, het platform waar al deze modellen verzameld worden. En MLX is een platform, sorry, een framework van Apple, waardoor die modellen veel sneller draaien op Apple hardware. Er zijn ook aanpassingen voor ander soort chips, voor Nvidia, die het natuurlijk allemaal maken, noem maar op. Je moet er wel rekening mee houden hè? Als je in die Apple wereld zit. Als ik een Mac heb, moet ik zoeken op MLX. Anders GGUF. GGUF. Dan hebben we Gemma 4. Dat is op zich logisch. De vierde generatie van de open source lijn van Gemini. Juist. Dus dit zijn Google's gratis lokale modellen. Omgezet door de MLX community in ander formaat. 31 biljoen parameters. Een beetje de grootte van het brein, zou je kunnen zeggen. En die draait dus alleen op een laptop van minimaal 3.500 euro. Ja, deze heeft 16 gig nodig. Ik heb 24. Dus anders crasht hij ook. Want dan is er niks meer over voor de rest van je laptop. Dus ik zoek hem een beetje uit op mijn laptop. Instruct. Ik moest zelf ook even denken. Wat is dit nou? Heb ik opgezocht. Dit gaat erover, je kunt ermee praten. En dan denk je misschien, dat mag ik hopen, dat ik ermee kan praten. Nee, ik weet nog, voordat wij de podcast begonnen, dat wij zaten te spelen met GPT, ver voor chat GPT. En dan moest je zeggen, dan moest je een soort begin maken van een nieuwsartikel en dan drukken op extend. Ja, hij kan alleen maar aanvullen. Je kan niet mee praten. Ja, dus op een gegeven moment hebben ze gezegd, we gaan fine tunes doen en dat ding laten aanvullen alsof hij praat. Dat zijn de instruct models. 4-bit stond erbij. Oké, sorry, Instruct. Ik heb dit al gezegd. Je kunt conversaties doen. En dan denk je misschien, oh, maar is dat standaard niet zo? Als je dit niet doet, ik heb het al heel vaak gehad, krijg je hele matige gesprekken. Want dat ding is helemaal niet gebouwd om gesprek te voeren. Hij heeft als het ware geen communicatievaardigheden meegekregen met mensen. 4-bit is de kwant. Daar hebben we het al eerder over gehad. Hoe lager de kwant, hoe matiger het model. Ja, 4 is niet top. Nee, maar 4 is wel het mooie gemiddelde. Oké. We gaan hem downloaden. Maar oké, jij hebt net geheimtaal uitgelegd en dan hoe we die moeten ontcijferen. Maar ik zag in die appstore gewoon de meest gelijkte. Waarom download ik die niet dan? Omdat je dan toch een beetje teleurgesteld gaat zijn in de kwaliteit. Dus jij zegt, er is nog best wat uitzoekwerk wat voor jouw laptop. Is dat iets wat je aan chat kan vragen? Van mijn laptop ziet er

ongeveer zo uit. Je maakt me zin af, absoluut. En dan zegt hij, dan is dit misschien een goed model voor jou. Er is ook een Hugging Face MCP. Als je die koppelt, gaat hij zelf met Hugging Face praten om erachter te komen wat bij jouw laptop past. Ah, oké. Daarom dus ook hier, het is niet een anti-clouten jij verhaal. Nee, zet die hele zware clouten jij in om jou te helpen om thuis wat in te richten. Ja, precies. Dat is de truc. Oké, dus die downloaden we nu. Even 18 fucking gigabyte. Ja, maar voor de duidelijkheid. Ik ga dat straks laten zien. Deze weet wel gewoon wie de kinderen van Willem-Alexander zijn. In die 18 gig. Maar ook wie de kinderen van de koning van Spanje zijn. Je bent Wikipedia aan het downloaden. Ja, want jij bent net de hele tijd aan het inleiden met dat ding is super dom. Maar het is dus niet per se zo. Want als je een groot model van 18 gigabyte downloadt, dan weet hij die dingen wel. Want daar heb ik nu nog niet echt gevoel voor. Wat is iets wat ik op mijn telefoon wel kan en wat is iets wat ik op mijn dure laptop niet kan. Dat is heel goed dat je dat zegt. Ik ga even praten met Gemma op mijn telefoon. Ik ben offline gegaan. Dat zit in de hoek, als jullie dat thuis ook kunnen zien. Ik ben nu offline. Ja, oké. Je hebt nu die, dat model staat nu live. Ja, ik moest hem wel downloaden natuurlijk. En jij zet hem nu aan in een modelpicker, zoals we dat ook gewend zijn van ChatPetit. Zelfde grijs ook, hè? Toch? Zelfde interface. Nu gaat hij hem inladen. Nu raakt hij in paniek. Geen zorgen, geen zorgen. Ik heb nu paniek. Oké, weten jullie nog Context Windows? Daar hebben wij jaren over gepraat al in de podcast. Het model neemt ruimte in je geheugen. en de rest van je geheugen wordt ingenomen... door hoe groot jij je context window wil hebben. Dus ik... daarom... nou ja, ik kan wel zeggen... ik open dit en ik denk... zo lekker man. Ik heb gewoon een laptop voor Gemma... 31B instruct in 4-bit quant. Ja, maar wel met een context window... van 4000 tokens. Nou, dat is helemaal niks... in wat er nu in de cloud kan. Maar voor wat ik wilde laten zien vandaag wel. Dus eigenlijk laten we iets hier zien... dat die een beetje te... GPU-poor noemen ze dat. Dat je een soort arm bent in je computerkracht. Oké. Maar even het context window... is dus de hoeveelheid geschiedenis... van een gesprek. Of hoeveel het grote documentje erin mag gooien. Hoeveel die kan onthouden. Dus een boek erin gooien vereist... een minimum hoeveelheid context. 100.000. En hier probeer jij... 4.000 te krijgen. En daar raakt hij al van in paniek. Dat wil ik er wel even bij zeggen... LM Studio probeert te gokken hoeveel geheugen die gaat innemen. Omdat anders als het teveel is, loopt je laptop vast. Hij weet dat voor Gemma 4 nog niet, omdat Gemma 4 te nieuw is. Dus staat hier, hey dude, ga jij maar gokken. Ik durf het niet, anders gaat je laptop stuk. Misschien heb je wel een belangrijk document aan te staan. Nee, je laptop gaat niet stuk. Nee, hij loopt vast. Hij gaat zeker niet stuk, dat is het verkeerde woord. Dus ik doe een klein contextwindowtje. Dan druk ik op load anyway, want ik skip all permissions. En 4000 heb je gewoon bedacht. Ja. Ja, want ik dacht, ik wil gewoon even een verhaaltje maken. Leuk joh. En wat krijg ik met 4000? Dat is ongeveer 4000, dus een brief,

een A4'tje. Ja, en als je daar dus buiten gaat, dan op een gegeven moment zie je dan, de marro has unloaded, want dan wordt die er gewoon uitgekikt door Apple waarschijnlijk, die zegt, wat ben jij aan het doen in je app? Jij gaat weg. En dan blijft die wel draaien. Dat is best wel een goed systeem. Ik hou dit in mijn achterhoofd, dat ik hier iets mee moet. Hm, licht aan. Niet als je de models pikt zoals ik straks nog ga doen. Kan je een kort verhaal schrijven wat ik kan voorlezen aan mijn neefje voordat hij gaat slapen? Klik erop en dan komt de verhaal over Bram en de slaapwolk uit. Natuurlijk. Deze praten ook heel erg als Gemini, want ja, zelde firma. Het zijn niet dezelfde modellen, daarom heten ze ook anders, maar ze komen wel een beetje uit hetzelfde trainingsregime. Er zit hier ook trouwens onderin een geel dingetje met Vision. Dat betekent dat hij heeft gedirecteerd dat je ook plaatjes erin kan gooien. Kom ik later op. Nu zeg ik... Oh, vet man. Zo, ik heb gewoon Chachaputie thuis. Dit is echt cool. Wie zijn de kinderen van Willem-Alexander? Nou, dan gaat hij laden. Oké. Ik dacht... Zo, not bad. Eerlijk. De namen kloppen. Ik ben op Wikipedia gaan kijken. Wauw, 2003, 2005, 2007 klopt ook. En dan... En toen zag ik dat hij de datum heeft verzonnen. Oké, de rest is bij elkaar bedacht. Ja, dus dan zie je een beetje de resolutie van het model. Hij zegt van, ja, die kinderen weet ik nog wel. Ik weet zelf wanneer ze geboren zijn. Maar die details hebben ze er van mij afgesnoeid. Ik ben een soort klein struikje wat in een vrachtwagen moest passen. Maar... Hij zegt niet dat hij dat niet weet. Hij vertelt het als een feit. Vind ik wel heftig. Het is dus heel erg spannend. Dit is een beetje de oude chat GPT thuis. Ja, het is zoals drie jaar geleden. Dus zet je oude antennes weer aan. Totale modellen zijn namelijk geen databases. En dit geldt ook voor de cloud modellen. Zeker als je niet betaalt, krijg je hele kleine chat GPT modelletjes erachter. Ze moeten zoveel mogelijk grond onder hun voeten krijgen. Grond onder hun voeten krijgen, wat bedoel ik daarmee? Als je bijvoorbeeld een videomodel aan het werk zet, moet je hem niet een lege kam vastgeven, maar een storyboard. Dit is bijvoorbeeld een jongen die doet met videomodellen zijn eigen films maken. Hij tekent storyboards en laat de storyboards uitrenderen. En dat zorgt ervoor dat je eigenlijk een soort bumpers bij de bowling geeft aan je model. Je zet hem als het ware vast. Op deze manier kan je ook zeggen, hier heb je de hele stamboom van de Oranjes. Ga nu maar mijn vraag beantwoorden. Oh, dit is een metafoor. Ja, sorry. Dit is een visuele metafoor. Dit is een visuele metafoor. Voor grounding. Oftewel, geef sturing en grond onder de voet en lijnen en kaders aan het model om binnen te genereren. En als je modellen binnen een kader laat genereren, bijvoorbeeld een kenniskader of een... een visueel kader, dan zie je dus dat het resultaat dichter komt... bij wat het uiteindelijke opdracht was. Dus als jij zegt, ik wil dat je een analyse doet van mijn notulen van mijn bedrijf... dan geef je een x-aantal notulen mee. Als je die niet meegeeft, dan gaat die bluftie dat hij jouw bedrijf kent... of gaat hij aan de hand van jouw website denken dat hij je notulen bij elkaar kan verzinnen. Dit is

waarom mensen zulke rare matige antwoorden uit AI-modellen krijgen... omdat ze zich niet goed beseffen... dat die aan de ene kant heel veel weet over algemene menselijke kennis en patronen, maar heel weinig weet over hoe laat het nu is, welke temperatuur het nu is en wie de kinderen zijn van wie. Dat kan je oplossen, grounding, door een hybride te maken tussen een lokaal en een cloud. Ik ga het visueel maken voor jullie. We zaten tot nu toe offline, dat zie je ook in de hoek, op ons laptop... met LM Studio en Gemma te praten... alsof we met ChatGPT zaten te praten in onze tent zonder internet. Nu gaan we hem openbreken. Maar we breken hem niet helemaal open. We zeggen, we willen wel weer naar de cloud, maar niet naar Google. Niet naar OpenAI. We gaan naar DuckDuckGo. Alsof we zelf even gaan opzoeken op internet... wie de kinderen zijn van Willem-Alexander. En dat betekent dat je... Sorry, daar kan je er heel veel van vinden... LM Studio heeft ondersteuning voor MCPs. Mogelijkheden om te connecten naar de buitenwereld. En DuckDuckGo is een van hun MCPs. Waarmee je hem eigenlijk, wat ze dan noemen, web grounding geeft. Oftewel, je geeft hem lijnen. Je geeft hem kaders in het complete internet. Ja. Dan gaat hij doen wat jij verwacht dat je AI zou moeten doen. Namelijk weten wie die kinderen zijn en wanneer ze geboren zijn. Nu ga ik een switch maken. En dat sluit mooi aan op jouw eerdere opmerkingen. Ik ga switchen naar GPT-OSS20B. Dat is het open source GPT-model van OpenAI. Dit is natuurlijk netjes de MLX-versie. Wat voor laptop heb ik daarvoor nodig? Hoe duur is die laptop? Ook wel 3.500. Oké. Ja, echt wel geld hoor. Maar dat is beter. Ja. Nee, dit is een model die slechter is in Nederlands schrijven. Slechter is in kennis. Maar veel beter is in het aanroepen van de juiste tools. Om informatie erbij te halen. Dat noemen ze de tool calling capabilities. Ik moet zeggen, veel van wat ik je nu achter de schermen laat zien. Wordt langzamerhand nu verstopt in die apps. En vanzelf gedaan. Maar het is het wilde westen. We moeten dit nu allemaal zelf doen. Over een jaar hoeft dit allemaal niet meer. Voor de duidelijkheid. Ergens bij OpenAI staat ook een model in een GPU geladen. Dit is niet een soort technisch andere versie van AI. Nee, dit is ook hoe het in het datacenter werkt. Best wel vet. Die zullen niet deze LM Studio gebruiken natuurlijk, maar you catch my drift. Ik vraag nu, wie zijn de kinderen van Willem-Alexander? Dan kies ik voor het GPT-OS OSS20B model. Dit beantwoordt jouw vraag, moet ik er meerdere installeren? Ja. Snap ik dan waarom ik moet switchen? Praat erover met Cloud AI, zeg maar. Dus AI in de Cloud. De GPT-modellen van OpenAI zijn goede toolcallers. En Gemma kan redelijk goed schrijven. Het zijn een soort collega's van je. Oké. Ja, Gemma is best een leuke schrijver, vind ik. Nou, dan laat ik het model in. Zie geen warnings. Want dit model... Over de context. Nee, want het is van die zelf inschatten. En hij doet er een lekkere context binnen bij van 131.000 tokens. Dat besluit hij zelf. Dus hier zie je helemaal niks roods. Veel minder spannend. Hij heeft trouwens ook zelf die quant gekozen. Prima. We gaan hem laden. Wie zijn de kinderen? Onderaan staat low

reasoning. Oh, vet. Het is een reasoning model. Het is een soort O1-achtig ding. Dus hij kan ook nog denken. Voor de duidelijkheid, dat kon die Gemma niet. Denken. Ah, oké. Dat thinking dingetje wat we gewend zijn uit JetGPT. Ja. Extended thinking. Hier kun je ook nog switchen naar reasoning modes. Best wel cool. En dan zegt die, zeg ik, sorry, dan druk ik op plus. Dan zet ik mijn DuckDuckGoSearch aan die ik heb geïnstalleerd via de site die je eerder zag. En dan krijg ik er een nieuw ding bij, namelijk DuckDuckGoSearch. En nu kan ik vragen en dan krijg ik dan terug, bam, best wel tof. Dan zou je zeggen, ja boeia, hij heeft het gewoon opgezocht op internet. Ja, maar dat is wel heel prettig als jij feiten wil hebben. Of in ieder geval dichter bij de feiten wil komen, laat ik het dan zo zeggen. Maar wat kan nu op je computer wel en niet? Dus als ik op mijn dure laptop van 3500 euro zit... hoe kan ik dit gebruiken voor mijn werk? E-mail triage, samenvattingen schrijven... spel grammatica controle... sparren over een gesprek met een collega... alles wat GPT-4 kon, bijna alles. Oké, maar dat is heel wezenlijk. Dat is heel wezenlijk. Goed dat je dit vraagt. Want dan zeg je dus eigenlijk de dingen waar mensen nu Chagipati voor gebruiken. Dat ze iets copy-pasten naar Chagipati om een mail bijvoorbeeld reply te schrijven. Dat is iets wat heel goed nu lokaal en dan eigenlijk als je het heel erg plat slaat. Dan zeg je dus, gebruik dan Gemma 4 op een Mac of op een Windows computer die zwaar genoeg is om Gemma 4 op te draaien met die formule die je net gaf. En dan kan je eigenlijk probleemloos e-mails laten schrijven door dat ding lokaal. Het is echt probleemloos. Nou ja. Het is echt geen Tolstoy. Het is geen Dostoevsky. Het is een statische vrij zielloze schrijver. Ja. Maar schrijft voor meels. Maar ik denk voor mensen die bijvoorbeeld jurist. Stel je bent nu. Je bent een pitter jurist. En je wil geen cloud model gebruiken hiervoor. Dan kun je dit dus. Dit kan nu gewoon. Ja. Dan is wel. Omdat ik het woord jurist hoor. Daar trigger ik natuurlijk op. Die willen misschien bijvoorbeeld. Ik vind het wel goed dat je dit zegt. Want eigenlijk zou je. Die had jij misschien ook verwacht vandaag. Een soort. capability matrix willen hebben. Het moeilijk van wat wel, wat niet. Ja, dat vind ik heel moeilijk om te begrijpen. Maar ik moet heel eerlijk zeggen dat... dit is op dit moment ook gewoon nog best wel pittig. Want we hebben de benchmarks wel. Alleen dan zie je heel vaak... Gemma. Gemma 4. En dan kijk ik dus als AI-nerd naar... welke quant? Hoeveel parameters? Want je hebt dus Gemma in allemaal sausjes en soorten. Dus de een is de ander niet. Er zit een soort... een certification op die modellen als het ware als ze kleiner worden. En ik merk dus dat eigenlijk jouw vraag, dit is heel erg nodig wat je nu zegt, want daardoor kan je ook het gevaar voorkomen dat iemand zegt, zo, ik heb Webinar van Wietse gekeken, ik heb een dure laptop gekocht, ik heb die Gemma ingeladen, ik heb gewoon al die screenshots, 4-bit quant, wow. En nu ga ik Discovery doen als jurist. Ik gooi er 100.000 tokens in aan documenten. Gemma is niet een hele denderende search daarin. Die doet het niet goed op needle in the haystack bijvoorbeeld. Namelijk het vinden van een naald

in een hooiberg. Dus al deze AI modellen zou ik misschien voor de voorzichtigheid inschatten als early chat GPT. Ja, maar dat betekent dus eigenlijk dat je hebt gewoon jouw use cases. Waar je dan wil bedenken, kan ik hier lokale AI voor gebruiken? En dan moet je dus eigenlijk gaan heel veel dingen gaan proberen. Ja, misschien is dat wel het mooiste advies. Het is toch een beetje lui misschien van mij. Maar ga ze vergelijken. En doe alsjeblieft dan wel even je wifi aan. En zet ze gewoon naast elkaar. En test wat. Want ik was bijvoorbeeld twee dagen terug... voor een vriend van mij een document aan het doornemen. En ik was hiermee bezig. Ik dacht, ik zal die niele in de eestrek doen. Toen bleek die GPT-OSS best wel goed te zijn in het vinden van die naald. Maar er geen analyse over te kunnen doen. Dus dan heb ik dat resultaat weer doorgegooid naar Gemma. Want je kunt het ook, dat is zo gaaf, die modellen met elkaar laten praten. Maar dat is wel een stapje te ver hoor. Ja, ik pak hem even over hier. In de livestream moesten we een beetje opschieten op dit moment, want we liepen uit de tijd. Maar achteraf hebben we nog wat meer tijd. Dus ik spreek even kort een stukje in en dan pakken we hem straks weer over uit de audio die we live hebben opgenomen. Smartphones. Uiteindelijk is het zo dat je natuurlijk ook die modellen op de smartphone kan draaien. Op iOS doe je dat met Locally AI. Dat staat gewoon in de App Store. En als je die opent krijg je eigenlijk al vrij snel de vraag. Wil je Apple Foundation gebruiken? Apple Foundation beschikbaar op alle iOS toestellen die Apple Intelligence ondersteunen. Uit mijn hoofd is dat vanaf de iPhone 14 Pro volgens mij. En de latere niet-pro iPhones hebben hem ook. Foundation is eigenlijk een mini-modelletje van Apple zelf... die standaard meegeleverd wordt als je Apple Intelligence activeert. Je ziet nu Gemma 4 in beeld staan. Dat is een beetje verwarrend, maar eigenlijk staat Apple Foundation... als model geselecteerd. En op het moment dat jij hier gaat chatten... terwijl je gewoon in vliegtuigmodus bent... kan je vragen stellen en krijg jij antwoord terug. Dat gaat best wel vlot ook. Dus het aantal tokens per seconde... is best wel indrukwekkend... voor zo'n lokaal smartphone modelletje. Het is wel zo, ik vraag hier... Hoe maak ik mijn koffie extra lekker? Waarom stel ik een beetje matige vragen? Omdat ik weet dat dit model eigenlijk weinig kennis heeft... en alleen maar vooral patroonherkenning kan. Dus samenvatten, teksten corrigeren, een agenda lezen, dat soort zaken. Maar op het moment dat jij feitenvragen gaat stellen... dus dit is wederom hoe heten de kinderen van Willem-Alexander... Dan krijg jij terug Willem-Alexander, de koning van Nederland, heeft twee kinderen. Hun namen zijn Prinses Maxima en Prinses Alexia. Dat klopt natuurlijk niet. Dat komt omdat dit model zo ver teruggesnoeid is om te passen op een smartphone, om te passen op de iPhone, dat het eigenlijk niet gezien moet worden als een kennismodel, maar alleen als een patroonherkenningsmodel. Crowning op mobiel is helaas nog moeilijk. Ik had het eerder over hoe je een call kan doen naar DuckDuckGo. Om wel daadwerkelijk data uit bijvoorbeeld Wikipedia te halen. Hoe je eigenlijk kunt zorgen dat die

gesnoeide modellen met weinig kennis corpus. toch ad hoc in het moment... gegrond kunnen worden. Het is gewoon nog moeilijk, omdat er zijn niet echt... goede apps. Locally AI bijvoorbeeld ondersteunt... geen toolcalling, dus je kan niet het web erbij halen. En er zijn wat experimentele apps. Op Android kan het wat beter. Dat is wel mooi. Dan multimodaliteit... Maar in locally AI kan je andere modellen laden dan Apple Foundation. Bijvoorbeeld de Gemma-serie. Dat kan je vinden onder instellingen en dan manage models. Staat op dit moment bovenaan Gemma 4. Sowieso een prettig voordeel van apps als locally AI, maar ook Google's AI-contact. Edge Gallery, ik weet het niet meer precies uit mijn hoofd, is dat je ook een curatie hebt. Dus eigenlijk een curatie van modellen die sowieso redelijk goed werken op mobiele apparaten, waardoor je niet door zo'n, nou ja, nagenoeg eindeloze lijst van Hugging Face hoeft te gaan scrollen. Dus hier heb je eigenlijk ook meteen een leuk overzichtje van kleine, redelijk kapabele modellen. Nou, ik kies hier voor Gemma 4. Ja. Die gaat hij downloaden. Dus uit mijn hoofd een 4 gigabyte ongeveer. En als je die eenmaal hebt gedownload. Kun je hem kiezen in de modelpicker. Dus die staat er nu bij. Het E2B modellen zijn modellen. Die eigenlijk gemaakt zijn. Om niet alleen maar teksten te verwerken. Maar ook naar afbeeldingen te kijken. Bijvoorbeeld en audio te luisteren. Hier kan je dan bijvoorbeeld in plaats van een vraag te stellen over Willem-Alexander, een vraag stellen over een foto. Ik vraag hier, wat zie je op deze foto? En dan gaat hij eventjes nadenken en na wat inference komt hij met zijn token. Ze zegt, ik zie een display met cosmetische producten en dan nog een heel verhaal. En dat klopt ook. Het is best wel knap hoe dit kleine baby garnaal modelletje toch weer best wel een knappe analyse van een foto doet. *** Dit kan natuurlijk als jij bijvoorbeeld slechtziend bent. En het zit ingebouwd in een bril. Best wel krachtig zijn. Dan. Waar eigenlijk alles samenkomt, is het moment dat je cloud AI, dus de krachtige AI modellen van de verschillende AI labs en bedrijven, gaat inzetten om een soort artifacten te maken. Dus applicaties voor jou, die dan daarna weer gebruik maken van lokale AI. Wat bedoel ik? Je hebt je laptop, daar staat in dit geval LM Studio op met Gemini als model. Nu is die nog offline. De cirkel staat er omheen en de datacenter heeft geen toegang. Dan ga je praten met bijvoorbeeld een model van OpenAI. En dan vraag je dat zware model om voor jou een applicatie te maken. Die vervolgens weer gebruik maakt van lokale AI. Wat bedoel ik daarmee? Je zegt tegen Codex. Wil jij voor mij een applicatie maken waardoor ik de teksten in Word kan redigeren? Dus ik selecteer een stuk tekst en ik druk op een bepaalde toetsencombinatie. En dan wordt die tekst geredigeerd. Dus de fouten worden eruit gehaald en wordt kort gemaakt. En dat doet hij op dat moment wanneer je reageert met een van je lokale modellen. Dus je hebt eigenlijk een tijdelijke verbinding nodig met de cloud AI om vervolgens een applicatie te laten maken die weer met de lokale AI aan de gang gaat. Waardoor je als die applicatie eenmaal klaar is, je deuren weer dicht kan doen, de luiken

weer dicht kan doen als het ware en weer verder kunt. Hoe doe je dat dan? LM Studio heeft ook een developer mode. Waarin je eigenlijk jouw LM Studio niet alleen maar beschikbaar stelt aan jezelf. Om ermee te kletsen, wat wij eerder hebben gedaan. Maar dat je je LM Studio beschikbaar stelt aan alle andere applicaties op jouw laptop of desktop. Daardoor kan een app die gebouwd is door een andere AI ook praten met jouw Gemini. Dus wat je doet is je gaat naar developer. Je kiest een model dat je inlaat, want die moet wel in je geheugen geladen worden, anders is die niet beschikbaar. En op dat moment wordt het model erin geladen, in dit geval GPT-OSS20B. En als die helemaal geladen is, is er op de achtergrond binnen jouw computer een endpoint, een toegangspunt, waarop dat taalmodel zichzelf beschikbaar stelt. Dan kan je naar Codex gaan of naar Cloud Code of naar Anti-Gravity. Welk van de harnassen jij ook prettig vindt. En dan stel je de vraag, ik heb een LM Studio Endpoint, want die heb je net aangezet. Kan je een native macOS app maken, waardoor ik tekst kan selecteren in iedere app. En die kan laten redigeren door een taalmodel naar keuze. Dan gaat Codex rommelen en die maakt eigenlijk al vrij snel voor jou een menulet. In het geval van de Mac, een applicatie die draait in je menubar. waar je op kan klikken en daar kan je dingen redigeren. Bijvoorbeeld redigeren, vervang selectie van tekst, redigeren naar klembord. Deze app staat offline. Je ziet het ook, ik heb het offline-icoon weer in de hoek gezet. Dus die is online gegenereerd en draait vervolgens offline verder. En kan je bijvoorbeeld in Word, zoals ik hier doe, gebruiken. Dus ik selecteer hier een stuk tekst. Dan doe ik een toetsencombinatie die bij die app hoort, die gemaakt is door Codex... En die vervangt de tekst door op de achtergrond het endpoint aan te roepen binnen LM Studio. Mocht je het allemaal niet meer helemaal volgen. Wat we hier voor elkaar hebben gekregen is dat we niet alleen maar kunnen praten met lokale AI modellen in LM Studio. Maar ook in andere applicaties op ons bestuurdersysteem die zelf de modellen die lokaal zijn aan kunnen roepen. En dit opent natuurlijk een enorme Pandora's box met interessante mogelijkheden. Ik heb hier even een voorbeeld laten zien waarin je ziet oké ik heb offline datacenter is niet nodig en binnen de laptop heb ik een LM Studio Endpoint met daarin een Gemini model geladen en een lijntje naar die redigeer app en die praten met elkaar allemaal binnen dezelfde zandbox eigenlijk zandbak als dat je gewend bent van lokale AI. *** Wat kunnen we dan nog meer bouwen? Nou, hier zeg ik, kan je een native MacOS applicatie voor me maken... die een map met scans van documenten verwerkt en omzet in tekstbestanden. Ik heb een LM Studio Endpoint met een OCR model. Dus ik heb hier op de achtergrond een andere Gemma geladen. Ik weet dat dat een model is die goed is in afbeeldingen. Je hebt hem net gezien. In de locally AI app eerder waar ik de afbeelding van de zeepjes en de product display liet zien. Datzelfde model laat ik hier lokaal in, alleen een grotere versie ervan. Dus die is een stuk krachtiger. Ik krijg een applicatie uit Codex. Ik zet mijn internet uit. En vanaf dat

moment kan ik mappen vol met in dit geval ingescande gedichten omzetten in selecteerbare tekst. Tekst waar ik vervolgens weer iets mee kan. Dus je gaat van een rauw ingescand gefotografeerd document naar daadwerkelijke tekst toe. En je kunt je voorstellen dat je hier foto-ontdubbelingsapplicaties... foto's recht zetten, zodat alle horizon recht staat. Foto's sorteren aan de hand van thema of kleuren. Voor iedere foto en of video en of audio transcript te genereren. Ja, eerlijk... Dit is best wel gaaf wat je eigenlijk kan als je je beseft dat je met je lokale model kan communiceren vanuit andere applicaties. En dat die applicaties ontwikkeld kunnen worden door een cloud model. Waardoor je applicaties super geavanceerd zijn. Maar nadat ze gegenereerd zijn eigenlijk die cloud AI niet meer aan hoeven roepen. Want je lokale model is voldoende. Dus ook hier zie je de OCR applicatie maakt veel meer. Verbinding met de endpoint van LM Studio die weer het model geladen heeft en de data heen en weer stuurt. Allemaal offline op je eigen laptop. Ja, sorry, ik moest je hierdoor injagen. Heel goed, heel goed. Want we gaan naar jullie vragen. Want Rien zegt, hoe kijken jullie naar compliance en juridische argumenten? Dan kom ik altijd bij mijn confrère Wietshagen. Om gewoon co-pilot te gebruiken binnen organisaties. Die beloven immers niet te trainen op je data en alles binnen je organisatie tenen te houden. In onze optiek nog steeds niet doen. Zeker niet bij Amerikaanse partijen. Ja, oké, maar... Dus hij zegt eigenlijk, wat zijn de compliance juridische argumenten om gewoon co-pilot te gebruiken als die ook allerlei veiligheidseisen hebben. Volledig afhankelijk van je eigen risicoassessment van je eigen organisatie op strategisch niveau. Als jullie glasvezel wordt doorgeknipt per ongeluk of iemand rijdt er met een vrachtwagen overeen en het hele bedrijf kan niet bij Co-Pilot. Hoeveel economische schade is dat? Dit soort vragen. Ja, het is als blijkt dat jouw data opgevraagd is door partij X of Y in een onderzoek. Is dat een probleem? Is dat een risico? Uiteindelijk is het het oude cloud met de... Het oude datacenter vraagstuk. Hoe afhankelijk ben ik van computers buiten mijn bedrijfsgebouw? Alleen dan on steroids keer 10. Omdat je langzaam zal zien dat je organisatie steeds meer zal gaan leunen op deze AI tooling. Ja. Hoe dicht wil je bij je energiecentrale zitten? Heb je een aggregaat staan? Dit zijn dezelfde vragen. Hoe afhankelijk wil je zijn van Amerikaanse bedrijven? En inderdaad over twee jaar. Kan ook Nederland betekenen. Ja. Het hoeft niet je achtertuin te zijn, niet wie het is in zijn tent. Joppe zegt, ik werk in een industrie waar intellectueel eigendom, contracten en andere gevoelige gegevens voor een groot deel van de werkzaamheden beïnvloeden, waardoor AI integreren in de dagelijks workflow een uitdaging is. Hoe kan AI in de dagelijkse workflow geïntegreerd worden wanneer een groot deel van de gegevens gevoelige data zijn? Ja, dat kan je dus in verschillende stappen doen. Ik sluit aan op de vorige vraag. Je kunt op papier, en dat bedoel ik niet cynisch, contractueel met je cloud provider afspraken maken over wat er met data gebeurt. Als je daar minder vertrouwen in hebt of het mag niet, kan je zeggen ik

ga voor een andere provider, een kleinere in de buurt. Desnoods een datacenter in dezelfde stad, ik noem maar iets, dezelfde regio, waar je een afspraak mee maakt om bijvoorbeeld datacenter te maken. dark fiber te leggen. Ik noem maar iets bizar. Het klinkt heel groot, maar dan ligt er een kabel van jouw bedrijf... naar hun datacenter. En dan weet je zeker dat het jouw kabel is. En daar staat jouw AI. Je kunt het on-premise doen binnen de organisatie. Dus je haalt een stuk van de computerarchitectuur daarheen. Of, wat ik binnen meer organisaties zie gebeuren... je geeft je personeel zwaardere laptops... waarop zij een deel van het werk kunnen doen. Al deze voorbeelden die ik net noem zijn geen heilige oplossingen. Ik wil alleen maar zeggen, AI hoeft niet per se heel ver van jou in een heel groot dataset te draaien. Even heel pragmatisch. Dit hele betoog is heel begrijpelijk. Alle voordelen van lokale AI. Maar ik moet toch wel na een uur concluderen dat het wel vroeg is nog. Absoluut vroeg. Want als jij zegt, ja, het ene model kan wel een beetje internetten en het andere model kan dat niet, dingen opzoeken, dan denk ik wel, kunnen ook nog een jaar wachten hiermee. Ja, en ik denk, ik heb heel erg... Dat is wel iets wat je in je achterhoofd moet houden. Supergoed, ja. De kans is heel groot dat je teleurgesteld gaat zijn als je het nu... Laten we het experimenteel noemen. Het is wel zo dat... Ik heb er dus over getwijfeld om dit wel of niet al te doen. Ik denk dat mijn doel vooral geweest is om te laten zien... dat AI in een groot datacenter niet inherent is aan AI. En daarmee wil ik zeggen dat er absoluut alternatieven zijn... dat ik een jaar geleden nog in de matigste app zat te rommelen zelf... waar ik niets van begreep en wiskunde moest begrijpen. Dit worden al best wel mooie apps. Ik zie het voor me, en dat zien we ook wel bewegen... dat de Apples, Microsoft en Googles van deze wereld... dit wat makkelijker gaan maken voor ons. Maar dat ligt natuurlijk voor... Dat is heel logisch dat dit over twee jaar allemaal veel eenvoudiger gaat zijn. Dat je niet meer een model hoeft te kiezen... dat die automatisch doet wat op jouw laptop perfect werkt. Dat die dat automatisch uitkiest. En dat betekent dus ook dat als jij... strategie wil bedenken voor jouw bedrijf... of voor jezelf... of dat je nu al rekening houdt met een scenario... waarin over anderhalf jaar de modellen zo goed zijn... als we het nu in de cloud hebben draaien. Dat is best een scenario waar je rekening mee zou kunnen houden. Goed genoeg zou ik zeggen om interessantere dingen te doen... dan die ik misschien vandaag heb laten zien. Want we zitten een beetje tussen documenten scannen... en korte verhaaltjes schrijven en onderzoek doen in. Ik snap dat dat niet super spectaculair is. Maar ik ben het met je eens. Dat wordt niet minder spectaculair de komende anderhalf jaar. Dat gaat spekken....structaculairder worden. Dus je kan nu al rekening gaan houden met.....oké, waar zitten die gevoeligheden dan? En wat zijn dan de dingen die ik... En als ik het nu vandaag ga proberen.....op kleine onderdeeltjes waarbij je gewoon.....kan copy-pasten, want het draait op je lokale computer.....dus het maakt niet uit... kun je al erachter komen wat wel en niet goed werkt. En

als ik jou goed begrijp is dat eigenlijk ook de... Jij roept mensen in eerste instantie op om nieuwsgierig daarna te zijn. Dat is goed begrijpen. En ten tweede is het ga spelen met die modellen. Want dat is wat nodig is. Ja, en ik denk dat mijn droom zou zijn als het ware... dat iemand gewoon in een gesprek zit ergens over strategie rondom AI. En dan zegt, maar het hoeft niet daar toch? Kunnen we niet zelf wat doen? Want ik zag wie ze daar toen mee bezig. Misschien is dat inmiddels wel verder of... Je vraagt hier om bijvoorbeeld een sterk gereguleerde zorgorganisatie. Zou die gebruik kunnen maken van lokale AI toepassingen? Waar denk je dan als eerste aan als je daarmee wil beginnen? Nou, als het bijvoorbeeld gaat om het transcriberen van patiënt-hulpverlener gesprekken. Hier zijn ook een aantal partijen mee bezig hoor. Dus dat kan je wel vinden als je even gaat zoeken online. Dan zeggen van oké, laten we de kleinere modellen. Bijvoorbeeld het probleem woorden naar tekst is opgelost lokaal. Het probleem samenvatten van die tekst is grotendeels opgelost lokaal. En op die manier... Maar dit is al huge. Want ik denk... Ik denk aan al die basisschooldocenten die rapportgesprekken moeten samenvatten totdat dat in een leerlingvolgsysteem staat. En die dat niet nu doen omdat ze denken, ja, wij mogen geen cloud eigen. Jij zegt nu even tussen neus en lippen door, dit is opgelost. Ja, maar je staat wel bij mij thuis in mijn garage te kijken naar een van de allereerste elektrische auto's. Je bedoelt, het is niet zo gebruiksvriendelijk als, ja, ja. Maar ik denk ook dat, want jij had het net over een soort trend naar dit gaan al die bedrijven natuurlijk ook aanbieden. Ik moet wel zeggen dat het wel zou helpen als wij als afnemer dit ook wat meer zouden verlangen. Om het even zo te zeggen. Er komen meer vegan opties op de kaart als er meer mensen vragen om salaris. Ja, dus die zorgorganisaties gebruiken software van grote softwareleveranciers. En het helpt als er druk ontstaat vanuit zorgorganisaties als voorbeeld. Kunnen we stukjes lokaal doen hiervan? Kan de audiopipeline wel lokaal? Waarom moet dit in de cloud? Maarten zegt, ik werk bij een universiteit met licenties en contracten. Hoe kunnen we een off-premise AI gebruiken op een veilige manier... die compatible is met de gevoeligheden? Ja, dit moet nu nog. Kijk, de hele grote organisaties kunnen zelfs de Googles van deze wereld inschakelen om on-premise Gemini te krijgen bijvoorbeeld. Dat bestaat ook. Maar dat gaat om budgetten die ver buiten jullie bereik zijn, vrees ik. Het zou interessant zijn om te kijken naar eigenlijk, want dit is een specifieke oplossing die als het ware ontworpen moet worden voor jouw situatie nu nog, best wel custom. waarbij je een soort consultancy-achtige partij moet hebben die dit gaat inrichten. En dat is niet heel gek, want bijvoorbeeld search appliances, oftewel ik wil wat Google doet, zoeken in heel veel data, maar dan binnen mijn eigen organisatie, dat bestaat gewoon. Dat iemand dat bij jou komt ophangen en dan komt help met documenten erin indexeren, is niet helemaal nieuw. Het idee dat we sommige dingen die we gewend zijn als consument in de cloud te hebben, dat we die binnen onze organisatie lokaal willen maken.

Alleen we zijn gewoon nog best wel vroeg. Ten slotte, als je nou wil starten hiermee, wat zijn dan ongeveer de ranges als je een laptop wil kopen? De ranges waarbij het het goedkoopst is, maar nog steeds relevant? En wat is een duurdere soort van upsell optie en wat kan je daar meer mee? Ja, ik zou de duurder optie gewoon niet doen. Want dan ga je richting een... Nvidia verkoop duur... Dan ga je richting de 10.000, 15.000 euro. En dan ben je wat mij betreft... Wat kun je dan? Als je bereid bent dat... Veel grotere context. Veel gewoon richting een half miljoen context. Modellen die in de buurt beginnen te komen... van een echt een... best wel een serieuze GPT-4, zeg maar. Dus zonder van de gaten erin die je vandaag zag. Maar ik vind alleen dat daar de prijs-kwaliteitsverhouding behalve... Kijk, als jij nu een onderzoeksinstituut bent of een universiteit, zou ik zeggen... Begin er gewoon alsjeblieft al mee. Maak er een onderzoeksbudget van. En haal wat AI-compute naar binnen. En begin met de kleine modellen, de baby-gornalen en werk op naar boven. Laten we er gewoon even vanuit gaan dat de kleinere modellen krachtiger worden en dat de machines betaalbaarder worden. Daar kan je al een lijntje voor tekenen. Dus mijn antwoord is niet richting die 10.000 euro gaan. Wat ik nu wel eens aan vrienden vertel, zeg als je in de markt bent voor een nieuwe laptop, zorg er dan even voor dat je wat meer RAM erin hebt zitten. 32 of misschien zelfs 64. Als de laptop van de baas betaald wordt, 128. Want dan kan je er grotere modellen in laden. En het wordt wel, ik zeg heel vaak tegen mensen, dat is ook de webinar vandaag. Het begint best interessant te worden eigenlijk inmiddels. Zeker met die hybride tussen ik durf ook wat cloud te gebruiken en wat lokaal door elkaar. Super. Dit is iets waar denk ik we allemaal heel geïnteresseerd in zijn. Dat dit op korte termijn heel veel beter gaat worden. En ook al is het nu nog Wild Westen. De eerste toepassingen zijn daar. Dat is wat ik vandaag heb gehoord. En dat is als je dan helemaal gaat denken over wat er dan gebeurt. Als je het combineert met elkaar. Of dat nou zoeken is. Of dat het nou software maken is met een cloud model. Die je lokaal gaat gebruiken. Denk aan een dagboek app. die jou ook nog eens proactief prompt stuurt over wat voor gedachten je allemaal wil delen met een AI, zonder dat dat naar servers gaat, maar dat die app wel is gemaakt met AI, dan komt er wel iets bij elkaar wat kan voldoen aan je nieuwsgierigheid en tegelijkertijd ook veilig en autonom blijft. Dat is toch bijzonder. Het is iets wat een half jaar geleden nog niet zo makkelijk kon. Ja, en ik denk misschien dat we zijn heel veel gewend nu van bijvoorbeeld Cloud in Word. En het is het Copilot in PowerPoint, noem het allemaal maar op. Dat bestaat dus niet voor Local AI. Dan kan je denken, ik ga wel wachten, dit bestaat niet leem. Of je kunt het gaan bouwen samen met Cloud AI. Dat wil ik gewoon even gezegd hebben. En dat, als je die schakel gaat maken, hoop ik vandaag een paar mensen, dan wordt het wel gaaf hoor. Dank voor jullie vragen. En dank voor het kletsen met elkaar in de chat. We willen over twee weken weer een webinar geven. En jullie mogen het onderwerp bepalen. En we gaan je binnenkort een mail sturen.

Over wat je voor onderwerp jij graag zou willen zien. Dat Wietse behandelt over twee weken. En op basis van jullie feedback gaan we het onderwerp kiezen. Dus hou je mail daarvoor in de gaten. Dank voor het kijken. Dank voor het meedoen met elkaar. Dat is zeer gewaardeerd. Dank Wietse. voor weer de enorme hoeveelheid compressie die je in zo'n complex onderwerp hebt gestoken. En we hopen jullie over twee weken weer te zien. Dank voor het kijken.